T2VWorldBench: A Benchmark for Evaluating World Knowledge in Text-to-Video Generation

Yubin Chen* Xuyang Guo† Zhenmei Shi‡ Zhao Song§ Jiahao Zhang¶

Abstract

文本到视频（T2V）模型在生成视觉上合理的场景方面表现出色，但其利用世界知识以确保语义一致性和事实准确性的能力仍未被充分研究。针对这一挑战，我们提出了 **T2VWorldBench**，这是第一个系统性的评估框架，用于评估文本到视频模型的世界知识生成能力，涵盖了 6 个主要类别、60 个子类别和 1200 个提示，涉及广泛领域，包括物理、自然、活动、文化、因果关系和物品。为了兼顾人类偏好和可扩展的评估，我们的基准结合了人类评估和使用视觉语言模型（VLMs）的自动评估。我们评估了目前 10 个最先进的文本到视频模型，从开源到商业模型，发现大多数模型无法理解世界知识并生成真正正确的视频。这些发现指出了当前文本到视频模型在利用世界知识方面的一个关键差距，提供了构建具有稳健常识推理和事实生成能力的模型的宝贵研究机会和切入点。

*abinzzz1227@gmail.com. San Jose State University.
†gxy1907362699@gmail.com. Guilin University of Electronic Technology.
‡zhmeishi@cs.wisc.edu. University of Wisconsin-Madison.
§magic.linuxkde@gmail.com. University of California, Berkeley.
¶ml.jiahaozhang02@gmail.com.

1 简介

最近生成模型的进展在多个方面显著提升了文本到视频（T2V）模型的性能，包括视频编辑、运动一致性和对象一致性，这促进了视频生成的探索。文本到视频模型并不生成静态图像，而是对真实物理世界进行建模，以创造出高度美观和逼真的视频。目前，先进的 T2V 模型如 Sora 和 Kling 能够根据用户提示生成符合物理定律的逼真视频。这些惊人的视频生成技术已经极大地改变了我们与视频互动的方式，让即使是业余爱好者也能以导演般的精确度创造出电影级的场景，并在公众和研究领域内获得广泛关注。

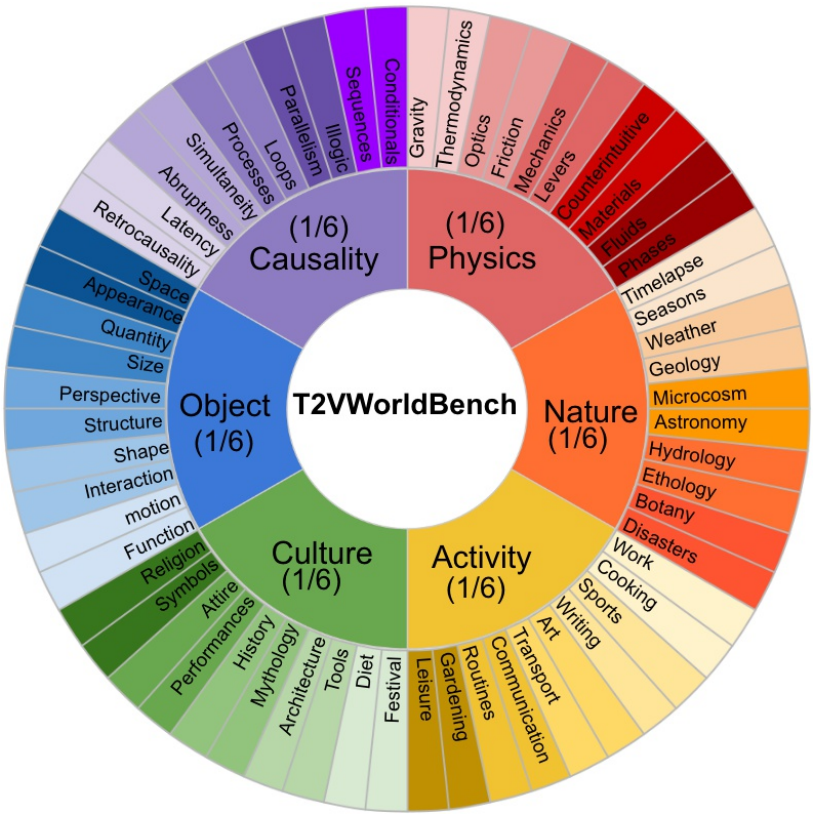

Figure 1: T2VWorldBench 的提示领域分布

尽管文本到视频的模型在语义理解和视频质量方面取得了突破 [HHY+24, LCL+24, ZHL+25]，但存在一个关键限制：大多数当前的文本到视频模型在虚构提示下表现良好，但未能探索模型根据世界知识生成视频的能力。尽管最近的研究已经调查了 T2I 模型基于世界知识生成图像的能力 [ZJX+25, NNZ+25]，但在文本到视频领域中相应的能力仍然研究不足。真正智能的视频生成需要对现实世界现象、因果关系和常识推理有深入理解，而不仅仅是像素操作 [SVC+24, YWPH+24]。例如，当为教育用途生成种子发芽的视频时，重要的是要直观地描绘一个连续的过程，其中胚根突破种皮并向上生长，直到绿叶最终长出，而不是仅展示种子突然变成幼苗。因此，迫切需要专门的研究和全面的评估来测试 T2V 模型对世界知识的总体理解和基于该知识进行推理的能力。

为此，我们提出了 T2VWorldBench，这是一个旨在评估 T2V 模型世界知识能力的综合基准。该基准包括来自六个主要类别的 1,200 个文本提示（完整的分类请参见图 1）。这些提示用于评估 10 个最先进的 T2V 模型，涵盖商业和开源系统，并反映截至 2025 年的文本转视频生成最新进展。为了在可扩展评估与人类偏好之间取得平衡，我们采用了一种混合评估协议，其中在视频质量、视频现实性、视频相关性和视频一致性这四个相同的标准上同时进行自动和人工评估。对于自动评估，人工注释员首先为每个提示提供基于现实常识知识的详细解释。然后，我们使用视觉语言模型

(VLMs) [LBPL19, LDF+20, ZYX+23] 评估生成的视频是否与这些期望一致，从而实现一种值得信赖和了解知识的评估，其超越了简单的质量指标。对于人工评估，多名注释员独立地逐帧审查每个生成的视频，并根据相同的四个标准分配分数。我们的主要贡献详述如下：

- 据我们所知，我们首次引入了一个基于世界知识的文本到视频基准，具有六个类别，包含 1200 个提示，涵盖物理、自然、活动、文化、因果关系和物体。
- 通过视觉模型评估和人工评估，我们从视频质量、视频真实感、视频相关性和视频一致性四个方面评估文本到视频的模型。我们发现，当前的文本到视频模型在基于世界知识的视频生成方面表现不佳，整体得分通常低于 0.70。

我们在第 2 节系统地回顾了此基准的相关工作。第 3 节给出了 T2VWorldBench 基准的详细描述。我们在第 4 节报告了我们的评估框架的主要评估结果。第 5 节给出了本文的若干总结性意见。

Model Name	Year	Organization	# Params	Open
Sora [Ope24]	2024	OpenAI	N/A	No
Mochi-1 [Gen24]	2024	Genmo	10B	Yes
PixVerse V4.5 [AIS25]	2024	AISphere	N/A	No
Kling [Kli24]	2024	Kuai	N/A	No
Dreamina [Byt24]	2024	ByteDance	N/A	No
Qingying [Zhi24]	2024	Zhipu	5B	Yes
LTX Video [HCB+24]	2024	Lightricks	2B	Yes
Pika 2.2 [Pik24]	2025	Pika Labs	N/A	No
Hailuo [Min25]	2025	MiniMax	N/A	No
Wan 2.1 [Ali25]	2025	Alibaba	14B	Yes

Table 1: 我们基准测试中评估的 10 个文本到视频模型概述。

2 相关工作

X 最近，基于文本生成图像模型的成功，扩散模型在文本生成视频（T2V）方面取得了显著进展 [BRL+23, CXH+23, LCZ+23, SPH+23] 。早期的文本生成视频工作主要集中在 GAN [GPAM+14, RMC16, KLA19] 和 VAE [KW14, RMW14, HMP+17] 上，受到模型泛化能力和语义理解能力的限制。如今，通过大规模数据的训练，T2V 模型可以生成真实且视觉上吸引人的视频 [YWL+23, WY24, OJK+24, NXZ+25] ，例如 Sora [Ope24] 通过其相似的扩散变换架构，将扩散模型的生成能力与变换的序列建模能力相结合，基于大规模的预训练生成符合人类审美的空间和时间序列真实视频。同样地，Kling [Kli24] 将物理建模、可控相机系统和高效的扩散架构集成在一起，使得模型能够生成令人信服且清晰的视频，同时确保语义的一致性。这些 T2V 模型在生成具有高视觉质量、语义一致性和场景多样性的视频方面展示了令人印象深刻的能力 [GZH+23, YTZ+24] 。然而，当前的 T2V 模型在将世界知识融入视频生成方面表现出局限性 [SPH+23, CXL+24, CWL+24] ，这也是我们的主要动机之一。

