分层细粒度偏好优化用于物理合理的视频生成

Harold Haodong Chen[1,2]，**Haojian Huang**[3]，
Qifeng Chen[1,2]，**Harry Yang**[†1,2]，**Ser-Nam Lim**[†2,4]
[1]The Hong Kong University of Science and Technology
[2]Everlyn AI, [3]The University of Hong Kong
[4]University of Central Florida
[†]Corresponding Author
✉ haroldchen328@gmail.com

Abstract

视频生成的最新进展使得创建高质量、视觉上引人注目的视频成为可能。然而，生成遵循物理定律的视频仍然是需要真实感和准确性的应用中的一个关键挑战。在这项工作中，我们提出了 PhysHPO，这是一种用于分层跨模态直接偏好优化的新框架，旨在通过启用细粒度偏好对齐来克服这一挑战，从而实现物理上合理的视频生成。PhysHPO 在四个分层粒度上优化视频对齐：a) 实例级别，确保整体视频内容与输入提示对齐；b) 状态级别，使用边界帧作为锚点确保时间一致性；c) 运动级别，为现实的动态建模运动轨迹；d) 语义级别，保持叙述与视觉之间的逻辑一致性。认识到真实世界的视频是物理现象的最佳反映，我们进一步引入了一种自动化的数据选择管道，以便高效识别和利用现有大规模文本-视频数据集中的" 良好数据"，从而消除高成本和耗时的数据集构建设的需求。在以物理为重点和一般能力的基准上的广泛实验表明，PhysHPO 显著改善了高级模型的物理合理性和整体视频生成质量。据我们所知，这是首次探索视频生成中细粒度偏好对齐和数据选择的工作，为更真实且符合人类偏好的视频生成模式铺平了道路。 PhysHPO Page

1 引言

视频生成技术最近在制作高质量、视觉吸引力强的 [39, 73, 90, 99, 64] 以及展示真实世界场景的长时间 [28, 19, 60, 96] 视频方面取得了显著进展。尽管有这些进步，生成符合物理定律的视频仍然是一个具有挑战性且关键的研究问题。创造出物理上合理的视频的能力对于从虚拟现实到模拟等注重现实感和准确性的应用来说至关重要。

当前增强文本到视频（T2V）生成的物理保真度的努力可以粗略地分为测试时基于反思的优化和训练时基于调优的优化。测试时基于反思的方法，例如 PhyT2V [85] ，使用大型语言模型（LLM）来迭代优化初始 T2V 提示。虽然有效，这些方法显著降低了计算效率，并且在自我纠错范式中受到模型能力上限的固有限制。相比之下，近期基于调优的方法（例如 WISA [75] 和 SynVideo [98] ）侧重于传统监督微调（SFT）范式。虽然有效，SFT 严重依赖于固定的监督信号，当针对特定能力优化 [43, 51, 23] 时表现出次优的有效性。

新兴的后训练技术例如直接偏好优化（DPO）[61] 已经展示了更高效的成对优化模式，以辨别人类偏好与非偏好样本之间的差异，从而增强视觉生成模型的特定方面，例如：安全性 [63, 50] 、定制化 [43] 、超分辨率 [9] 。这表明 DPO 在物理上合理的视频生成方面具有潜力，但这一领域仍然基本未被探索。然而，最近在视频生成 [51, 94, 36, 16] 中的 DPO 工作主要集中于视频在实例层面的粗粒度对齐，这可能导致次优的偏好对齐 [31] 。我们强调，为了实现最佳对齐，特别是对于物理上合理的视频生成，需要细粒度的偏好对齐，超越视觉吸引力以纳入详细建模。

基于这些见解，我们提出了一个新颖的框架，即用于物理上合理视频生成的分层跨模态直接偏好优化，称为 PhysHPO。PhysHPO 增强了不同层次的粒度之间视频偏好的对齐。具体

Preprint. Under review.

Figure 1: PhysHPO 显著提升了视频生成的物理合理性。文本提示来自于 VideoPhy [6] 。（顶部）流体-流体：蜂蜜扩散到热牛奶中。（中间）固体-流体：一个苹果掉入一罐红酒中。（底部）固体-固体：削皮器削苹果。

而言，我们设计了四个对齐层次：❶ 实例层次：通过将总体提示内容与最合适的视频进行匹配以确保全面的对齐。❷ 状态层次：利用边界帧作为关键锚点来建立合理的状态。❸ 运动层次：通过视频中的结构信息建模运动，从而实现超越单纯像素外观的对齐。❹ 语义层次：确保描述内容与视觉展示之间的逻辑一致性。

此外，与现有的用于生成物理合理视频的 SFT 方法 [75, 98] 相比，这些方法在数据集构建中消耗了大量资源，我们认为流行的" 一模型一数据集" 范式可能不是最佳选择。不像条件引导（例如，跳舞 [103, 30] ）或风格聚焦（例如，卡通 [82, 19] ）视频生成任务，这些任务需要额外的数据注释或特定领域的数据，现实世界的视频本身就蕴含着丰富的物理动态，这表明更高效地利用数据具有潜力。为此，我们提出了一种新颖的自动化数据选择流程，以有效处理现有的大规模文本-视频数据集，从而避免为了生成物理合理的视频而进行繁琐的新数据收集工作。与现有的用于高质量视频大规模预训练的数据处理流程不同（例如，在 Open-Sora [99] 中），我们的核心想法是从大型、高质量的原始数据中选择一个与期望目标要求密切匹配的" 好数据" 子集——直观上，在其中物理规律得到明显反映的现实世界视频。据我们所知，以前的工作中没有研究过视频生成领域中的数据选择工作。总结而言，这项工作有三个方面的贡献：

- 我们引入了一种新颖的层次化跨模态 DPO（PhysHPO）框架用于视频生成，这是一种更细粒度的 DPO 策略，用于增强视频之间的对齐，优化物理上合理的视频生成。
- 我们主张利用真实世界的视频，而不是从头开始构建数据集，以实现物理上合理的视频生成。这种方法直观地反映了物理现象，并首次将数据选择问题引入到视频生成中。
- 在以物理为重点（即 VideoPhy [6] ，PhyGenBench [54] ）和一般能力（即 VBench [34] ）为基准的大量实验中表明，PhysHPO 显著提高了现有高级模型的物理合理性和整体视频生成能力。

2 相关工作

尽管生成视觉上引人注目的视频已有所进展，实现物理合理性仍然具有挑战性，正如用户和基准测试所指出的。现有的图像到视频（I2V）研究集中于从图像中解析对象并通过考虑物理属性来估计其运动，并且某些研究仅探索对象的自由落体。然而，这些方法仅限于固定的物理类别或静态场景，这限制了它们的普适性。最近的研究旨在增强 T2V 模型的更广泛

物理合理性。PhyT2V 引入了一个 LLM 用于在测试时迭代地优化提示。然而，极大增加的推理开销和固有的性能限制制约了其效果。后续研究探索了传统 SFT 以提高模型性能。具体来说，WISA 构建了一个 32 K 视频数据集，SynVideo 使用计算机图形管线合成视频数据，这两者都需要大量手动干预，耗费资源。直观上，真实世界的视频自然反映物理现象。因此，高效利用现有数据集而不产生不必要的数据效率低下是一个有趣的问题。为此，我们首次引入了视频生成的数据选择概念，自动化这一过程以有效利用现有数据资源。尽管 SFT 在预训练中表现出色，我们建议采用 DPO 后训练范式进一步建模视频对之间的差异，从而更深入地探索物理信息。

“优质数据”的数据选择 数据选择是一种关键技术，可以在不牺牲性能的情况下高效地训练模型 [3, 74] ，这在预训练 [7, 70] 和后训练 [14, 47] 阶段都得到了强调。最近的研究强调了语言模型的有效性源于大规模预训练和经过精心策划的小型指令数据集的结合 [69, 17, 18, 77, 101] 。针对语言模型，已经提出了各种复杂的方法，包括基于质量的 [45, 20, 46] 、注重多样性的 [24, 53] 、复杂性考虑 [84, 67, 35] ，以及较简单的启发式方法，如选择较长的响应和基于梯度的核心集选择 [80, 95, 59] ，这些共同证明了在语言模型训练中具有显著的优势。然而，现有的方法主要针对语言模型，视频生成中的潜力仍然很少被探索。在这项工作中，我们率先探索了专门针对后训练阶段视频生成的数据选择策略。我们提出了一种新的自动化数据选择流程，明确强调现实性、物理真实和多样性，以增强物理上合理的视频生成。

用于视频生成的直接偏好优化 DPO [61] 已经成为一种可替代传统 RLHF [58] 的有前途的方法，能够在不需要额外奖励模型的情况下增强语言模型 [31, 27, 49] 和图像生成模型 [72, 78] 。最近的工作已经将 DPO 应用于视频生成。像 VideoDPO [51] 这样的开创性工作遵循 Diffusion-DPO [72] 范式，引入了用于自适应视频评分的 OmniScore，而 HuViDPO [36] 利用反馈将输出与人类偏好对齐。OnlineVPO [94] 使用以视频为中心的模型进行离策略优化，MagicID [43] 使用 DPO 进行基于 ID 的定制。Cheng et al. [16] 采用无鉴别器的方法进行直接优化，而 GAPO [102] 为动漫视频生成引入了 AnimeReward。最近的基础模型如 Seaweed-7B [64] 和 SkyReels-V2 [10] 也在后期训练中结合了 DPO，进一步突显了其潜力。然而，现有 DPO 工作仅关注于视频实例级别的粗粒度对齐，忽视了更细致的细节。在本文中，我们旨在通过建模精细对齐来增强 DPO 的有效性，特别是用于生成物理上合理的视频。为了实现这一点，我们提出了一种新颖的分层跨模态偏好优化框架，名为 PhysHPO，能够有效捕获视频的细粒度细节。

3 预备知识

扩散模型的直接偏好优化 Diffusion-DPO [72] 将人类偏好对齐应用于迭代生成过程，消除了显式的奖励建模。给定扩散模型 $p_\theta(y|x,t)$ 和参考模型 $p_{\text{ref}}(y|x,t)$ ，具有偏好约束的去噪目标变为

$$\max_{p_\theta} \mathbb{E}_{t,x,y\sim p_\theta}[r(x,y,t)] - \beta D_{\text{KL}}\left(p_\theta(y|x,t) \parallel p_{\text{ref}}(y|x,t)\right), \tag{1}$$

其中 $r(\cdot)$ 表示时间相关的奖励函数，x 表示输入条件，y 表示生成的样本，t 表示时间步。DPO 通过去噪路径建立了轨迹级别的奖励映射：

$$r(x,y,t) = \beta \log \frac{p_\theta(y|x,t)}{p_{\text{ref}}(y|x,t)} + \beta \log Z(x,t), \tag{2}$$

其中 β 控制 KL 约束强度，$Z(x,t)$ 为时间相关的分区函数。偏好优化目标通过替代奖励参数化并应用负对数似然损失来得出：

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \log \sigma\left(u(x,y_w,y_l,t)\right), \; u = \underbrace{\log \frac{p_\theta(y_w|x,t)}{p_{\text{ref}}(y_w|x,t)}}_{\text{Preferred path score}} - \underbrace{\log \frac{p_\theta(y_l|x,t)}{p_{\text{ref}}(y_l|x,t)}}_{\text{Non-preferred path score}}. \tag{3}$$