随着 T2V 模型的发展加速，在各个领域测试 T2V 模型的性能变得越来越重要。这使我们能够探索此类生成模型的基本限制，并指出许多未来的方向，如高阶流匹配、惰性传播和理论保证。最初，使用 Inception Score (IS)、Fréchet inception distance (FID) 和 Fréchet Video Distance (FVD) 作为评价视频质量的指标。为了评估语义一致性，引入了 CLIPScore 作为指标，通过利用 CLIP 模型来评价文本提示和生成视频的相似性。尽管早期的评估指标在低级感知和静态语义对齐方面表

现良好，但它们在捕捉时间一致性、物理建模和细粒度解释上仍面临挑战。为改善 T2V 评价，提出了几个新的基准，包括全面评价、数值约束、动态一致性、细颗粒度评价、多属性结合、物理原理约束。具体而言，[JXTH24] 引入了一个时间动态基准，对 16 个关键的时间维度进行分层评价，包括多个评价指标如 CLIPScore，BLIPScore 和 VQA Score。[HHY+24] 提出了一种全面的评测基准，通过多维度、人类对齐和丰富见解的特性来全面评估 T2V 模型。尽管之前的基准在评估 T2V 模型能力的几个方面表现出了成效，但大多数主要关注文字提示的语义对齐，这忽略了对文本和世界知识更深层次的整合。为了应对这一挑战，最近的研究开始探索能够测试模型整合和根据世界知识推理能力的基准 [MLT+24, NNZ+25, ZJX+25] 。然而，这些基准主要评估 T2I 模型。相比之下，文本到视频模型中世界知识的整合和推理没有被充分重视，这是我们工作的主要动机。

3 基准测试

我们在本节中介绍我们研究中提出的基准——T2VWorldBench。第 3.1 节描述了基线模型。第 3.2 节提供基准提示。之后，第 3.3 节展示了我们基准的评估协议。

3.1 基线模型

在我们的工作中，我们对 2024 年至 2025 年间发布的十个最先进的文本到视频生成模型进行了综合评估，其中包括商业和开源系统。这个选择确保我们的工作反映了 T2V 模型的最新进展，同时揭示了目前 T2V 模型在整合和推理世界知识方面面临的挑战。模型的详细信息在表 1 中。

为确保一致性，视频以每个 T2V 模型的最低可用分辨率生成，通常为 720p。所有视频均被限制为 16:9 的宽高比，并限制在大约 5 秒。实施细节在附录 A 中提供。

3.2 基准提示

为了全面评估当前 T2V 模型整合和推理世界知识的能力，提示需要设计得超越字面意义。具体来说，这些提示应挑战 T2V 模型对隐性知识的掌握，测试其推理能力，并探讨其对现实世界物理定律和客观事实的理解。例如，与“一个男人在走路”这样的简单提示不同，“一个人在走路时踩到香蕉皮”这一提示超越了字面描述，要求模型能够推断出香蕉皮的滑溜特性，预测由此产生的失衡，并生成符合真实物理法则和因果逻辑的连贯视频。在我们的工作中，我们精心构建了 T2VWorldBench，它包括了 6 个知识领域：物理、自然、活动、文化、因果关系和物体。每个领域包含 10 个细分领域，每个细分领域包含 20 个提示，共计 1,200 个精心策划的提示。T2VWorldBench 的详细组成如图 1 所示。图 2 显示了具有代表性的提示和生成的视频结果。

在使用 VLMs 自动评估世界知识时的关键挑战在于，VLMs 自身可能不具备准确的世界知识，从而使它们的评估可能不可靠。为了解决这个问题，我们为每个视频提供了人工撰写的解释，详细说明所需的世界知识和推理链。这些解释不仅旨在阐明每个提示背后的隐含世界知识和推理链，还作为参考点，用于与 T2V 模型的生成结果进行比较。例如，在类似“一个足球运动员在比赛中罚点球”的提示中，相应的解释会具体说明视频显示的是一个足球运动员在球场上，把球放在禁区内的点球点上，退后几步，助跑，然后将球踢向球门，而守门员试图扑球。

物理学。 T2VWorldBench 的物理领域评估 T2V 模型理解和应用基本物理原理（例如重力、运动和力）以生成视频的能力。我们的目标是调查模型是否真正理解现实世界中的物理原理，而不仅仅是基于大量训练数据中提示的字面意思对视频进行逼近的拟合。例如，T2V 模型应能够理解水到冰的状态变化、玻璃和粘土的材料，以及现实世界中的基本物理定律（如牛顿定律）。

自然。 T2VWorldBench 的自然领域旨在探索 T2V 模型对支配生物和非生物世界的自然法则和规律的理解。我们的自然领域由 10 个子领域组成，如季节、天气和地质。通过生成与自然现象相关的视频，我们试图调查 T2V 模型是否理解自然现象的原因和机制，而不仅仅是基于提示生成具有表面意义的视频。例如，当提示为“向日葵从种子发芽、生长、开花到结籽的整个生命周期”

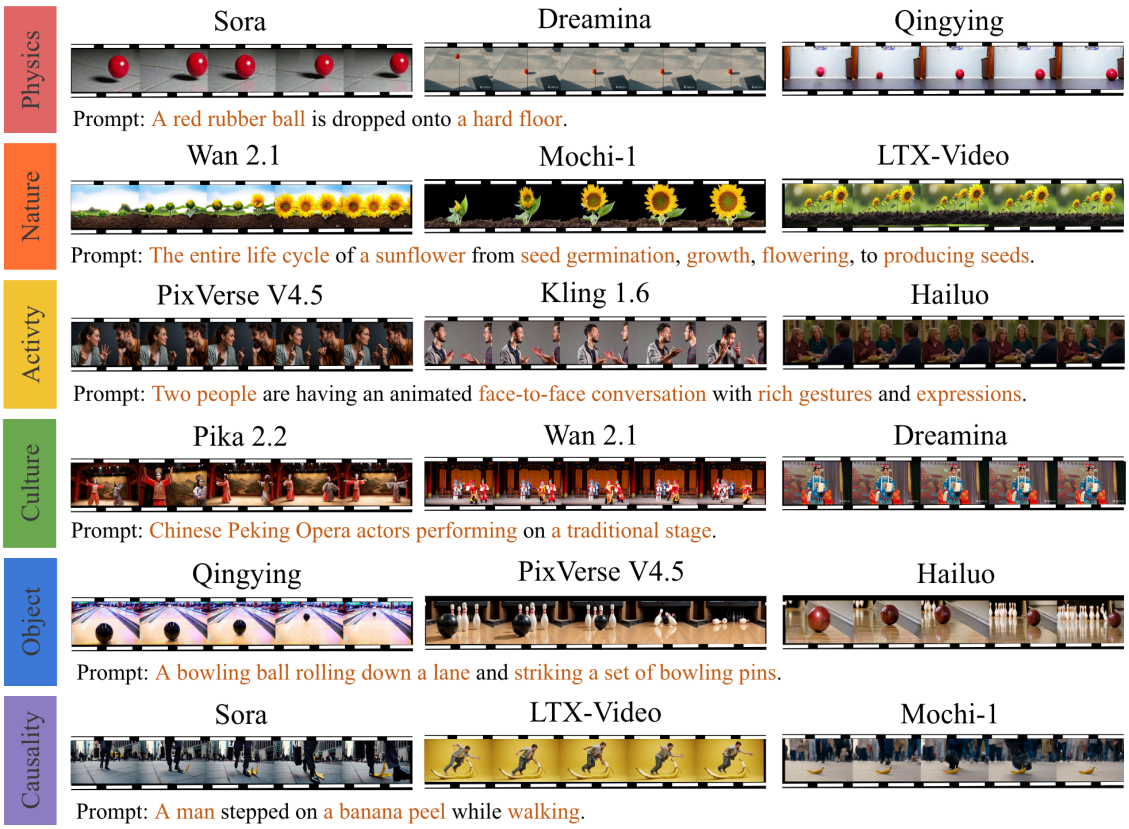

Figure 2: 使用来自不同文本到视频模型的视频示例来解释此基准中评估的所有 6 个领域。

时，一个理想的文本到视频模型应该能够理解并以视觉方式呈现植物发育的生物过程，准确捕捉从发芽、茎生长到开花的每个阶段，并在生成的视频中体现视觉现实和语义一致性。

活动。我们与世界的互动以及理解世界的方式与活动深刻相关，T2VWorldBench 的活动领域评估 T2V 模型理解和生成涉及活动的连贯序列的能力，这些活动范围从烹饪等日常行动到更结构化的运动。活动领域提出了特殊的挑战，因为它要求模型捕捉物体交互的时间动态和连续行为。仅仅识别动作的表面层次是不够的；T2V 模型必须理解动作如何随时间展开，并在视频帧之间保持一致性。对于提示“运动员在跑道上奔跑。”一个合理的生成应该描绘出运动员开始移动的动态序列视频，保持一致的跑步姿势，并顺利地沿着跑道前进。