4 数据选择：从物理到多样化

现有的文本-视频数据集已经经过严格的数量和质量筛选 [99, 57, 48, 15] 。虽然为不同任务创建特定的数据集很流行 [93, 33, 19, 92] ，但对于物理上合理的视频生成来说，真实世界的视频本身反映了物理定律。因此，选择现有的数据可能比构建新数据集更优 [75, 98] 。在本节中，我们初步探讨视频生成中的数据选择，特别关注识别能有效反映物理定律的“优质数据”。

视频生成模型经历预训练来学习世界知识。对这些模型进行微调使其与特定目标对齐，类似于语言模型中的指令微调。数据选择已被证明对通过识别一个小的高质量数据集来使语

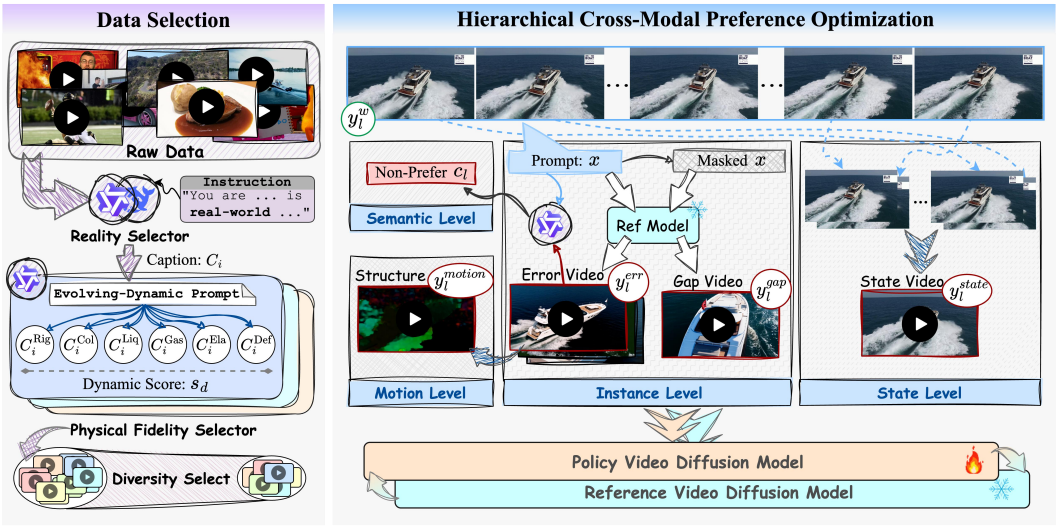


Figure 2: 我们提出的 (左) 数据选择和 (右)PhysHPO 框架概述。

言模型对齐是有效的。考虑一个大的数据池，其中每个数据点由一个视频及其对应的字幕组成。目标是选择一个大小为 m 的子集，以最大化后训练的性能：

虽然在特定领域中有多种潜在的数据选择方法，但我们的目标是使过程尽可能简单以具备可操作性，如图 2 （左）所示。接下来我们将详细说明该过程。

4.1 现实世界物理选择

给定一个大规模高质量的文本-视频数据集 D （在这项工作中，我们采用了普遍用于后训练的 OpenVidHD-0.4M [57] ），第一步是将真实世界的视频过滤到一个现实数据池 D' 。具体来说，考虑到真实世界视频和虚拟视频之间的显著内容差异，我们采用视觉语言模型（VLMs）（即 Qwen2.5-VL [4] ）来判断视频 V_i 是否是真实世界的视频。我们进一步委托 DeepSeek-VL2 [79] 进行双重检查以确保准确性。表格 ?? 显示了从 D 随机抽取的 1,000 个视频的准确率。

物理逼真度 为了与以前的工作对齐，我们采用了来自于 WISA [75] 的三种物理类别的 17 物理现象：❶ 动力学（刚体运动、碰撞、液体运动、气体运动、弹性运动、变形）、❷ 热力学（熔化、凝固、汽化、液化、燃烧、爆炸）和 ❸ 光学（反射、折射、散射、干涉和衍射、不自然光源）。与广泛使用的基于启发式分类的数据选择方法（在语言模型中需要一个目标数据集作为参考）不同，我们遵循流行的 LLM-as-a-Judge 范式 [53, 26, 102] 使用 LLMs 进行自动数据评估。

为了最小化消耗，我们在字幕层级进行筛选。给定一个视频字幕 C_i ，一个简单的方法是让一个大型语言模型（LLM）直接对样本是否清楚地反映物理定律进行评分。然而，由于缺乏参考，LLM 可能会给大多数样本分配相似的分数 [32] 。为了解决这个问题，我们提出采用深入演变提示 [84] 来增强字幕。与先前的工作 [53, 84] 在单一维度（例如，复杂性）增强样本不同，我们探索多个维度。具体来说，如图 2 所示，我们首先使用设计的演变动态提示和 LLM（即，Qwen2.5 [86] ）来增强字幕，强调“动态”类别中的七种物理现象中的每一种。这有助于模型更精确地区分字幕之间的差异，实现细粒度评分。然后我们要求 LLM 对这七个样本进行排序和评分，得到对应于字幕 C_i 的“动态”分数 s_d 。同样，可以获得“热力学”分数 s_t 和“光学”分数 s_o 。这一策略提供了对物理现象更为细致的区分。详细信息见附录 §B 。

总得分计算为 $s = s_d \times s_t \times s_o$ 。然后，D' 中的所有样本根据 s 进行排序，生成排序池 $S' = \{x'_1, x'_2, \cdots, x'_k\}$ ，其中 x'_0 是得分最高的样本。

4.2 选择多样性

为了确保一个先进的生成模型能够处理各种用户提示，在给定预算下，数据最好保持最大程度的多样性。然而，现实世界的数据常常表现出冗余。为此，我们引入了一种迭代方法，遵循 Deita 的指导，确保在选定的现实世界数据中保持多样性。简单地说，关键思想是在数据池中迭代选择样本，只有当它们对数据集的多样性有贡献时，才将它们添加到数据集中。形式上，我们定义了一个指示函数，如果满足多样性标准则为，否则为。使用 LLaMA-1 13B，对字幕进行编码生成嵌入，然后计算每个候选样本与其最近邻居之间的余弦距离。如果满足条件，则一个样本对多样性有贡献。

我们将不同数据选择策略的性能与原始的 DPO [72] 进行比较，同时展示了我们的方法 PhysHPO，如图 ?? 所示。总结了几个关键观察：观察 ❶ 数据选择的优越性。我们的数据（21K），因实际性和物理逼真性被选择，展示出比其他策略更优的质量，如直接评分、随机选择，甚至是手动构建的数据集 WISA-32K [75] 和未选择的原始数据（433K）。`"Our Data (21K)+DPO"` 优于 `"Direct Scoring (21K)+DPO"` 、 `"Random Selection (21K)+DPO"` 、 `"WISA-32K+DPO"` 和 `"Raw Data (433K)+DPO"` ，突出了我们选择的有效性。观察 ❷ 数据多样性的重要性。`"Our Data (21K)+DPO"` 与 `"Ours w/o diversity (59K)+DPO"` 的比较强调了多样性在数据选择中的关键作用。尽管样本较少，多样性的数据集优于更大的数据集，表明多样性增强了模型的能力。观察 ❸ 战略数据和优化的协同作用。我们选择的数据与 PhysHPO（ `"Our Data (21K)+PhysHPO"` ）的结合实现了最高性能，显示了战略数据选择与高级优化之间的强大协同作用。更详细的分析见附录 §B 。接下来我们详细介绍我们的 PhysHPO 框架。

5 方法论

为了解决具有挑战性的物理可信视频生成问题，我们提出了 PhysHPO，它在四个层次上实现了分层偏好优化（图 2 ）：(i) 实例级整体偏好优化，调整视频级偏好以确保整体质量（§5.1 ）；(ii) 状态级边界偏好优化，在开始和结束帧锚定物理状态以保持稳定性（§5.2 ）；(iii) 运动级动态偏好优化，利用结构信息准确建模和对齐运动表示（§5.3 ）；(iv) 语义级一致性偏好优化，确保叙述与视觉之间的细致连贯性（§5.4 ）。

5.1 实例级整体偏好优化

与仅依赖于其基础模型进行偏好对视频数据 (y_w, y_l) 进行优化的 VideoDPO [51] 不同，我们的方法利用通过我们数据选择过程选择的数据作为偏好视频 y_w 及其文本提示 x 。对于非偏好视频 y_l ，最近的工作 [102, 94] 通常使用基础模型基于 x 生成这些视频，并遵循方程 (3) 作为目标函数。然而，非偏好视频中的复杂依赖性往往被忽略。通常，视频生成面临两个主要问题：① 在物理方面的缺陷或错误，例如运动；② 无法完全表达提示的信息，例如缺失对象。为此，我们在优化过程中引入了两种类型的非偏好视频：

$$\mathcal{L}_{\text{Instance}} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \log\sigma\left(\beta\log\frac{p_\theta(y_w|x,t)}{p_{\text{ref}}(y_w|x,t)} - \beta\log\frac{p_\theta(y_l|x,t)}{p_{\text{ref}}(y_l|x,t)}\right), \tag{4}$$

其中

$$\log\frac{p_\theta(y_l|x,t)}{p_{\text{ref}}(y_l|x,t)} \leftarrow \sum_{i\in\{err,gap\}} \beta_i \log\frac{p_\theta(y_l^i|x,t)}{p_{\text{ref}}(y_l^i|x,t)}. \tag{5}$$

在此，y_l^{err} 表示与偏好样本在语义上对齐的易出错或不完善的非偏好视频。相反，y_l^{gap} 代表与偏好样本在语义上存在差异的非偏好视频。具体来说，为了构建 y_l^{err} ，我们使用基础模型为每个 x 生成三个视频，并通过相似度计算选择与偏好视频 y_w 最相似的作为被拒绝的视频 y_l^{err} 。对于 y_l^{gap} ，我们在生成过程中在提示 x 中引入随机掩码，以创建与 y_w 的语义差异。

5.2 状态级边界偏好优化

虽然实例级的对齐捕捉了整体内容，但细粒度的对齐对于生成物理上合理的视频是至关重要的，而以前的方法常常忽略了这一点。在视频生成中，从序列的开始到结束保持一致的物理状态对于确保现实性和连贯性是必不可少的。初始帧和最终帧特别关键，因为它们锚定了视频的物理叙述。为此，我们提出了状态级边界对齐，旨在增强模型对视频开始和结束时物理状态的关注。具体而言，我们替换首选视频 y_w 的前 N 帧和最后 N 帧，以构建状态级的非首选样本 y_l^{state} 。状态级目标函数可以定义为

$$\mathcal{L}_{\text{State}} \sim \log\sigma\left(u_{\text{State}}(x, y_w, y_l^{state}, t)\right). \tag{6}$$

。

5.3 运动级动态偏好优化

虽然视频的视觉外观对于质量至关重要，但过度关注外观可能会掩盖模型与物理动态的对齐。为了解决这一问题并增强模型对动态信息（如运动）的学习，我们建议从首选 y_w 和非首选 y_l （具体为 y_l^{err} ）视频中提取结构信息（例如光流）。这些结构特征通常用于条件引导的视频生成 [83, 22, 100] ，允许更明确地表示首选和非首选样本之间的物理运动差异。根据公式 (6)，运动层次的目标函数 $\mathcal{L}_{\text{Motion}}$ 由 $u_{\text{Motion}}(x, y_w \to y_w^{motion}, y_l \to y_l^{motion}, t)$ 导出，其中 y_w^{motion} 和 y_l^{motion} 分别表示从 y_w 和 y_l 中提取的结构信息。