文化。文化领域侧重于评估 T2V 模型对不同文化的理解。不同地区文化习俗的显著差异对 T2V 模型带来了巨大挑战。为了准确生成与文化相关的提示，模型不仅需要广泛的世界知识，还需要深入了解各地文化的细节、习俗及社会方面。例如，当面临诸如“在印度庆祝排灯节”和“中国京剧表演”等提示时，T2V 模型必须识别相关场景元素、手势、动作及其背后的文化意义，以便生成符合客观事实的文化主题视频。

因果关系。 T2VWorldBench 的因果领域旨在评估 T2V 模型理解和生成在时间上和逻辑上连贯的序列的能力，其中事件之间通过因果关系相连接。该领域的核心挑战在于模型是否能够识别单一动作并模拟一个动作如何触发一系列随时间展开的可观察后果。对于像“一个人打翻了玻璃杯，水洒到了桌子上”这样的提示，T2V 模型必须准确捕捉时间序列的连续性，并根据提示推理过程的视觉后果，从打翻杯子的场景开始，并连贯地生成水洒出的场景。

最后，T2VWorldBench 具有一个对象域，用于评估 T2V 模型是否能超越基本的对象识别，展示对视觉和物理对象属性的更深刻理解，包括空间、外观、数量、大小、视角、结构、形状、交互运动和功能。例如，当给出提示“一个人用剪刀剪一张纸”时，一个理解世界知识的文本到视频模型不仅应该识别个体对象（人、剪刀和纸），还应该理解它们的形式特征和功能交互。这包括正确

描绘人握持剪刀的方式、剪刀的机械运动、剪纸过程中纸的变形和断裂，并在整个剪切过程中保持时间上的连贯性，这些共同代表了模型整合世界知识来建模对象属性、用途及其交互逻辑的能力。

3.3 评估协议

以前的基准评估使用单一指标 [UVSK+19] 来对模型生成结果进行评分，这被认为是评估 T2V 模型生成结果的有效和直观的方法。然而，这样的单一评估指标过于简化了对 T2V 模型能力的评估，可能无法全面展示 T2V 模型在多个领域的具体能力。在我们的工作中，我们引入了四个评价维度：视频质量、视频真实感、视频相关性和视频一致性。

为了确保评估的稳健性和可靠性，我们采用了人工与自动评估相结合的评估策略。我们对人工和自动评估协议进行了详细介绍如下：

人工评估。

为了更好地符合人类偏好并获得出色的评估结果，我们在评估过程中引入了两位独立的人类标注者，他们在 AI 领域具有专业知识并且具有正常的视觉能力（没有残疾）。这些人类标注者从多个维度独立评估生成的结果，包括视频质量、视频真实性、视频相关性和视频一致性。对于每个评估维度，我们提供五个等级的评分，随后归一化为 [0, 1] 范围内的分数：

- **Level 1** （评分 0.2）：较差，视频质量非常低，存在严重的运动伪影，如模糊和拖影，缺乏视觉真实性，并且不是根据提示生成的。视频帧明显不一致，导致内容不连贯且难以理解。
- **Level 2** （评分 0.4）：一般，视频有明显的质量问题，如焦点模糊和画面不清晰。视觉效果有些人工化，缺乏令人信服的真实性。虽然生成的视频与提示相关，但联系较弱。帧之间的过渡不够连续。
- **Level 3** （评分 0.6）：可接受，视频达到了基本的质量水平，内容大体上可辨识，视觉缺陷可以接受，视频大致抓住了提示的主要意思，尽管有些细节可能不完整。视频的真实性可以接受，大体上符合人工偏好。视频的连贯性总体上可以接受，但仍有一些不连贯的场景。
- **Level 4** （评分 0.8）：良好，视频画面清晰，仅有少量轻微的视觉瑕疵，并且很好地与提示对齐，生成的视频符合世界知识。视频保持了可信的真实性，帧之间的进展大多流畅，创造了连贯且吸引人的视频。
- **Level 5** （评分 1.0）：优秀，视频以生动、高质量的视觉效果和对细节的高度关注而突出。不仅准确匹配了提示，还通过逼真且富有表现力的内容传达了其隐含的世界知识。每一帧自然过渡到下一帧，反映了时间上的高度一致性。

自动评估。

为了支持文本到视频模型的可扩展和高效评估，并从之前的优秀工作中学习自动评估技术 [HHY+24, SHL+25] ，我们引入了当前表现优异的多模态 LLaVA1.6-34B [LLLL24] 模型用于自动化评价，以充分利用其在视觉与语言融合理解方面的优势，实现对生成视频内容的准确分析和评分，并增强评估过程的自动化和一致性。

为了指导自动化评估，我们为每个提示构建了细粒度的参考解释。这些解释详细描述了基于世界知识和相关模式的理想视频应该是什么样子。通过将评估基于明确的、基于知识的期望，我们确保评估不仅捕捉到表面的视觉特征，还捕捉到语义一致性和与提示的相关性。重要的是，自动设置中使用的评估指标与人工评估中使用的指标保持一致，以确保两种评估方法之间的可比性和可靠性。

我们使用 VLM 的一般方法是从元提示开始，旨在指导模型对生成的视频进行结构化评估。具体来说，我们采用由基础提示和一组特定任务的提示构成的两阶段提示策略，这些提示对应于四个关键评估指标：视频质量、真实感、相关性和一致性。首先，为了评估整个视频，我们将其分割为网格，每个网格由 9 个连续的帧组成，排列成 3×3 的布局。然后，VLM 依次评估视频中的每个

You are an AI video quality evaluator. Analyze this 3×3 grid showing 9 CONSECUTIVE video frames arranged chronologically from left to right, top to bottom.

CRITICAL: These are 9 consecutive frames extracted from a continuous video sequence, NOT randomly sampled frames.

FRAME ARRANGEMENT:
- Row 1: Frame N → Frame N+1 → Frame N+2
- Row 2: Frame N+3 → Frame N+4 → Frame N+5
- Row 3: Frame N+6 → Frame N+7 → Frame N+8

GENERATION CONTEXT:
- Prompt: `{original_prompt}`
- Explanation: `{explanation}`

Figure 3: **基础提示模板**。基础提示建立了所有评估任务的基础上下文。它指示 AI 评估者其角色，定义输入结构为一个 3 × 3 的网格，由连续的九个视频帧组成，右侧有视觉示例，并提供原始提示和解释。这个提示作为建立所有特定维度评估的共同基础。

网格，并将所有网格中的最低分作为视频的最终得分。对于每个网格，评估从建立上下文的基础提示开始。该提示包括原始的文本到视频提示，以及从世界知识中得出的细粒度解释，这两者共同定义了视频的预期语义。参见图 3 以供参考。

在自动评估的第二步中，我们为评估模型提供了详细的、针对特定指标的提示，分别对应于四个关键的评估维度（视频质量、真实性、相关性和一致性）。每个提示都经过精心设计，以引导模型关注视频的特定方面，从而使自动评估更加有针对性和可解释。每个维度的自动评估提示具体如下：

- **Quality** ： 此评估提示评估生成视频的技术保真度。一个符合人类偏好的高质量视频应呈现清晰画面，且失真或噪声最小。评估模型被指示专注于低层次视觉属性，如分辨率、清晰度、明确性，以及渲染瑕疵的存在（例如毛刺、模糊或几何断裂）。

- **Realism** ： 此方面关注生成内容的视觉可信度。重点在于视频是否看起来真实且自然发生，考虑的因素包括真实的物体纹理、光照和阴影、物理交互，以及遵循常识物理的情况。目标是确保生成的场景和物体不会显得虚假、荒谬或违反物理常识。

- **Relevance** ： 此维度关注输入提示与生成视频内容之间的对齐，以及细粒度解释。评估重点在于基于世界知识提示推断出的关键实体、动作和场景是否在生成的视频中正确展示。视频应准确且全面地反映提示所传达的语义和细粒度细节，并能够完美展示提示中的相关背景知识。

- **Consistency** ： 此评估提示旨在评估视频中连续帧的时间一致性，分析生成视频中的对象是否随着时间保持其状态、位置、外观和运动连续性。高度的一致性对于传达可信的视频意义至关重要。

Stage 2 评估提示的模板在附录 B 中提供。

为了获得对 T2V 模型性能的稳健和平衡的最终评价，我们采用了结合人工评估和自动评估的混合评分协议。对于每个评估维度，即质量、真实性和相关性，我们首先通过计算评级者分配的平均分数来汇总人工注释。然后，将人工获得的分数与自动评估框架生成的相应分数相结合，以确保语义敏感性和一致性。每个维度的最终得分是通过取人工和自动评估的平均值来获得的。随后，生成视频的综合评价得分通过取四个维度的平均值来获得：

$$S_{overall} = \frac{1}{4} \sum_{d \in D} S_d, \tag{1}$$

其中 $D = \{$ Quality, Realims, Relevance, Consistency $\}$。本评价策略实施了一个结合主观人类洞察与可扩展且可重复的自动评估的综合评估过程。

4 实验

我们在本节中展示了 T2VWorldBench 的主要实验结果。我们在第 4.1 节中分析和讨论了从综合结果中得出的见解。我们在第 4.2 节中提供了正确和错误生成之间的定性比较。最后，我们在第 4.3 节中分析了人类注释者之间的差异。

4.1 总体世界知识结果

Model	Physics	Nature	Activity	Culture	Causality	Object	Avg.
Wan2.1	0.70	0.68	0.71	0.67	0.62	0.72	0.68
LTX Video	0.65	0.68	0.73	0.66	0.65	0.68	0.68
Kling1.6	0.66	0.68	0.74	0.69	0.62	0.66	0.67
Dreamina	0.63	0.68	0.69	0.68	0.63	0.69	0.67
Mochi-1	0.63	0.72	0.68	0.63	0.62	0.68	0.66
Sora	0.64	0.69	0.67	0.67	0.57	0.64	0.65
Hailuo	0.60	0.62	0.68	0.65	0.58	0.65	0.63
PixVerse V4.5	0.59	0.64	0.66	0.62	0.58	0.68	0.63
Qingying	0.57	0.57	0.63	0.64	0.56	0.68	0.61
Pika2.2	0.61	0.73	0.60	0.57	0.56	0.56	0.60

Table 2: 模型在六个维度及总体平均的表现。

每个评估维度的得分是依据第 3.3 节所示的评估协议得出的。结果呈现在表格 2 和图 4 中。

如表格 2 所示，目前的文本生成视频模型在基于需要整合世界知识的提示生成视频时仍然面临显著挑战。即使是在我们基准测试中表现最好的模型，如 Wan2.1 和 LTX Video，其表现也仅为中等水平，平均得分约为 0.68，这表明仍有很大的改进空间。这为我们带来了以下的见解：

Observation 4.1. *最先进的文本到视频（T2V）模型的整体评分并不理想，且最先进的 T2V 模型仍然远未掌握生成世界知识密集型内容的能力，突显出它们在结合世界知识推理方面存在巨大差距。*

在我们的基准测试中，所有文本到视频模型在活动和对象等评估领域中表现相对较强，这表明当前的文本到视频模型更擅长捕捉表面级别的动作和视觉对象。相比之下，因果关系和文化等评估维度的得分始终较低，这突显了对抽象推理和文化场景建模的困难。例如，Sora 和 Pika2.2 在因果关系维度上的得分都低于 0.60，强调了在处理多个因素相互作用的复杂事件时的局限性。此外，尽管一些专用的文本到视频模型，如 LTX Video 和 Wan2.1，在大多数评估维度上表现竞争力，但其他模型在不同的评估维度上表现出显著的差异性。这表明当前的文本到视频模型通常在狭义的能力上表现出色，但缺乏在不同类型提示上需要对世界知识进行细致理解的广泛鲁棒性。这些发现指向以下观察：

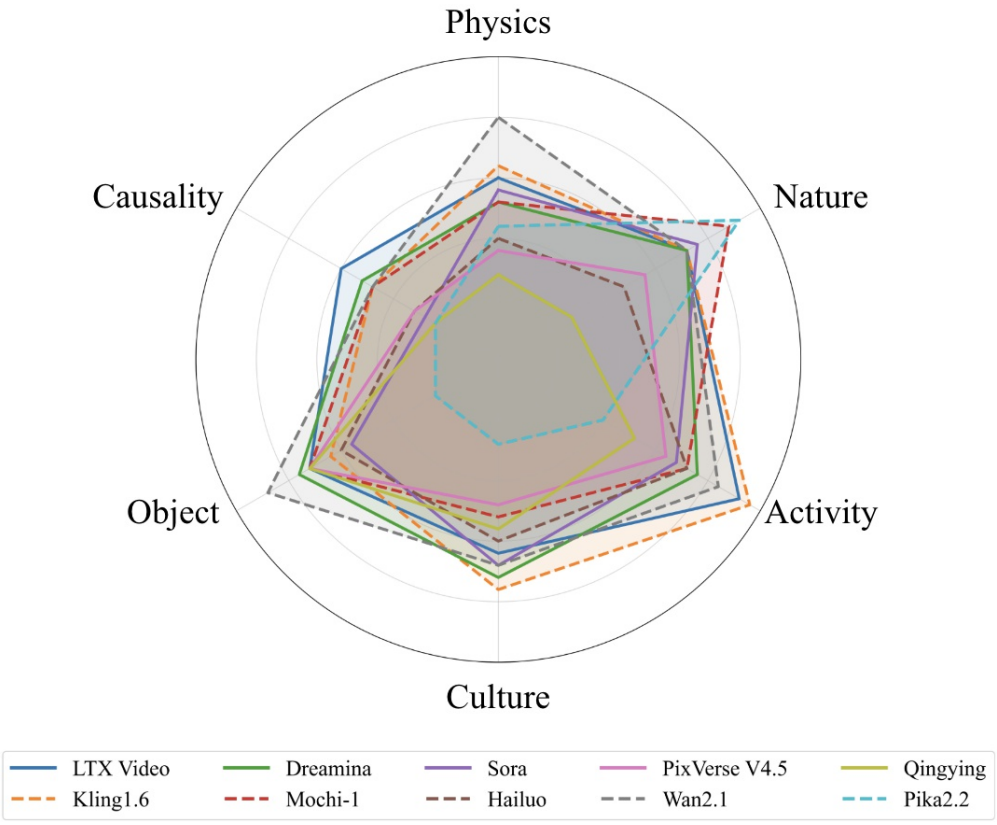

Figure 4: **模型在 6 个评估维度上的性能雷达图。**

Observation 4.2. *尽管目前的文本到视频模型在视觉层面表现得相当不错，但它们在世界知识推理方面表现不佳，并且表现出有限的泛化能力。*

4.2 定量研究

为了更好地理解文本到视频模型在面对世界知识提示时的表现，我们基于我们提出的基准 T2VWorldBench 进行了定性研究。这与现有的基准不同，后者倾向于关注诸如物体运动或视觉上的简单活动等通用场景。T2VWorldBench 要求模型基于真实世界的理解、文化和背景知识创建视频。例如，一个典型的提示可能是“东亚人在吃饭时常用的餐具类型”，这需要模型正确地将文化知识（如筷子的使用）与视觉生成关联起来。向世界知识视频合成的这种转变，使我们能够更具挑战性和意义地评估模型是否能够超越字面上的相关性，将普遍的、文化的和自然的知识融入一致的视觉输出中。

此外，我们分析了不同 T2V 模型在我们的基准测试中，基于相同的提示生成的正确和错误结果，突出当前 T2V 模型在理解和推理世界知识方面的优缺点，如附录 C 中的图 9 和图 10 所示。