Table 1: 在物理专注的基准上进行评估，即 VideoPhy [6] 、PhyGenBench [54] ；以及通用质量基准，即 VBench [34] 。原始的 DPO 算法用我们选择的数据实现。我们用粗体显示最佳结果，“↑”表示数值越高越好。

Method	VideoPhy [6]				PhyGenBench [54]					VBench [34]		
	SS ↑	SF ↑	FF ↑	Over. ↑	Mechanics ↑	Optics ↑	Thermal ↑	Materials ↑	Overall ↑	Total ↑	Quality ↑	Semantic ↑
CogVideoX-2B [90]	12.7	21.9	25.4	18.6	0.38	0.43	0.34	0.39	0.39	81.6	82.5	77.8
+ PhyT2V [85]	14.1	19.9	28.6	18.9	0.45	0.48	0.34	0.50	0.45	82.0	82.9	78.4
+ Vanilla DPO [72]	15.5	19.2	28.6	19.2	0.43	0.50	0.40	0.50	0.46	82.2	83.1	79.0
+ PhysHPO (Ours)	20.4	24.7	42.9	25.9	0.50	0.56	0.47	0.58	0.53	82.5	83.3	79.3
CogVideoX-5B [90]	24.4	53.1	43.6	39.6	0.39	0.55	0.40	0.42	0.45	81.9	83.1	77.3
+ PhyT2V [85]	25.4	48.6	55.4	40.1	0.45	0.55	0.43	0.53	0.50	82.3	83.3	78.3
+ Vanilla DPO [72]	28.2	50.0	51.8	41.3	0.48	0.60	0.47	0.58	0.54	82.4	83.3	78.7
+ PhysHPO (Ours)	32.4	54.1	58.9	45.9	0.55	0.68	0.50	0.65	0.61	82.8	83.7	79.3

5.4 语义级一致性偏好优化

大多数之前在扩散模型中的 DPO 工作主要依赖视觉对齐进行优化。然而，我们提出利用语言来更精确地描述视觉差异，提供文本语义层次上的特定参考信息。受 LMs 中的 DPO 启发 [61] ，我们通过进一步在文本层次上建模视频的语义信息，引入了语义层次的一致性对齐。具体来说，对于每个偏好的视频对 y_w 和 y_l ，我们首先将提示语 x 视为偏好的说明 c_w 。然后，一个 VLM（即 Qwen2.5-VL [4] ）负责基于 y_l 修改 c_w ，调整不一致的部分以生成 c_l 。该过程确保 c_w 和 c_l 仅在视频不一致的地方在文本描述上有所不同，从而促进更有针对性的优化。然后，在 $\mathcal{L}_{\text{Semantic}}$ 中的 $u_{\text{Semantic}}(c_w, c_l, y_w, t)$ 被公式化为：

$$u_{\text{Semantic}} = \log \frac{p_\theta(y_w|c_w,t)}{p_{\text{ref}}(y_w|c_w,t)} - \log \frac{p_\theta(y_w|c_l,t)}{p_{\text{ref}}(y_w|c_l,t)}. \tag{7}$$

通过集成实例、状态、运动和语义层次的偏好优化，PhysHPO 的整体损失函数定义如下：

$$\mathcal{L}_{\mathcal{PHYSHPO}} = \mathcal{L}_{\text{Instance}} + \lambda\mathcal{L}_{\text{State}} + \rho\mathcal{L}_{\text{Motion}} + \mu\mathcal{L}_{\text{Semantic}}, \tag{8}$$

，其中 λ 、ρ 和 μ 表示损失权重。

6 实验

在本节中，我们进行了广泛的实验以回答以下研究问题：（RQ1）PhysHPO 是否增强了生成视频的物理合理性？（RQ2）PhysHPO 是否影响其他性能方面？（RQ3）PhysHPO 对其关键组件的敏感性如何？（RQ4）PhysHPO 在效率、效果和广泛适应性方面是否比 SFT 更有优势？

6.1 实验设置

我们将 PhysHPO 应用于高级模型 CogVideoX-2B 和 5B。由于基于 SFT 方法的开源代码和模型权重不可用，即 WISA 和 SynVideo，我们将比较重点放在以下基准：PhyT2V、普通的 DPO、SFT 以及各自的基础模型。在 HunyuanVideo 上的实现和结果显示在附录中。

评估 为了评估 PhysHPO 的有效性，我们采用了专注于两个关键方面的基准测试：❶ 物理聚焦：(i) VideoPhy [6] 用于固体-固体、固体-流体和流体-流体的相互作用。(ii) PhyGenBench [54] 用于力学、光学、热学和材料学。❷ 通用能力：VBench [34] 用于整体质量和语义。

我们在 份

实现细节 选定的数据集上训练基础模型，使用全局批量大小为 8 ，并采用 AdamW 优化器和学习率为 $2e-5$ 。实例级非首选权重设置为 $\beta_{err} = 0.7$ 和 $\beta_{gap} = 0.3$ ，状态级样本的权重设为 $N = 2$ 。损失权重 λ 、ρ 和 μ 分别设置为 0.4 、0.3 和 0.2 。所有实验均在 8 台 NVIDIA H100 GPU 上进行。

6.2 PhysHPO 的物理与总体性能

为了解答 RQ1 和 RQ2，我们在物理相关和一般基准测试（表 1 ）上全面比较了 PhysHPO 与三种基线方法，并在图 1 、 3 和图 4 （左）中分别展示了定性结果和用户研究。我们的观察总结如下：观察 ❹ 表明 PhysHPO 在增强物理真实性和一般质量方面表现优异。如表 1 所示，我们的 PhysHPO 在三个物理聚焦基准测试中持续优于基线方法（即 PhyT2V [85] 和普通 DPO [72] ），这些方法在每个节中都针对物理合理性的不同维度只表现出微小甚至负面的性能提升。表 1 进一步突出了它在 VBench [34] 上对于一般能力的稳健性。图 1 和图 3 中的定性比较提供了对 PhysHPO 能力的视觉确认，展示了它比基础模型的明显优势。观察 ❺ 表明 PhysHPO 有效地使视频生成与人类偏好对齐。图 4 （左）展示了在物理和一般维度上进行的用户研究，其中 PhysHPO 在视频生成与人类偏好对齐方面持续优于基线方法，证明了其在偏好对齐方面的优越性。

Figure 3: PhysHPO 的定性结果。由于篇幅限制，其他基线的结果在附录 §D 中提供。文本提示来源于 PhyGenBench [54] 。（1）材料：一个脆弱、易碎的鸡蛋被用很大的力气投向坚固的岩石表面，并在碰撞时发生撞击。（2）力学：一个有弹性的网球被用力扔向地面，展示其与表面的动态相互作用。（3）光学：放大镜逐渐靠近硬币，揭示压纹设计的复杂细节和纹理。（4）热学：一个时间流逝视频捕捉到生奶在水壶中温度快速上升至 100°C 以上时的变

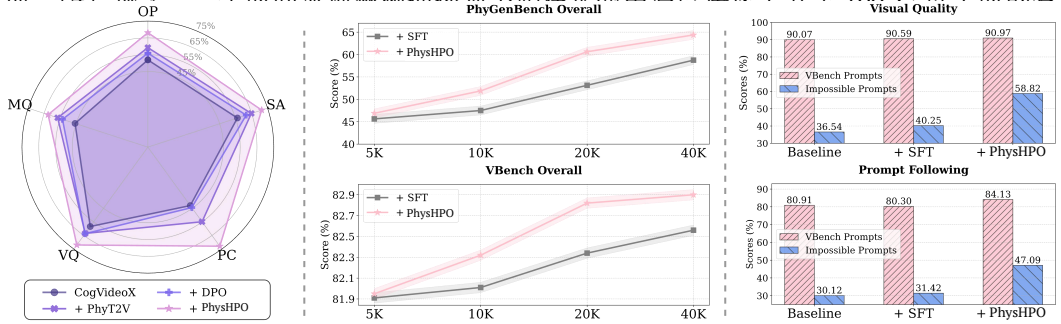


Figure 4: （左）用户研究涵盖五个维度：总体偏好、语义遵循、物理常识、视觉质量和运动质量。（中）在不同数据量下，PhysHPO 和 SFT 性能比较。（右）使用 IPV-Txt [5] 的鲁棒性测试。

6.3 消融分析

为了回答 RQ3，我们在 VideoPhy 上进行了 PhysHPO 各个级别及其组合贡献的评估，结果如表格 ?? 所示。我们给出以下观察：观察 ❻ 分层偏好优化的有效性。表格 ?? 显示，当逐步移除 $\mathcal{L}_{\text{Semantic}}$ 、$\mathcal{L}_{\text{Motion}}$ 和 $\mathcal{L}_{\text{State}}$ 时，性能呈现出稳定的下降。这突显了分层偏好优化框架的有效性，其中每个级别都对提高物理真实性有独特贡献。观察 ❼ 细粒度对齐的重要性。细粒度对齐，无论是显式的还是隐式的，对于优化都是至关重要的：i）去除实例级间隙视频

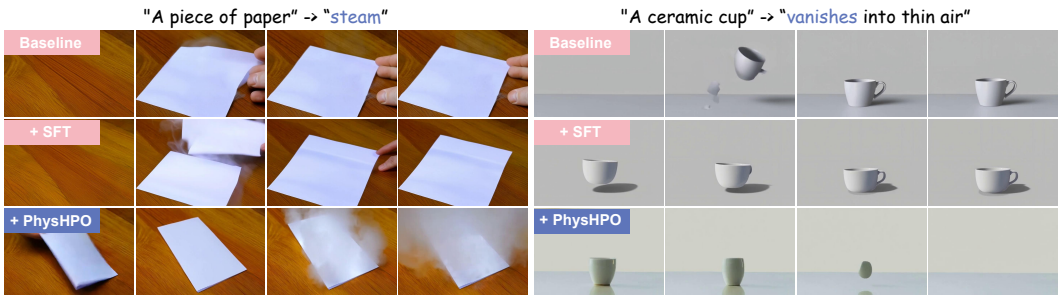


Figure 5: 稳健性测试演示。详细提示可以在附录 §C 中找到。

对齐（"Only w/ $\mathcal{L}_{\text{Instance}}$ " 对比 "Vanilla DPO w/ Our Data" ）突出显示了在填补生成与真实语义和动态之间差距时视觉显式建模的重要性。ii）排除 $\mathcal{L}_{\text{Semantic}}$ 揭示了隐式文本-视频对齐对于捕捉文本提示与生成视频之间微妙且连贯关系的必要性。

6.4 PhysHPO 与 SFT 的分析

为了解答 RQ4，我们使用 PhyGenBench 上的相同数据比较了 PhysHPO 和 SFT 的物理真实度，使用 VBench 比较一般质量，如图 4 （中）所示。受 [5] 的启发，我们进一步用不可能的提示（例如，“一辆车穿过海洋，好像在飞一样”）评估两种方法，以评估物理真实度的改进是否也能更好地处理富有想象力或物理上不可能的场景，结果在图 4 （右）和图 5 中展示。我们给出以下观察：观察 ❽ ，PhysHPO 是一个数据高效的视频生成增强器。图 4 （中）表明，在不同数据量的物理聚焦和一般质量指标中，PhysHPO 一直优于 SFT。这反映了 PhysHPO 通过细粒度信息利用更有效地利用训练数据以实现卓越性能的能力。观察 ❾ ，PhysHPO 使得超越物理定律的创造性泛化成为可能。如图 4 （右）和图 5 所示，PhysHPO 在不可能的提示上表现更好，生成的视频更接近语义意图，同时保持内部一致性。这表明提高物理真实度不仅改善了对现实物理学的遵循，还赋予模型更强的泛化和适应富有想象力场景的能力，展示了超越简单物理规则的学习。

7 结论

在本文中，我们提出了 PhysHPO，一个用于分层跨模态直接偏好优化的新框架，增强了视频在四个层次上的细粒度对齐：实例、状态、运动和语义。认识到真实世界的视频是物理现象最佳的反映，我们引入了一个自动化数据选择管道，以高效利用大规模文本-视频数据集，消除了对详尽数据集构建的需求。在物理重点和一般基准上的广泛评估表明，PhysHPO 显著提升了现有视频生成模型的物理合理性和整体质量，解决了物理上合理的视频生成中的关键挑战。

References

[1] Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*, 2023.

[2] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

[3] Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, et al. A survey on data selection for language models. *arXiv preprint arXiv:2402.16827*, 2024.

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[5] Zechen Bai, Hai Ci, and Mike Zheng Shou. Impossible videos. *arXiv preprint arXiv:2503.14378*, 2025.

[6] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024.

[7] David Brandfonbrener, Hanlin Zhang, Andreas Kirsch, Jonathan Richard Schwarz, and Sham Kakade. Color-filter: Conditional loss reduction filtering for targeted language model pretraining. *Advances in Neural Information Processing Systems*, 37:97618–97649, 2024.

[8] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL `https://openai.com/research/video-generation-models-as-world-simulators`.

[9] Miaomiao Cai, Simiao Li, Wei Li, Xudong Huang, Hanting Chen, Jie Hu, and Yunhe Wang. Dspo: Direct semantic preference optimization for real-world image super-resolution. *arXiv preprint arXiv:2504.15176*, 2025.

[10] Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Juncheng Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengchen Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025.

[11] Haodong Chen, Lan Wang, Harry Yang, and Ser-Nam Lim. Omnicreator: Self-supervised unified generation with universal editing. *arXiv preprint arXiv:2412.02114*, 2024.

[12] Harold Haodong Chen, Harry Yang, and Ser-Nam Lim. Beyond generation: Unlocking universal editing via self-supervised fine-tuning. *arXiv preprint arXiv:2412.02114*, 2024.

[13] Harold Haodong Chen, Haojian Huang, Xianfeng Wu, Yexin Liu, Yajing Bai, Wen-Jie Shu, Harry Yang, and Ser-Nam Lim. Temporal regularization makes your video generator stronger. *arXiv preprint arXiv:2503.15417*, 2025.

[14] Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*, 2023.

[15] Shoufa Chen, Chongjian Ge, Yuqi Zhang, Yida Zhang, Fengda Zhu, Hao Yang, Hongxiang Hao, Hui Wu, Zhichao Lai, Yifei Hu, Ting-Che Lin, Shilong Zhang, Fu Li, Chuan Li, Xing Wang, Yanghua Peng, Peize Sun, Ping Luo, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Goku: Flow based video generative foundation models. *arXiv preprint arXiv:2502.04896*, 2025.

[16] Haoran Cheng, Qide Dong, Liang Peng, Zhizhou Sha, Weiguo Feng, Jinghui Xie, Zhao Song, Shilei Wen, Xiaofei He, and Boxi Wu. Discriminator-free direct preference optimization for video diffusion. *arXiv preprint arXiv:2504.08542*, 2025.

[17] Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

[18] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, et al. Ultrafeedback: Boosting language models with scaled ai feedback. *arXiv preprint arXiv:2310.01377*, 2023.

[19] Karan Dalal, Daniel Koceja, Gashon Hussein, Jiarui Xu, Yue Zhao, Youjin Song, Shihao Han, Ka Chun Cheung, Jan Kautz, Carlos Guestrin, et al. One-minute video generation with test-time training. *arXiv preprint arXiv:2504.05298*, 2025.

[20] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*, 2023.

[21] Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, et al. What's in my big data? *arXiv preprint arXiv:2310.20707*, 2023.

[22] Yutang Feng, Sicheng Gao, Yuxiang Bao, Xiaodi Wang, Shumin Han, Juan Zhang, Baochang Zhang, and Angela Yao. Wave: Warping ddim inversion features for zero-shot text-to-video editing. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024.

[23] Jinlan Fu, huangfushenzhen, Hao Fei, Xiaoyu Shen, Bryan Hooi, Xipeng Qiu, and See-Kiong Ng. CHip: Cross-modal hierarchical direct preference optimization for multimodal LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=7lpDn2MhM2`.

[24] Yuan Ge, Yilun Liu, Chi Hu, Weibin Meng, Shimin Tao, Xiaofeng Zhao, Hongxia Ma, Li Zhang, Boxing Chen, Hao Yang, et al. Clustering and ranking: Diversity-preserved instruction selection through expert-aligned quality estimation. *arXiv preprint arXiv:2402.18191*, 2024.

[25] David Grangier, Simin Fan, Skyler Seto, and Pierre Ablin. Task-adaptive pretrained language models via clustered-importance sampling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=p6ncr0eTKE`.

[26] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

[27] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143, 2024.

[28] Yuwei Guo, Ceyuan Yang, Ziyan Yang, Zhibei Ma, Zhijie Lin, Zhenheng Yang, Dahua Lin, and Lu Jiang. Long context tuning for video generation. *arXiv preprint arXiv:2503.10589*, 2025.

[29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[30] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024.

[31] Haojian Huang, Haodong Chen, Shengqiong Wu, Meng Luo, Jinlan Fu, Xinya Du, Hanwang Zhang, and Hao Fei. Vistadpo: Video hierarchical spatial-temporal direct preference optimization for large video models. *arXiv preprint arXiv:2504.13122*, 2025.

[32] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

[33] Yuzhou Huang, Ziyang Yuan, Quande Liu, Qiulin Wang, Xintao Wang, Ruimao Zhang, Pengfei Wan, Di Zhang, and Kun Gai. Conceptmaster: Multi-concept video customization on diffusion transformer models without test-time tuning. *arXiv preprint arXiv:2501.04698*, 2025.

[34] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[35] Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A Smith, Iz Beltagy, et al. Camels in a changing climate: Enhancing lm adaptation with tulu 2. *arXiv preprint arXiv:2311.10702*, 2023.

[36] Lifan Jiang, Boxi Wu, Jiahui Zhang, Xiaotong Guan, and Shuang Chen. Huvidpo: Enhancing video generation through direct preference optimization for human-centric alignment. *arXiv preprint arXiv:2502.01690*, 2025.

[37] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *Advances in Neural Information Processing Systems*, 37:48955–48970, 2024.

[38] Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*, 2024.

[39] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[40] Simon Le Cleac'h, Hong-Xing Yu, Michelle Guo, Taylor Howell, Ruohan Gao, Jiajun Wu, Zachary Manchester, and Mac Schwager. Differentiable physics simulation of dynamics-augmented neural objects. *IEEE Robotics and Automation Letters*, 8(5):2780–2787, 2023.

[41] Chenyu Li, Oscar Michel, Xichen Pan, Sainan Liu, Mike Roberts, and Saining Xie. Pisa experiments: Exploring physics post-training for video diffusion models by watching stuff drop. *arXiv preprint arXiv:2503.09595*, 2025.

[42] Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E Gonzalez, et al. Worldmodelbench: Judging video generation models as world models. *arXiv preprint arXiv:2502.20694*, 2025.

[43] Hengjia Li, Lifan Jiang, Xi Xiao, Tianyang Wang, Hongwei Yi, Boxi Wu, and Deng Cai. Magicid: Hybrid preference optimization for id-consistent and dynamic-preserved video customization. *arXiv preprint arXiv:2503.12689*, 2025.

[44] Jialuo Li, Wenhao Chai, Xingyu Fu, Haiyang Xu, and Saining Xie. Science-t2i: Addressing scientific illusions in image synthesis. *arXiv preprint arXiv:2504.13129*, 2025.

[45] Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. *arXiv preprint arXiv:2308.12032*, 2023.

[46] Ming Li, Lichang Chen, Jiuhai Chen, Shwai He, Jiuxiang Gu, and Tianyi Zhou. Selective reflection-tuning: Student-selected data recycling for llm instruction-tuning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 16189–16211, 2024.

[47] Ming Li, Yong Zhang, Shwai He, Zhitao Li, Hongyu Zhao, Jianzong Wang, Ning Cheng, and Tianyi Zhou. Superfiltering: Weak-to-strong data filtering for fast instruction-tuning. *arXiv preprint arXiv:2402.00530*, 2024.

[48] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024.

[49] Aiwei Liu, Haoping Bai, Zhiyun Lu, Yanchao Sun, Xiang Kong, Xiaoming Simon Wang, Jiulong Shan, Albin Madappally Jose, Xiaojiang Liu, Lijie Wen, Philip S. Yu, and Meng Cao. TIS-DPO: Token-level importance sampling for direct preference optimization with estimated weights. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=oF6e2WwxX0`.

[50] Runtao Liu, Chen I Chieh, Jindong Gu, Jipeng Zhang, Renjie Pi, Qifeng Chen, Philip Torr, Ashkan Khakzar, and Fabio Pizzati. Safetydpo: Scalable safety alignment for text-to-image generation. *arXiv preprint arXiv:2412.10493*, 2024.

[51] Runtao Liu, Haoyu Wu, Zheng Ziqiang, Chen Wei, Yingqing He, Renjie Pi, and Qifeng Chen. Videodpo: Omni-preference alignment for video diffusion generation. *arXiv preprint arXiv:2412.14167*, 2024.

[52] Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision*, pages 360–378. Springer, 2024.

[53] Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=BTKAeLqLMw`.

[54] Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*, 2024.

[55] Antonio Montanaro, Luca Savant Aira, Emanuele Aiello, Diego Valsesia, and Enrico Magli. Motioncraft: Physics-based zero-shot video generation. *Advances in Neural Information Processing Systems*, 37:123155–123181, 2024.

[56] Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models learn physical principles from watching videos? *arXiv preprint arXiv:2501.09038*, 2025.

[57] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024.

[58] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[59] Xingyuan Pan, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, and Shanbo Cheng. G-dig: Towards gradient-based diverse and high-quality instruction data selection for machine translation. *arXiv preprint arXiv:2405.12915*, 2024.

[60] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. *arXiv preprint arXiv:2310.15169*, 2023.

[61] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

[62] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[63] Shouwei Ruan, Zhenyu Wu, Yao Huang, Ruochen Zhang, Yitong Sun, Caixin Kang, and Xingxing Wei. Towards nsfw-free text-to-image generation via safety-constraint direct preference optimization. *arXiv preprint arXiv:2504.14290*, 2025.

[64] Team Seawead, Ceyuan Yang, Zhijie Lin, Yang Zhao, Shanchuan Lin, Zhibei Ma, Haoyuan Guo, Hao Chen, Lu Qi, Sen Wang, et al. Seaweed-7b: Cost-effective training of video generation foundation model. *arXiv preprint arXiv:2504.08685*, 2025.

[65] Dian Shao, Mingfei Shi, Shengda Xu, Haodong Chen, Yongle Huang, and Binglu Wang. Finephys: Fine-grained human action generation by explicitly incorporating physical laws for effective skeletal guidance. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1905–1916, 2025.

[66] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015.

[67] Haoran Sun, Lixin Liu, Junjie Li, Fengyu Wang, Baohua Dong, Ran Lin, and Ruohui Huang. Conifer: Improving complex constrained instruction-following ability of large language models. *arXiv preprint arXiv:2404.02823*, 2024.