对于提示“一个红色橡胶球落在硬地板上”，Mochi [Gen24] 通过生成一个描述预期运动的视频展示了准确的理解。相比之下，PixVerse [AIS25] 未能捕捉到背后的动力，而是生成了一个红色球在地板上滚动的画面，错过了提示中暗示的运动和互动。同样，对于需要文化或自然世界知识的提示，差异变得更加明显。在提示“美国总统的工作场所”时，Qingying [Zhi24] 生成了一个与这个概念大致相关的白色宫殿式结构，而 Sora [Ope24] 成功识别并展示了白宫，显示出更高水平的知识基础和运用世界知识的推理能力。在另一个例子中，提示“沙漠中最常见的带刺植物”揭示了 Kling [Kli24] 将描述与仙人掌联系的能力，展示了正确的视觉语义对齐，而 Dreamina [Byt24] 描绘了一个巨大的、带刺的、抽象的绿色植物，缺乏现实世界的合理性。在物理因果场景“一个人走

路时踩到了香蕉皮”中也可见类似的对比：Wan [Ali25] 通过现实的滑倒情节捕捉到了完整的因果序列，而 Hailuo [Min25] 仅仅展示了一个人路过香蕉皮，未能捕捉到提示中暗示的因果序列和嵌入的预期结果。这让我们得到以下见解：

Observation 4.3. 对于当前的文本到视频模型，即使它们能够在处理包含世界知识的提示时理解语义，它们在生成阶段常常存在偏差，输出的视频不符合现实或逻辑，暴露出理解和生成之间的显著差距。

4.3 人工标注者方差分析

对于每个评价维度，我们取两位标注者评分的平均值，以产生更稳定和更具代表性的评估。为了评估两位标注者之间的一致性和可靠性，我们通过公式 (2) 计算他们在所有评价维度上的得分的皮尔逊相关系数（r）。

我们为每个评价指标定义了两个注释者的分数为 $X = \{X_1, X_2, \ldots, X_n\}$ 和 $Y = \{Y_1, Y_2, \ldots, Y_n\}$，其中 n 是样本的数量。他们分数之间的皮尔逊相关系数 r 的计算如下：

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2} \cdot \sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}. \tag{2}$$

Metric	Pearson (r)	Agreement Level
Quality	0.623	Moderate agreement
Realism	0.617	Moderate agreement
Relevance	0.728	Substantial agreement
Consistency	0.758	Substantial agreement

Table 3: **皮尔逊相关系数**。此表显示了每个评估指标下两个人工注释者之间的皮尔逊相关系数，并附有相应的协议程度的定性解释。

5 结论

在我们的研究中，我们提出了 **T2VWorldBench** ，这是一个创新的基准，用于系统地评估文本到视频模型理解和整合世界知识的能力，由来自 6 个维度和 60 个子维度的 1200 个提示组成。我们对 10 个文本到视频模型的评估，涵盖了商业和开源模型，揭示了在生成过程中对世界知识的有效利用方面仍然存在显著的不足。即使是迄今为止总体表现最好的文本到视频模型，也很难在复杂和依赖知识的推理场景中展示出卓越的视频生成性能。我们希望我们的工作能为未来的研究提供参考，并激发进一步的探索和改进，以增强文本到视频模型对世界知识的理解和综合能力。

Appendix

路线图。节介绍了这十个基线文本到视频模型的实现细节。在 B 节中，我们展示了评估提示模板。在 C 节中，我们展示了一系列视频示例。

A 实现细节

本节提供了关于以下所列 10 个基准文本到视频模型的详细信息：

- **Sora** [Ope24]：由 OpenAI 团队在 2024 年开发的 Sora 是一个闭源的生成模型。该模型能够生成每秒 30 帧的视频，并可选择 5、10、15 或 20 秒的持续时间。它支持一系列的输出格式，包括从 480p 到 1080p 的分辨率以及多种纵横比（1:1、16:9 和 9:16）。Sora 提供风格预设，并可从单个提示生成四种不同的视频变体。此外，还有一个“宽松模式”，每个视频的处理延迟大约为 30 秒。

- **Dreamina Video 3.0** [Byt24]：Dreamina Video 3.0 是由 ByteDance 团队在 2024 年发布的闭源生成器，支持 5 秒和 10 秒的视频。它支持多种纵横比（16:9、21:9、4:3、1:1、3:4、9:16），并利用 DeepSeek-R1 [GYZ+25] 进行提示增强。

- **Qingying** [Zhi24]：青影是智谱开源 CogVideo 模型 [HDZ+23, YTZ+24] 的商业化实施。它能够在 1:1、9:16、16:9、4:3 和 3:4 五种纵横比中以 30/60 FPS 生成 5 秒的视频。青影支持两种模式：质量和快速。此外，它提供了对视频风格、情感氛围和摄像机运动的细致控制，并支持 AI 生成的音频和视觉效果。

- **Wan2.1 Plus** [Ali25]：万 2.1 Plus 是一个阿里巴巴集团在 2025 年发布的开源生成模型 [Wan25]，支持多种纵横比（1:1、3:4、4:3、9:16、16:9）。它还提供额外功能，如“灵感模式”和“音效”。

- **Mochi-1** [Gen24]：由 Genmo 在 2024 年发布的 Mochi-1 是一个开源模型。其标准输出包括分辨率为 480p、纵横比为 16:9 的 5 秒、24 FPS 视频。Mochi-1 支持用于可重复性的种子功能，并包含随机提示建议功能。它能同时生成两个视频，每个视频的处理时间大约为 3 分钟。

- **LTX Video** [HCB+24]：由 Lightricks 于 2024 年开发，LTX Video 是一个开源模型。它生成 5 秒钟、24 帧每秒的视频，分辨率为 512p，支持 16:9、1:1 和 9:16 的宽高比。LTX Video 能够对位置、镜头类型、参考和风格进行精细控制，甚至支持旁白集成。

- **PixVerse V4.5** [AIS25]：PixVerse V4.5 是 AISphere 于 2025 年发布的一个闭源模型。它生成时长为 5 或 8 秒的视频。PixVerse V4.5 支持多种分辨率，包括 360p、540p、720p 和 1080p，并提供五种宽高比：16:9、4:3、1:1、3:4 和 9:16。

- **Kling 1.6** [Kli24]：Kling 1.6 是由快手于 2024 年发布的闭源生成模型。它生成时长为 5 或 10 秒的视频输出，支持 16:9、1:1 和 9:16 的宽高比。它具有两种生成模式：标准模式和限制的高质量模式。Kling 支持高级提示功能，包括负提示、用于重现的固定种子、提示字典和 AI 辅助提示建议。对于生成，Kling 可以从单个提示生成 4 个视频。每个视频的处理时间约为 4 分钟，最大批量为 5 个视频。

- **Hailuo 01-Director** [Min25]：Hailuo 01-Director 是由 Minimax 于 2025 年发布的一个用于文本到视频生成的闭源模型。其标准输出为 6 秒、24 帧每秒、720p 分辨率的视频，通常默认宽高比为 16:9。

- **Pika 2.2** [Pik24]：Pika2.2 是 Pika Labs 于 2025 年发布的闭源生成模型。它提供了 Pikawaps、Pikaaddition、Pikaaffects、Pikaframes 和 Pikascenes。Pika 2.2 可以生成分辨率为 720p 或 1080p 的 5 秒或 10 秒的视频，支持多种宽高比（16:9、9:16、1:1、4:5、4:3、5:2）。为了控制生成过程，它支持负面提示和种子输入。Pika 2.2 能够同时生成 4 个视频，每个视频大约需要 30 秒处理。

B 评估提示

我们采用基于提示的框架，使用 LLaVA 模型对 AI 生成的视频进行全面评估。这一方法包括将基本提示（见图 3）与针对特定评估维度的专门提示相结合。例如，通过将基本提示与质量提示（见图 5）配对来评估视频质量。类似地，视频真实感是通过基本提示加上真实感提示（见图 6）进行评估的。同样的方法应用于使用相关性提示（见图 7）评估视频相关性和使用一致性提示（见图 8）评估一致性。这确保了每项评估都在一致的背景下进行，同时允许对视频的每个不同属性进行集中、独立的评分。

YOUR TASK: Evaluate the TECHNICAL QUALITY of these consecutive frames.

Quality (1-5): Technical Excellence
Check for artifacts, resolution, clarity, color balance, and rendering quality across all consecutive frames.
- 1: Severe technical issues affecting most frames
- 2: Multiple obvious flaws impacting viewing experience
- 3: Acceptable with minor flaws
- 4: High quality with trivial imperfections
- 5: Flawless professional-grade

CRITICAL OUTPUT FORMAT REQUIREMENT:
You MUST end your response with exactly this format (no variations allowed):

Reasoning: [Your detailed analysis of technical quality]
Quality: X

Where X is ONLY a single digit from 1 to 5. Do not add brackets, extra text, or explanations after the number.