[68] Xin Tan, Yuetao Chen, Yimin Jiang, Xing Chen, Kun Yan, Nan Duan, Yibo Zhu, Daxin Jiang, and Hong Xu. Dsv: Exploiting dynamic sparsity to accelerate large-scale video dit training. *arXiv preprint arXiv:2502.07590*, 2025.

[69] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model, 2023.

[70] Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari Morcos. D4: Improving llm pretraining via document de-duplication and diversification. *Advances in Neural Information Processing Systems*, 36:53983–53995, 2023.

[71] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[72] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024.

[73] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[74] Jiahao Wang, Bolin Zhang, Qianlong Du, Jiajun Zhang, and Dianhui Chu. A survey on data selection for llm instruction tuning. *arXiv preprint arXiv:2402.05123*, 2024.

[75] Jing Wang, Ao Ma, Ke Cao, Jun Zheng, Zhanjie Zhang, Jiasong Feng, Shanyuan Liu, Yuhang Ma, Bo Cheng, Dawei Leng, et al. Wisa: World simulator assistant for physics-aware text-to-video generation. *arXiv preprint arXiv:2503.08153*, 2025.

[76] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.

[77] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

[78] Xianfeng Wu, Yajing Bai, Haoze Zheng, Harold Haodong Chen, Yexin Liu, Zihao Wang, Xuran Ma, Wen-Jie Shu, Xianzu Wu, Harry Yang, et al. Lightgen: Efficient image generation through knowledge distillation and direct preference optimization. *arXiv preprint arXiv:2503.08619*, 2025.

[79] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

[80] Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *International Conference on Machine Learning (ICML)*, 2024.

[81] Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. Data selection for language models via importance resampling. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=uPSQv0leAu`.

[82] Jinbo Xing, Hanyuan Liu, Menghan Xia, Yong Zhang, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Tooncrafter: Generative cartoon interpolation. *ACM Transactions on Graphics (TOG)*, 43(6):1–11, 2024.

[83] Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, et al. Make-your-video: Customized video generation using textual and structural guidance. *IEEE Transactions on Visualization and Computer Graphics*, 2024.

[84] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, 2023.

[85] Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. *arXiv preprint arXiv:2412.00596*, 2024.

[86] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[87] Nianzu Yang, Pandeng Li, Liming Zhao, Yang Li, Chen-Wei Xie, Yehui Tang, Xudong Lu, Zhihang Liu, Yun Zheng, Yu Liu, et al. Rethinking video tokenization: A conditioned diffusion-based approach. *arXiv preprint arXiv:2503.03708*, 2025.

[88] Xindi Yang, Baolu Li, Yiming Zhang, Zhenfei Yin, Lei Bai, Liqian Ma, Zhiyong Wang, Jianfei Cai, Tien-Tsin Wong, Huchuan Lu, et al. Vlipp: Towards physically plausible video generation with vision and language informed physical prior. *arXiv e-prints*, pages arXiv–2503, 2025.

[89] Yu Yang, Aaditya K Singh, Mostafa Elhoushi, Anas Mahmoud, Kushal Tirumala, Fabian Gloeckle, Baptiste Rozière, Carole-Jean Wu, Ari S Morcos, and Newsha Ardalani. Decoding data quality via synthetic corruptions: Embedding-guided pruning of code data. *arXiv preprint arXiv:2312.02418*, 2023.

[90] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[91] Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*, 2025.

[92] Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactory: Creating new games with generative interactive videos. *arXiv preprint arXiv:2501.08325*, 2025.

[93] Shenghai Yuan, Jinfa Huang, Yujun Shi, Yongqi Xu, Ruijie Zhu, Bin Lin, Xinhua Cheng, Li Yuan, and Jiebo Luo. Magictime: Time-lapse video generation models as metamorphic simulators. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[94] Jiacheng Zhang, Jie Wu, Weifeng Chen, Yatai Ji, Xuefeng Xiao, Weilin Huang, and Kai Han. Onlinevpo: Align video diffusion model with online video-centric preference optimization. *arXiv preprint arXiv:2412.15159*, 2024.

[95] Jipeng Zhang, Yaxuan Qin, Renjie Pi, Weizhong Zhang, Rui Pan, and Tong Zhang. Tagcos: Task-agnostic gradient clustered coreset selection for instruction tuning data. *arXiv preprint arXiv:2407.15235*, 2024.

[96] Lvmin Zhang and Maneesh Agrawala. Packing input frame contexts in next-frame prediction models for video generation. *arXiv preprint arXiv:2504.12626*, 2025.

[97] Peiyuan Zhang, Yongqi Chen, Runlong Su, Hangliang Ding, Ion Stoica, Zhenghong Liu, and Hao Zhang. Fast video generation with sliding tile attention, 2025. URL `https://arxiv.org/abs/2502.04507`.

[98] Qi Zhao, Xingyu Ni, Ziyu Wang, Feng Cheng, Ziyan Yang, Lu Jiang, and Bohan Wang. Synthetic video enhances physical fidelity in video synthesis. *arXiv preprint arXiv:2503.20822*, 2025.

[99] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

[100] Xiaojing Zhong, Xinyi Huang, Xiaofeng Yang, Guosheng Lin, and Qingyao Wu. Deco: Decoupled human-centered diffusion video editing with motion consistency. In *European Conference on Computer Vision*, pages 352–370. Springer, 2024.

[101] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021, 2023.

[102] Bingwen Zhu, Yudong Jiang, Baohan Xu, Siqian Yang, Mingyu Yin, Yidi Wu, Huyang Sun, and Zuxuan Wu. Aligning anime video generation with human feedback. *arXiv preprint arXiv:2504.10044*, 2025.

[103] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision*, pages 145–162. Springer, 2024.

A 混元视频的更多结果

为了进一步验证我们提出的 PhysHPO 的优越性，在本节中，我们进一步将 PhysHPO 应用于由 FastVideo [97] 实现的 HunyuanVideo-540p [39] 。

Method	VideoPhy [6]				PhyGenBench [54]					VBench [34]		
	SS ↑	SF ↑	FF ↑	Over. ↑	Mechanics ↑	Optics ↑	Thermal ↑	Materials ↑	Overall ↑	Total ↑	Quality ↑	Semantic ↑
HunyuanVideo [39]	19.7	43.2	42.9	33.4	37.5	58.0	36.7	45.0	45.6	81.4	83.1	74.9
+ PhyT2V [85]	21.1	45.9	50.0	36.3	40.0	58.0	33.3	47.5	46.3	80.5	82.0	74.5
+ PhysHPO (Ours)	26.8	50.7	55.4	41.6	47.5	62.0	43.3	60.0	54.4	81.9	83.5	75.7

Table 2: 在 HunyuanVideo [39] 上评估 PhysHPO。

定量结果 表 2 展示了 PhysHPO 在 HunyuanVideo 上的量化结果。一致于我们在主要内容中的观察 ❹ ，PhysHPO 在增强物理逼真度和整体视频质量方面表现出色。这些结果进一步验证了其有效性。

我们在图 6 中展示了结果。关于人类动作的更多结果展示在第 §D 节。

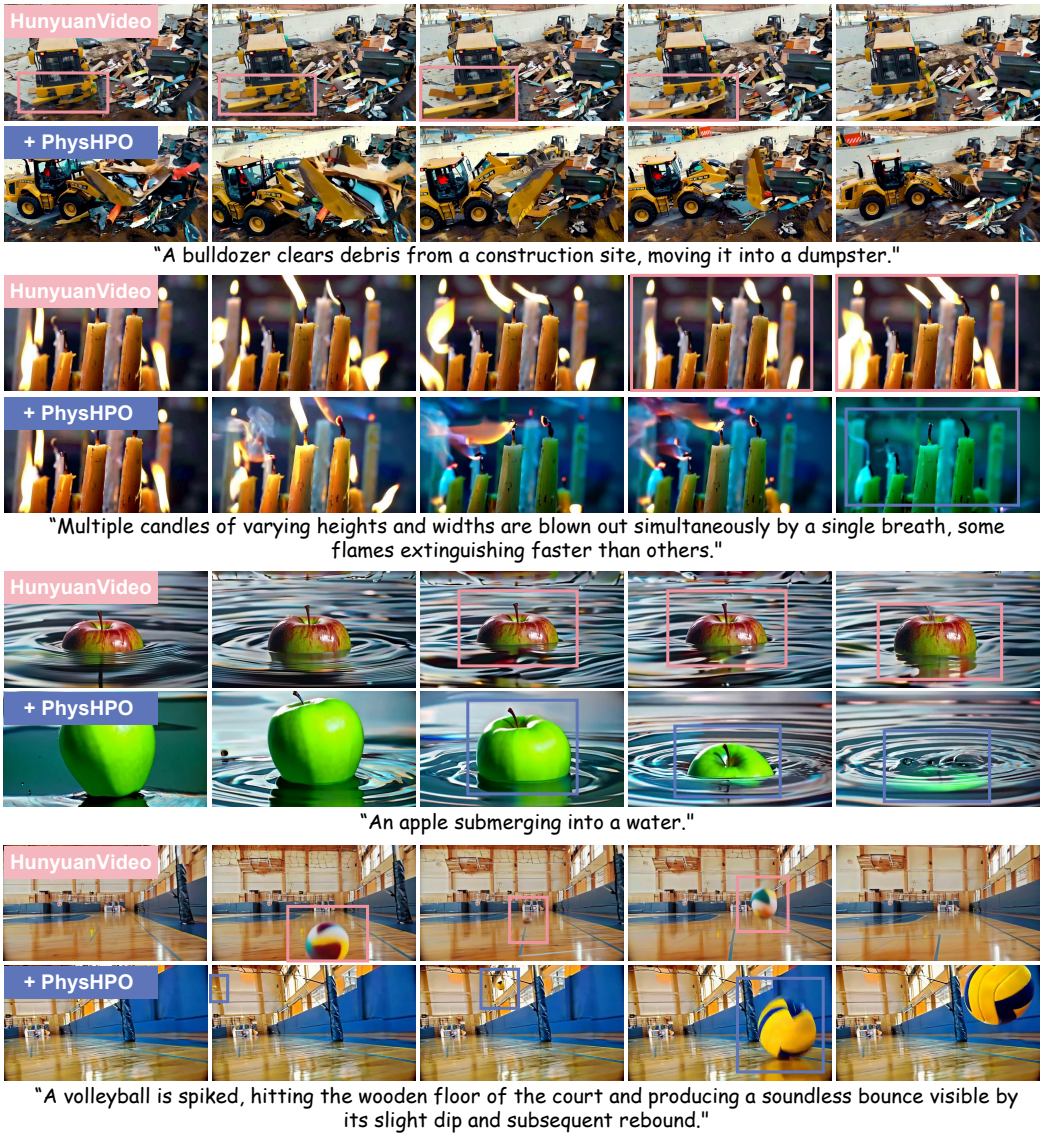


Figure 6: 在 HunyuanVideo [39] 上对 PhysHPO 的定性展示。

B 数据选择的更多细节

深入演变提示 为了实现对现实世界视频物理逼真度的更细粒度评分，我们采用了深入的演变提示策略 [84] 。这种方法通过补充示例描述来强调特定的物理现象。我们设计的演变动态提示示例如图 7 所示。