EXAMPLE:
Reasoning: The frames show good resolution and color balance with minor compression artifacts.
Quality: 4

Figure 5: **质量提示模板**。质量提示是专为评估视频质量而设计的。它指导 AI 评估者评估客观属性，如瑕疵、分辨率、清晰度和色彩平衡，使用 1 到 5 的评分来量化从严重缺陷到专业级别的视频质量。

C 视频示例

在本节中，我们提供了由我们提出的基准提示生成的视频的广泛示例。图 9 和图 10 展示了我们质量研究的结果。图 11 - 20 显示了我们基准中每个文本到视频模型的生成结果，其中从每个视频中

YOUR TASK: Evaluate the REALISM and believability of the content.

Realism (1-5): Physical Plausibility
Assess believability and natural appearance throughout the consecutive sequence.
- 1: Severe physics violations, obviously fake appearance
- 2: Multiple unnatural elements, clearly AI-generated look
- 3: Generally plausible with some artificial aspects
- 4: Very natural with minimal artificial tells
- 5: Photorealistic perfection

CRITICAL OUTPUT FORMAT REQUIREMENT:
You MUST end your response with exactly this format (no variations allowed):

Reasoning: [Your detailed analysis of realism and believability]
Realism: X

Where X is ONLY a single digit from 1 to 5. Do not add brackets, extra text, or explanations after the number.

EXAMPLE:
Reasoning: The video appears very natural with realistic physics and minimal artificial elements.
Realism: 4

Figure 6: **现实性的提示模板**。现实性提示指导视频现实性的评估。评估基于物理合理性，指导评估者识别任何物理违反、非自然元素或其他人为痕迹。1 到 5 的评分量化内容与照片般真实完美的接近程度。

提取五个具有代表性的帧，并按顺序排列形成视觉条带。这些展示的实例与第 4 节中讨论的实验设置相一致。

YOUR TASK: Evaluate how well the frames match the generation goals.

Relevance (1-5): Adherence to Goals
Compare the video sequence against the Prompt: `{original_prompt}` and Explanation: `{explanation}`.
- 1: Completely unrelated content
- 2: Weak connection, missing major elements
- 3: Captures general concept, lacks details
- 4: Accurately represents most elements
- 5: Perfect alignment with all requirements

CRITICAL OUTPUT FORMAT REQUIREMENT:
You MUST end your response with exactly this format (no variations allowed):

Reasoning: [Your detailed analysis of how well it matches the prompt and explanation]
Relevance: X

Where X is ONLY a single digit from 1 to 5. Do not add brackets, extra text, or explanations after the number.

EXAMPLE:
Reasoning: The video accurately depicts most elements from the prompt but lacks some specific details.
Relevance: 4

Figure 7: **相关性提示模板**。相关性提示专注于评估视频与用户提示和解释的相关性。该提示指导 AI 评估者将视觉输出与提供的原始提示和解释进行比较，根据生成内容对所需元素和世界知识的捕捉程度，按 1 到 5 的等级进行评分。

YOUR TASK: Evaluate the TEMPORAL CONSISTENCY between consecutive frames.

Consistency (1-5): Temporal Coherence Between Consecutive Frames
IMPORTANT: Since these are consecutive frames, analyze smooth transitions and logical progression from frame to frame.
- 1: Chaotic inconsistency - objects teleport, backgrounds change randomly, no logical flow between consecutive frames
- 2: Major temporal disruptions - significant jumps or morphing between adjacent frames, jarring transitions
- 3: Generally stable progression with some noticeable but minor temporal inconsistencies between frames
- 4: Smooth temporal flow with natural progression, only very minor variations between consecutive frames
- 5: Perfect temporal continuity - seamless, natural progression that could be from real video footage

CRITICAL OUTPUT FORMAT REQUIREMENT:
You MUST end your response with exactly this format (no variations allowed):

Reasoning: [Your detailed analysis of frame-to-frame consistency and temporal flow]
Consistency: X

Where X is ONLY a single digit from 1 to 5. Do not add brackets, extra text, or explanations after the number.

EXAMPLE:
Reasoning: The consecutive frames show smooth transitions with natural progression and minimal temporal inconsistencies.
Consistency: 4

Figure 8: **一致性提示模板**。一致性提示用于评估视频的一致性。核心任务是分析连续帧之间的连贯性，重点关注过渡的平滑性以及对象和动作的逻辑进展。1 到 5 的评分衡量了视频的时间流，从混乱和脱节到无缝和自然。