Evolving Dynamic Prompt

```
prompt: '''You are an expert in physics and natural language processing, specializing in analyzing and enhancing textual
descriptions based on physical phenomena. Your task is to process a given {caption} and perform the following steps.

### Step 1: Enhance the caption based on six physical phenomena
For the given {caption}, enhance it to emphasize each of the following six dynamic phenomena. The enhancement must strictly be
based on the original caption's content. Do not invent new events, objects, or scenarios. Instead, focus on rephrasing and
elaborating on the existing information in the caption to highlight the specific characteristics of each phenomenon. Ensure
that the enhanced captions are vivid, precise, and aligned with the corresponding physical behavior:

1. Rigid body motion: Focus on the movement of solid objects that maintain their shape and structure, such as rotation or
translation.
2. Collision: Emphasize interactions where two or more objects come into contact and exchange momentum or energy.
3. Liquid motion: Highlight any aspects of the caption that involve fluid-like behaviors, such as flow, splashing, or other
liquid characteristics.
4. Gas motion: If the caption involves air or gas, describe behaviors such as diffusion, expansion, or turbulence. If no gas
motion is mentioned, rephrase the caption to make any relevant gaseous interactions clearer without adding new elements.
5. Elastic motion: Focus on stretching, compressing, or oscillatory motion of elastic materials if present in the caption.
Elaborate on any such motions to make them more vivid.
6. Deformation: Highlight changes in shape or structure of materials under stress, such as bending, twisting, or breaking.

Output six enhanced captions, each tailored to one of the above phenomena. Ensure that all enhancements remain faithful to the
original caption's content.

### Step 2: Score the enhanced captions
After generating the six enhanced captions, evaluate each caption based on how well it represents the corresponding dynamic
phenomenon. Provide a score from 1 to 5 (1 = poor, 5 = excellent) for each caption. Additionally, provide a brief reason for
the score, explaining why the caption is effective or where it could be improved.

### Final Output Format:
Your output should follow this exact format:
Caption of rigid body motion: [Enhanced caption]; Score: [1-5]; Reason: [Explanation].
Caption of collision: [Enhanced caption]; Score: [1-5]; Reason: [Explanation].
Caption of liquid motion: [Enhanced caption]; Score: [1-5]; Reason: [Explanation].
Caption of gas motion: [Enhanced caption]; Score: [1-5]; Reason: [Explanation].
Caption of elastic motion: [Enhanced caption]; Score: [1-5]; Reason: [Explanation].
Caption of deformation: [Enhanced caption]; Score: [1-5]; Reason: [Explanation].

### Example:
...

'''
```

Figure 7: 精心设计的动态进化提示的展示。

同样，演化热力学和演化光学提示遵循相同的结构。

除图 ?? 中显示的进化评分和直接评分的性能比较外，我们还在图 ?? 中提供了在不同策略下具有详细评分的示例。具体来说，我们比较了一个动态搅拌的样本（顶部）与另一个相机拉远但场景保持静止的样本（底部）。如所示，直接评分策略对这两个截然不同的视频赋予了相似的分数，而我们的进化评分则提供了更细致的区分和准确的评估。

在我们的数据选择过程中，我们首先进行了基于视频的筛选以确保真实性，然后进行了基于字幕的筛选以增强物理逼真性和多样性。在这里，我们进一步验证了我们基于字幕的选择策略的效率和有效性。

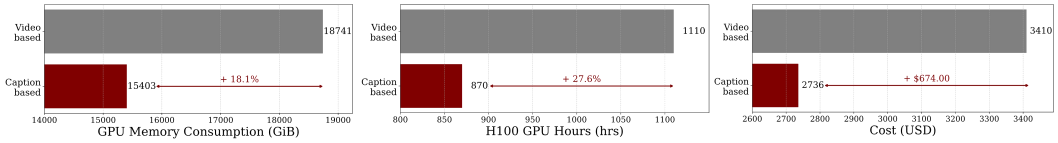


Figure 8: 基于视频的选择与我们的基于字幕的选择的效率比较。（左）GPU 内存消耗（GiB）。（中）H100 GPU 小时数（小时）。（右）成本（美元）。

图 8 首先展示了我们基于字幕的选择与基于视频的选择的效率比较。具体来说，❶ GPU 内存消耗（左）：基于字幕的选择需要显著更少的 GPU 内存，与基于视频的选择相比，使用量减少了 18.1 %（15, 403 GiB 对比 18, 741 GiB）。这一减少突显了我们方法在内存占用方面的计算效率。❷ H100 GPU 小时（中）：基于视频的选择增加了 27.6 % 的计算时间（3, 410 小时对比 2, 736 小时），表明基于字幕的选择在时间上更有效，并且更适合于大规模处理。❸ 成本（右）：在大规模情况下，基于视频的工作流程比基于字幕的工作流程多出 $ 674.00 的成本，这进一步强调了我们方法的成本效益。

为了进一步验证我们基于字幕的选择方法的有效性，我们从实境选择的数据池中随机抽取了 100 个数据点进行人工评分。然后，我们通过将人工评分结果与三种评分方法进行比较来分析评分准确性：基于视频的评分、基于字幕的直接评分以及我们基于字幕的演变评分。如

图 ?? 所示，我们基于字幕的演变评分在所有三个评估维度中与人工评分的对齐度最高。这些结果进一步证实了我们的方法不仅提高了计算效率，还保证了卓越的评分可靠性。

为了更清晰地理解我们数据选择过程的每个阶段，我们在表 ?? 中呈现了统计细分。原始数据直接来自于 OpenVidHD-0.4M [57] 。在接下来的工作中，我们将使用我们的数据选择管道处理更多现有的高质量文本-视频数据集（例如，MiraData [37] ，InternVid [76] ），以促进未来的研究。

数据样本演示 真实世界的视频是物理现象的直接反映。在此，我们展示了一些我们选择的视频作为说明性例子，如图 9 所示。这些例子突出了多样的物理场景，为我们数据集中捕获的现象类型提供了一个参考。

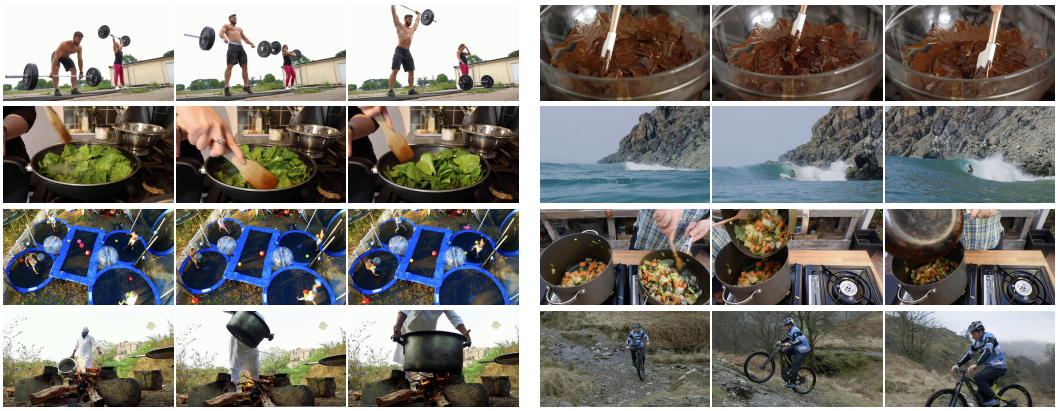


Figure 9: 从我们选择的数据集中随机抽取的数据示例。

C 更多实验设置和分析

C.1 实验设置的更多细节

我们使用两个广泛采用的物理专注基准来评估生成视频的物理保真度：VideoPhy [ICLR'25] [6] 和 PhyGenBench [ICML'25] [54] 。根据 WISA [75] 中介绍的评估协议，我们利用 VideoPhy [6] 中的 VideoCon-Physics 来评估生成视频的两个关键指标：物理定律一致性（PC）和语义连贯性（SA）。具体而言，我们使用来自 VideoPhy 的 344 个精心策划的提示和 PhyGenBench 的 160 个提示，这些提示均旨在反映多样的物理原则和场景。

为了量化性能，我们遵循各个基准论文中列出的评分标准。对于 VideoPhy 和 PhyGenBench 来说，我们把 PC 和 SA 值大于或等于 0.5 的情况视为 PC = 1 和 SA = 1，而把值小于 0.5 的情况视为 PC = 0 和 SA = 0。具体来说：

- 对于 VideoPhy，我们报告生成视频中同时满足 PC = 1 和 SA = 1 的比例，这反映出其与物理定律和语义意图的同时一致性。
- 对于 PhyGenBench，我们专注于物理常识对齐（PCA）得分，该得分在 PC = 1 的条件下评估物理推理的一致性。

此评估框架确保了与以往工作的一致性，并提供了对于生成视频的物理现实性和语义连贯性的稳健评估。

受有趣视频 [ICML'25] [5] 的启发，我们进一步使用来自 IPV-Txt [5] 的提示来评估 PhysHPO 是否超越了仅仅拟合固定的物理现象，并展示出更强的泛化能力和鲁棒性。结果如图 4 （右）和图 5 所示。

对于图 4 （右）中显示的定量比较，我们遵循 [5] 中的评估协议，基于两个关键指标评估生成的视频：

- 视觉质量（VQ）：该指标通过组合来自 VBench [34] 的六个因素得出，包括主体一致性、背景一致性、运动平滑度、美学质量、成像质量和动态程度。这些因素被聚合为一个单一的分数，以反映生成视频的整体视觉质量。
- 不可能提示跟随 (IPF)：该指标评估生成的视频与这些不可能提示的语义意图的对齐程度。根据 [5] ，我们利用 GPT-4o 为每个视频提供基于提示遵循性进行的二值判断，并计算积极判断的比例作为最终的 IPF 分数。

对于图 5 中呈现的定性比较，我们在此进一步详细说明提示：

- 左边："一张纸在木桌上神秘地变成了蒸汽。这个超现实的情景以某人将纸放下开始，接着是人手的一次轻触，触发了意外反应——坚实的纸张瞬间蒸发为丝丝白色蒸汽，并消散在空气中。"
- 右图："在这段逼真的画面中，一个陶瓷杯神秘地从平坦的桌面上凭空消失。没有任何可见的原因，突然的消失以一种不可思议的方式违背了物理学。这一场景以清晰、逼真的细节和自然光线被捕捉下来。"

我们进行了一项用户研究，使用平均意见评分（MOS）和直接成对比较来评估人类偏好。具体而言，我们设计了一个用户友好的界面，以便于评估过程并从总共 15 名志愿者中收集反馈。提供给参与者的详细指示如下所示。

User Study: Physically Plausible Video Generation

Thank you for participating in our user study! Please follow these steps to complete your evaluation:

1. Video Generation: Carefully read the target prompt provided, and then view the provided videos.
2. Scoring Criteria: Assign a score to each generated video based on the following aspects (1 being the lowest, 5 being the highest):
   - 整体偏好：对生成视频的整体满意度进行全面评价，包括语义、物理推理、视觉质量和运动质量等方面。
   - 语义贴合度：视频与输入的语义描述或文本指令的吻合程度。
   - 物理常识：视频内容是否符合基本的物理常识，比如运动定律、物体交互和逻辑行为。
   - 视觉质量：视频的视觉外观，包括分辨率、清晰度、色彩表现和纹理细节。
   - 运动质量：视频中运动的平滑性和自然性，包括物体或角色的轨迹和速度变化。
3. Submission: Click the "Submit Scores" button to submit your scores.

Notations:
1. We observe that the edge browser is not fully compatible with our interface. Chrome is recommended.
2. Remember to click the "Submit Scores" button after your evaluation.
3. If you see that videos and the score sliders are not aligned, shrinking your page usually works.
4. If the video seems to be stuck, usually waiting for a few seconds will sovle this.
5. If the page is not responsive for a long time, please try to refresh it.
6. If you have any questions, please directly contact us. Thank you for your time and effort!

C.2 更多分析

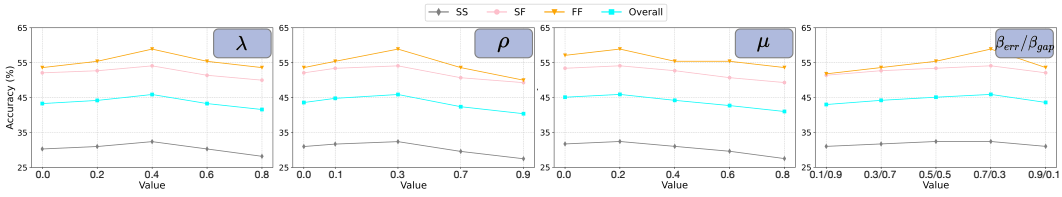


Figure 10: 在 VideoPhy [6] 上 PhysHPO 的超参数分析。

超参数分析 图 10 详细分析了各种超参数对 PhysHPO 在 VideoPhy [6] 基准上性能的影响。a）第一张图检验了状态级损失权重（λ），显示准确性随着 λ 增加而提高，在 $\lambda = 0.4$ 达到顶峰，然后下降。这表明适度强调状态级一致性可以提高性能，而过多的权重可能导致对低级细节的过拟合。b）同样，第二张图中的运动级损失权重（ρ）展示了类似的趋势，性能在 $\rho = 0.3$ 达到最佳。这突显了捕捉运动动态的重要性，但也强调了过分强调这一方面的潜在危害。c）第三张图探讨了语义级损失权重（μ），准确性随着 μ 增加而提高，直至 $\mu = 0.2$ ，

然后表现趋于平稳或略微下降。这强调了对齐语义级信息以确保连贯且上下文准确的视频生成的必要性。d ）最后，第四张图分析了实例级负采样中错误（β_{err} ）和空隙（β_{gap} ）样本之间的平衡。结果表明，权重（$\beta_{\text{err}}/\beta_{\text{gap}} = 0.7/0.3$ ）达到最高准确性，因为这两种样本类型都有助于模型的稳健性。过分强调错误（$\beta_{\text{err}}/\beta_{\text{gap}} = 0.9/0.1$ ）或空隙样本（$0.1/0.9$ ）会降低性能，这可能是由于训练信号失衡所致。

非首选样本分析 非首选样本的选择决定了它们与首选样本的对比，并显著影响模型对偏好学习的关注。在此，我们进一步探讨在实例、状态和运动层面选择非首选样本的策略，同时确保与主文本中对应的首选样本配置保持一致。具体来说，i) 实例层面：对应于一个首选视频，生成的错误视频的数量，即生成多少视频以选择错误视频；ii) 状态层面：交换多少边界帧以构造非首选样本；iii) 运动层面：选择结构化表示，如深度图、光流和 Canny 边缘。

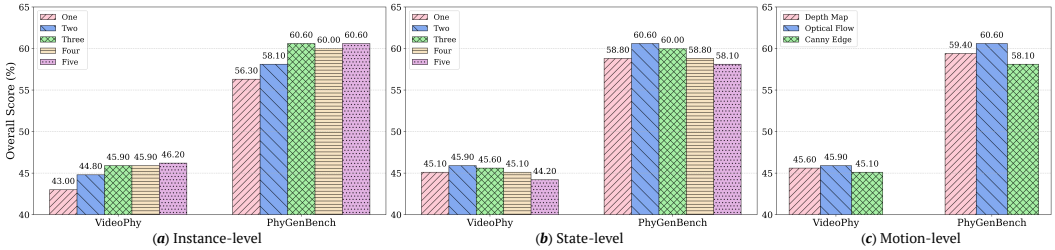


Figure 11: 非优选样本分析：(a) 实例级，(b) 状态级，和 (c) 运动级。

图 11 展示了不同非优选样本选择策略对模型性能的影响：(a) 实例级：我们观察到选择三个生成样本可以达到最佳性能。虽然增加样本数量可能会带来轻微的性能提升，但同时也会导致更高的计算成本，因此选择三个样本是一种实际的平衡。(b) 状态级：交换适量的边界帧可获得最佳结果，因为过多或不足的帧交换会削弱优选样本和非优选样本之间的对比。(c) 运动级：光流在整体得分上达到最高，显示出其在捕捉运动动态方面的有效性。相比之下，深度图和 Canny 边缘表示的性能相对较低，这可能是由于其运动信息不够详细。这些结果强调了在选择非优选样本时仔细平衡计算效率和性能的重要性。每个级别上的适当配置确保了模型性能的有效性和稳健性。

D 展览板

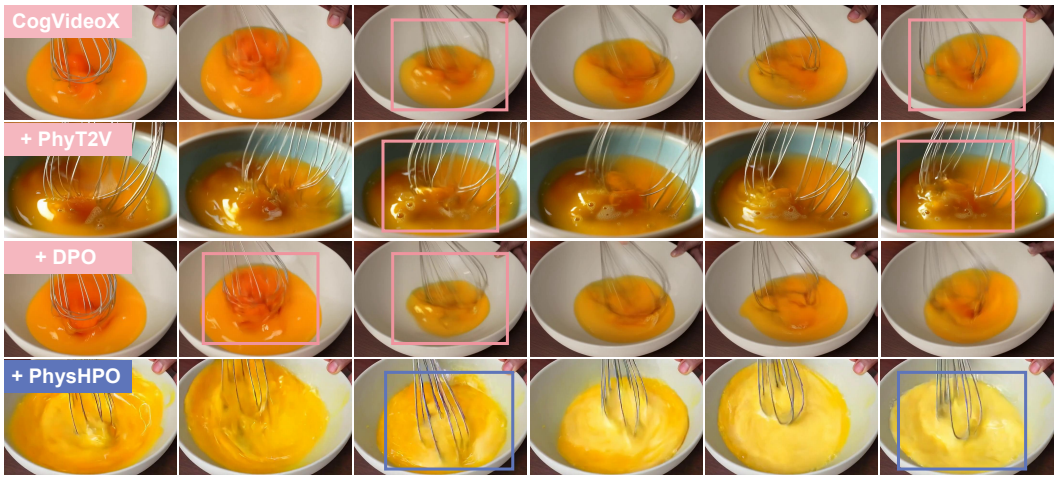
在这里我们提供更多比较结果：图 12 （与基线相比）和图 13 （与物理聚焦提示相比），以及关于人类动作/运动的结果：图 14 （HunyuanVideo [39] ）和图 15 （CogVideoX [90] ）。我们强烈推荐观看 HTML 演示和我们的附录。

E 局限性和未来工作

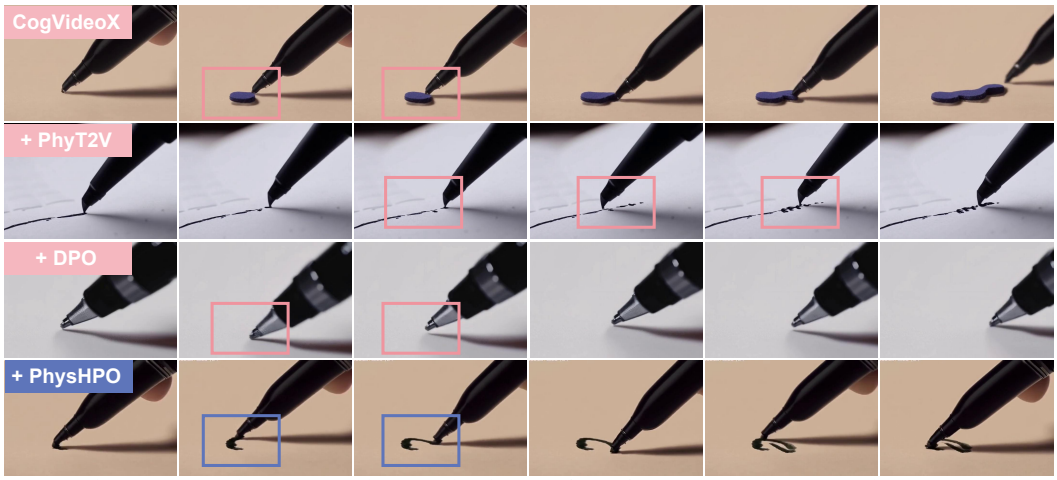
虽然 PhysHPO 在视频对齐上具有细粒度的精确性，但训练大规模视频生成模型所需的巨大的计算成本仍然是一个潜在的限制，尤其是对于资源有限的个人和组织而言。未来的研究应探索更轻量级的架构（例如，FramePack [96] ）或进一步探索测试时偏好对齐，以在减少计算开销的同时实现更大的性能提升。此外，本文引入的数据选择策略为了实用性而被设计得相对简单。未来的工作应根据特定的任务需求对其进行进一步优化。

F 伦理影响

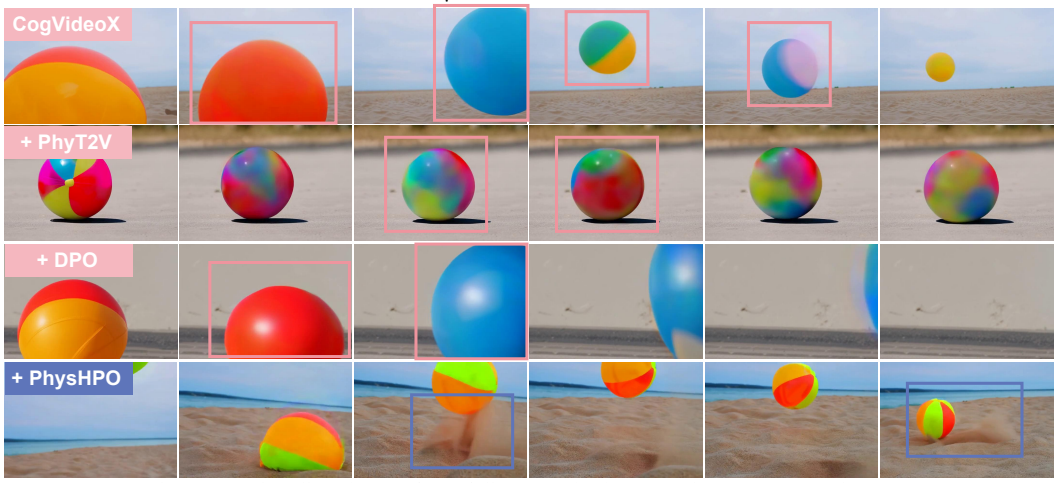
PhysHPO 被开发为一种仅用于研究的分层偏好优化策略。它仍可能引发重要的伦理考虑，特别是在内容生成方面。生成高质量视频的能力可能被滥用来制作误导或有害的内容。为减轻此风险，我们建议在生成的视频中加入如水印等的保护措施，以确保透明性和真实性。此外，还应制定负责任使用的指南，强调其在伦理和创造性背景下的应用，比如在教育、艺术或研究场景中，同时阻止其用于欺骗性或有害目的。



"A whisk mixes an egg in a bowl."



"A black pen is used to write on the smooth, white surface of a notebook, showcasing the interaction between the pen and the notebook surface."



"A vibrant, elastic beach ball is thrown forcefully towards the ground, capturing its dynamic interaction with the surface upon impact."