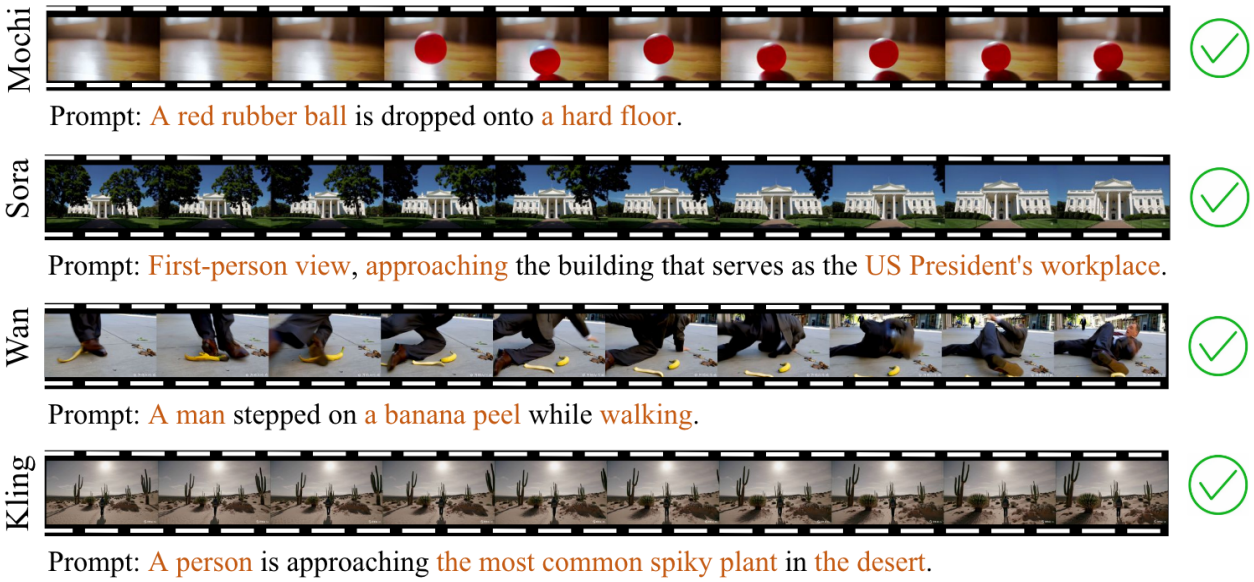

Figure 9: 成功理解世界知识的例子。

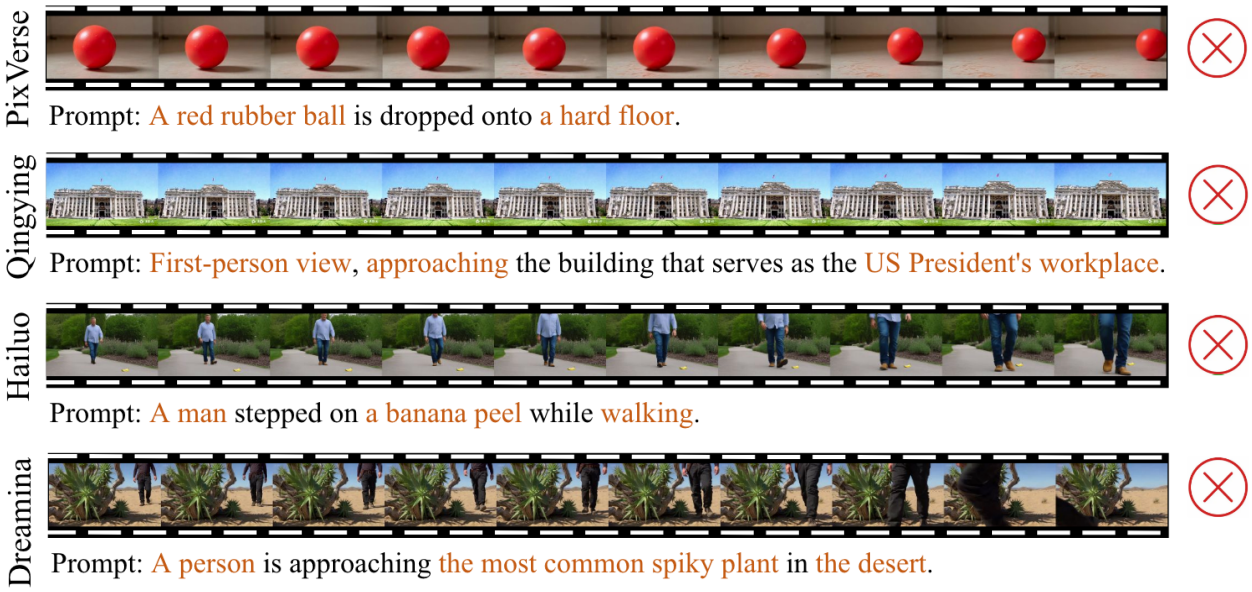

Figure 10: 理解世界知识失败的例子。

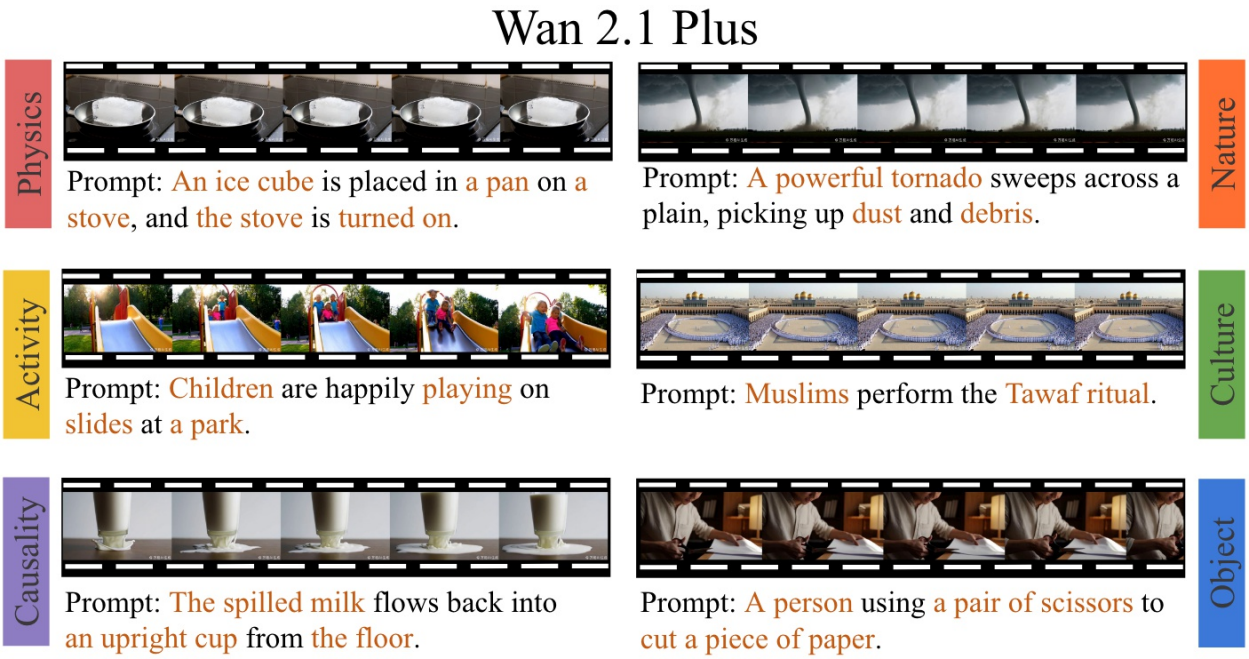

Figure 11: 智歌 2.1 Plus 的视频生成。

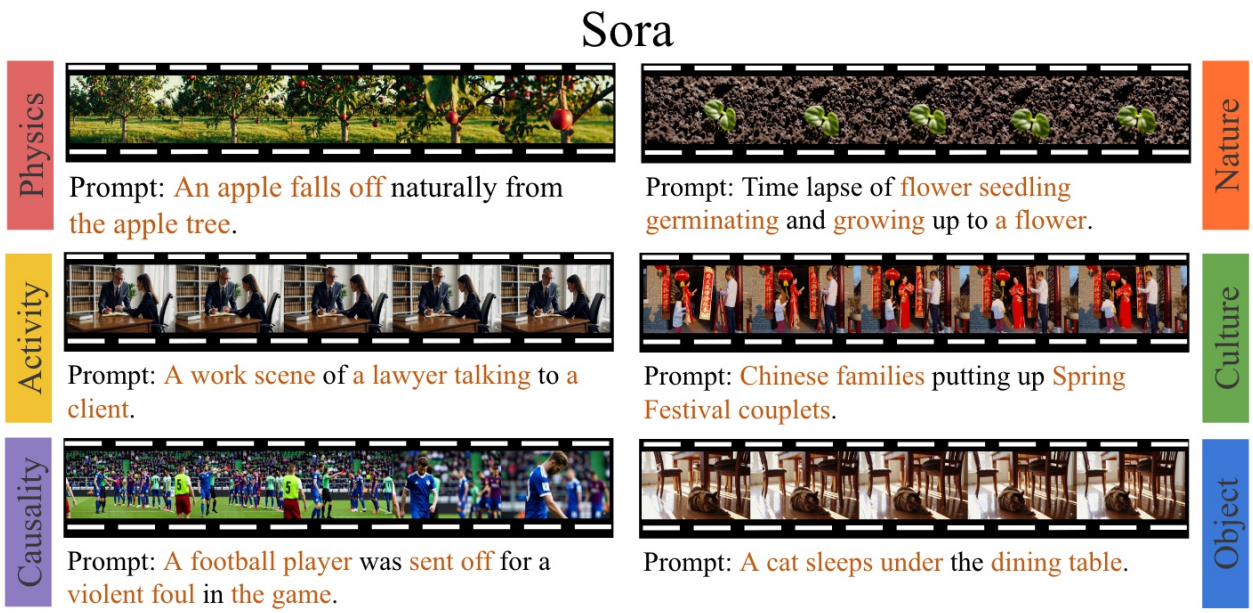

Figure 12: Sora 的视频生成。