Figure 12: 更多与基线的比较展示。

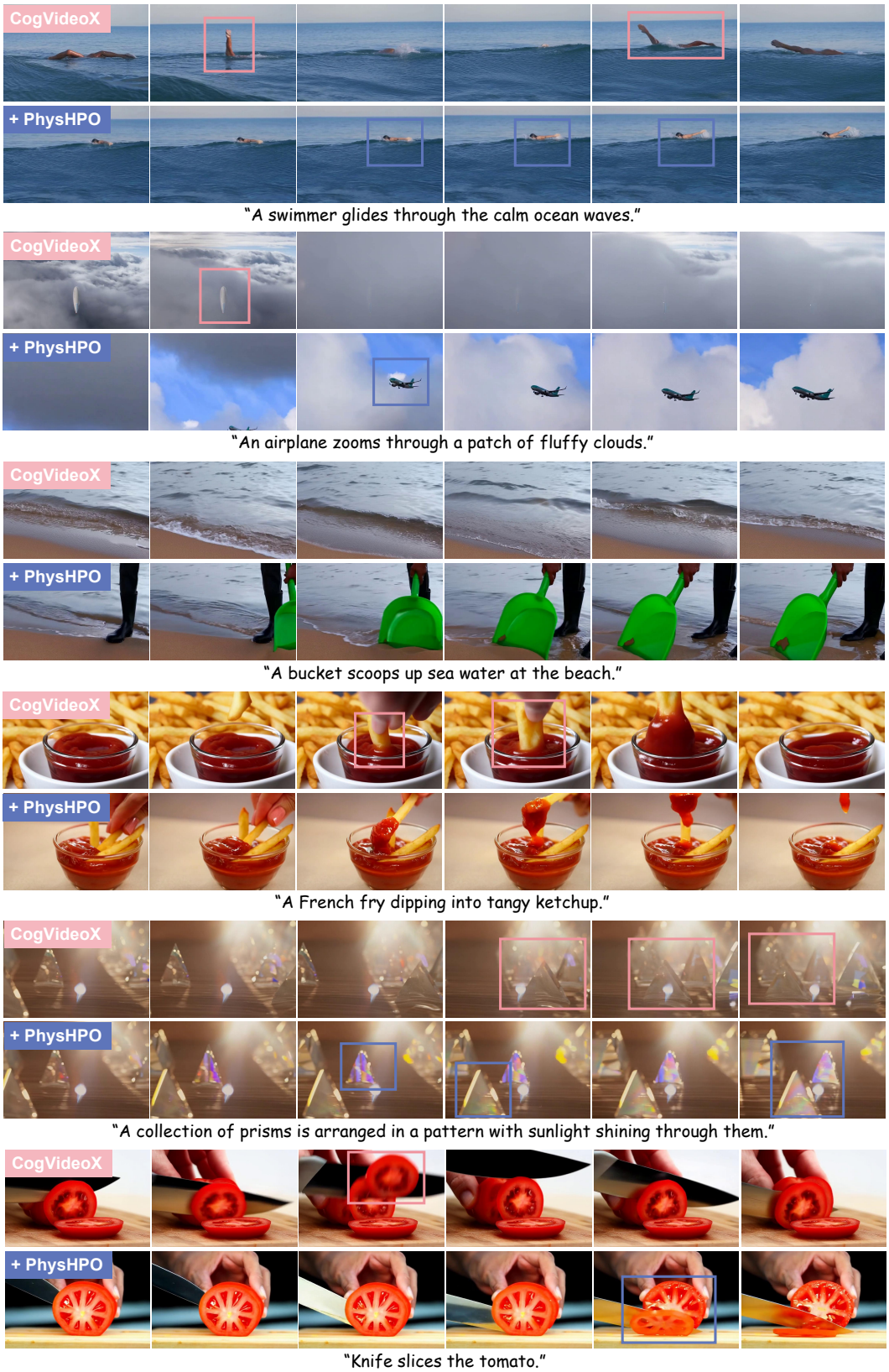


Figure 13: 更多的结果演示来自 VideoPhy [6] 的提示。

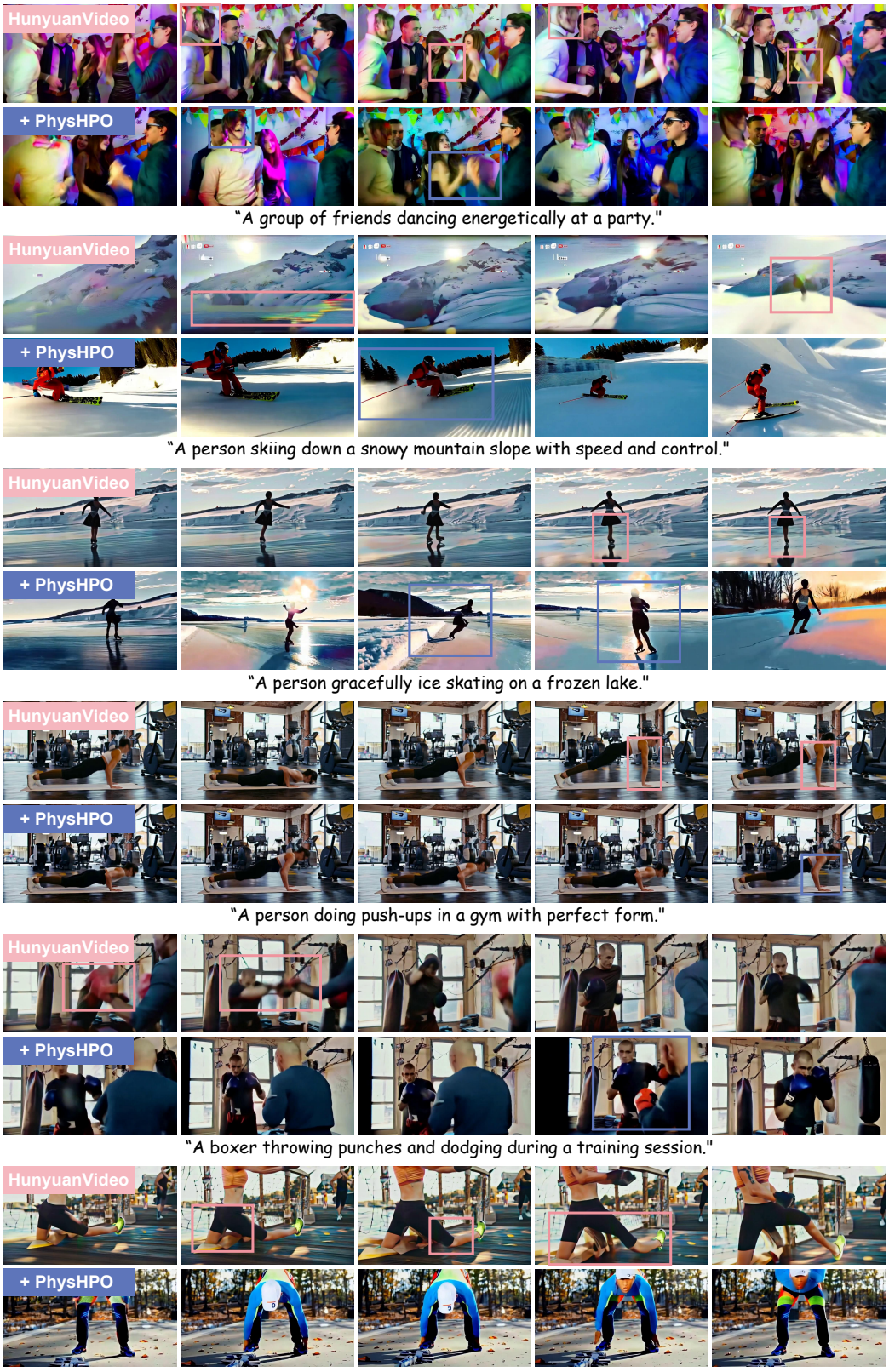


Figure 14: 更多展示结果与人类动作/运动为重点的提示。

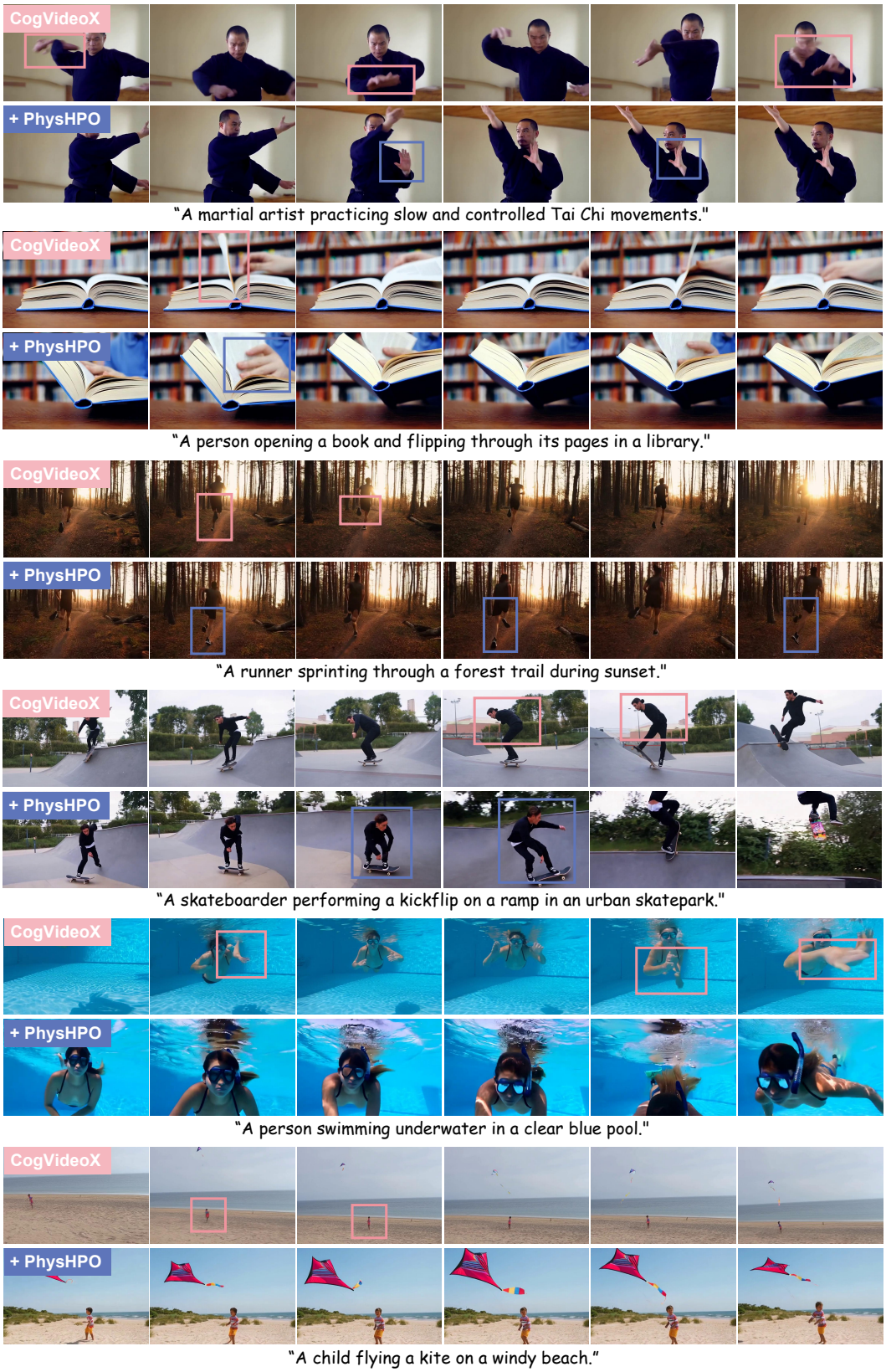


Figure 15: 更多关于人类动作/运动为重点的提示的结果演示。