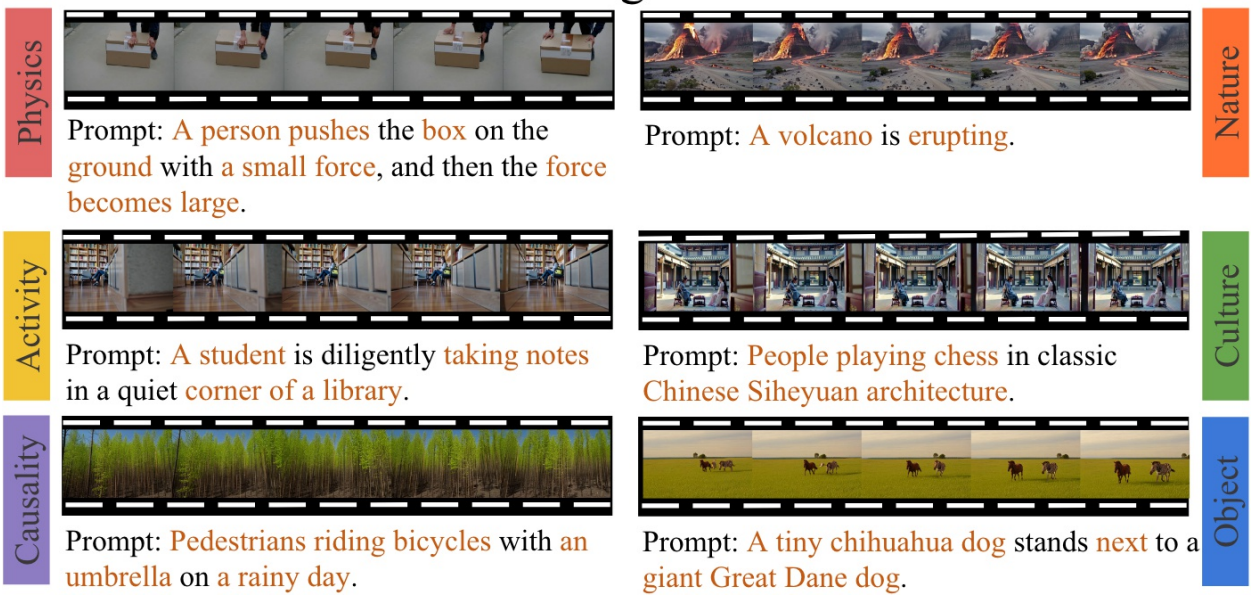

Figure 13: 克林 1.6 的视频生成。

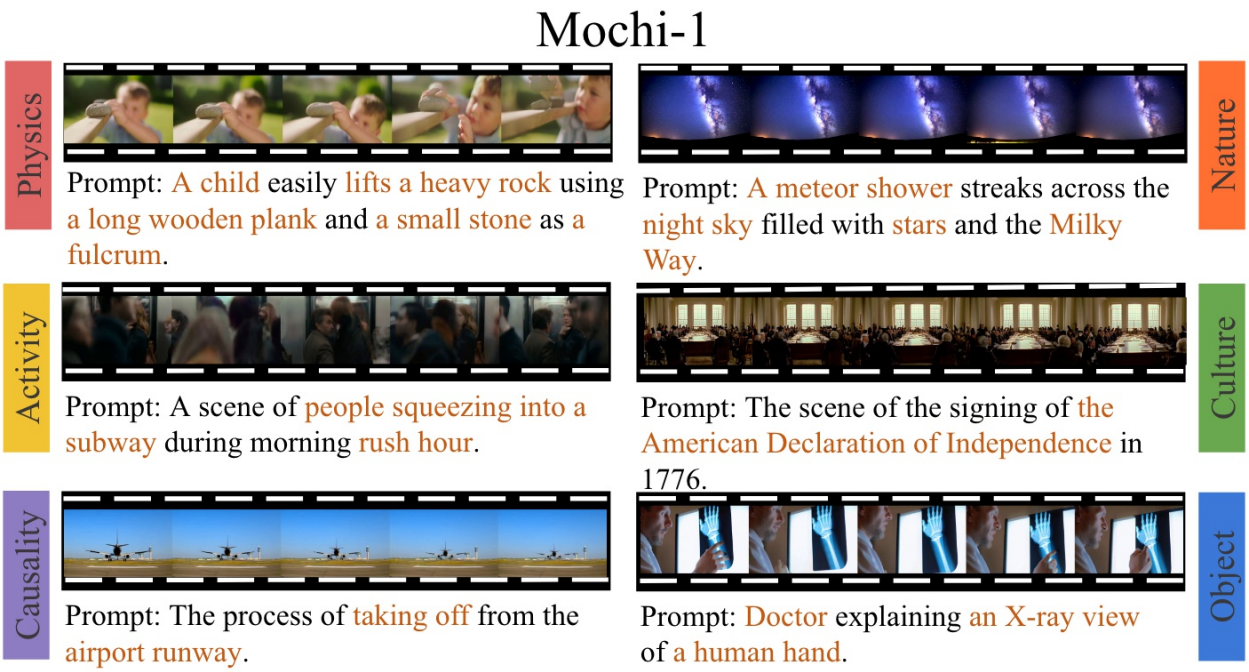

Figure 14: 糯米团 1 的视频生成。

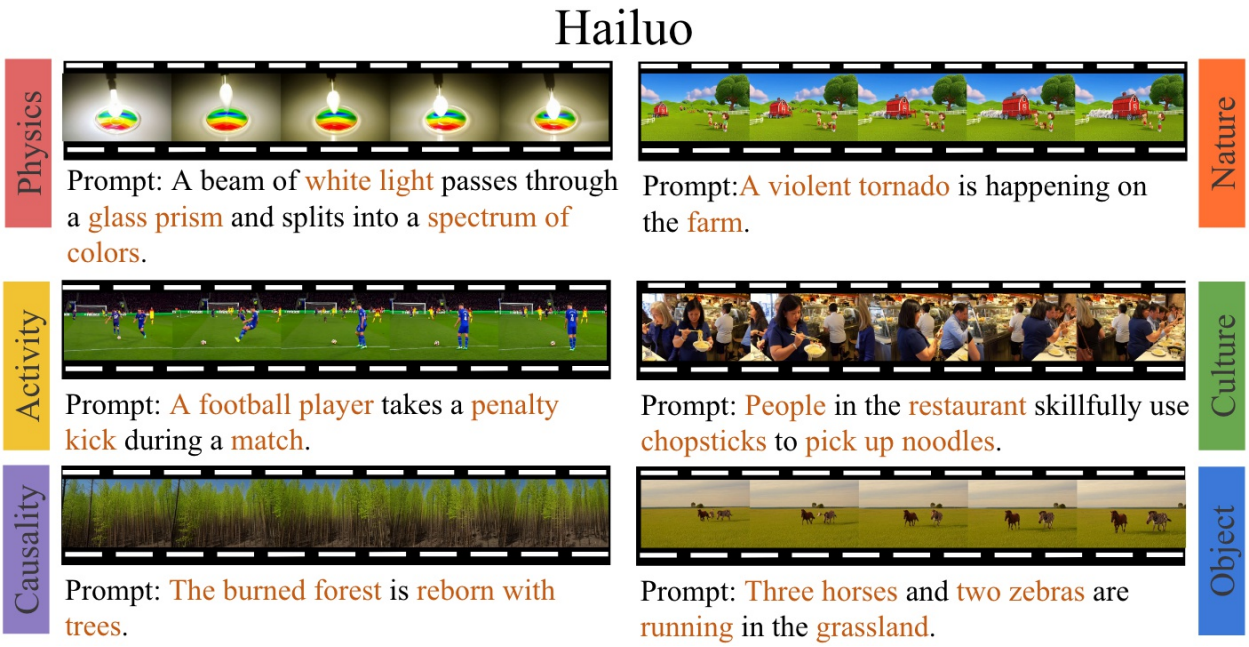

Figure 15: 海螺的视频生成。

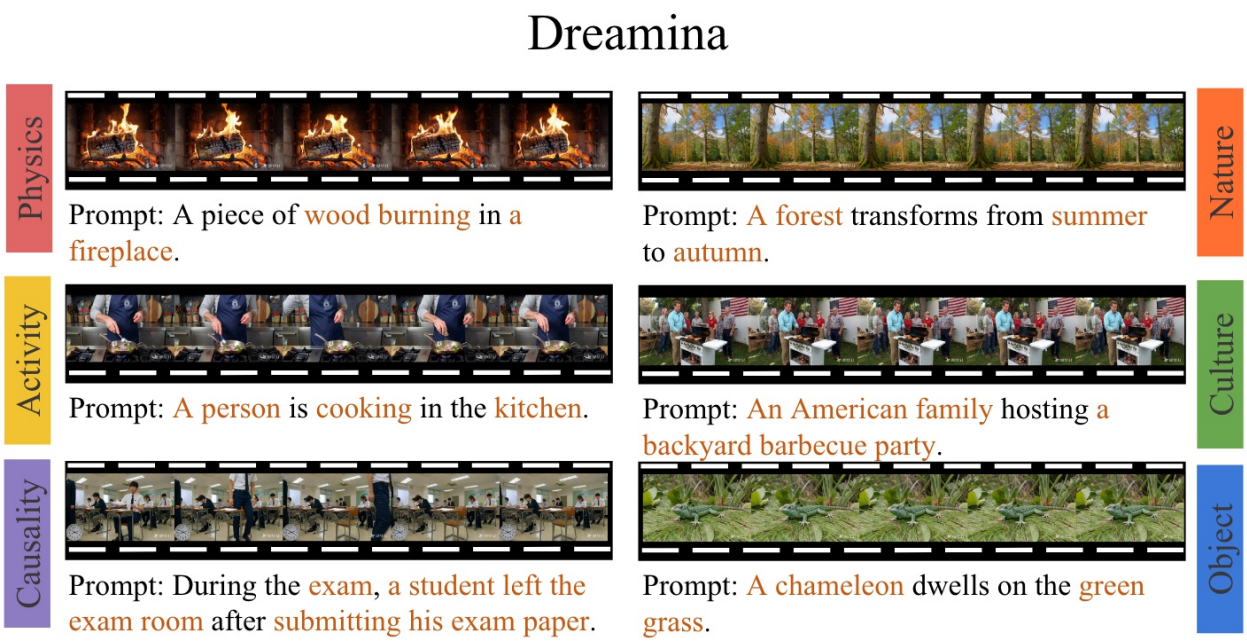

Figure 16: Dreamina 的视频生成。

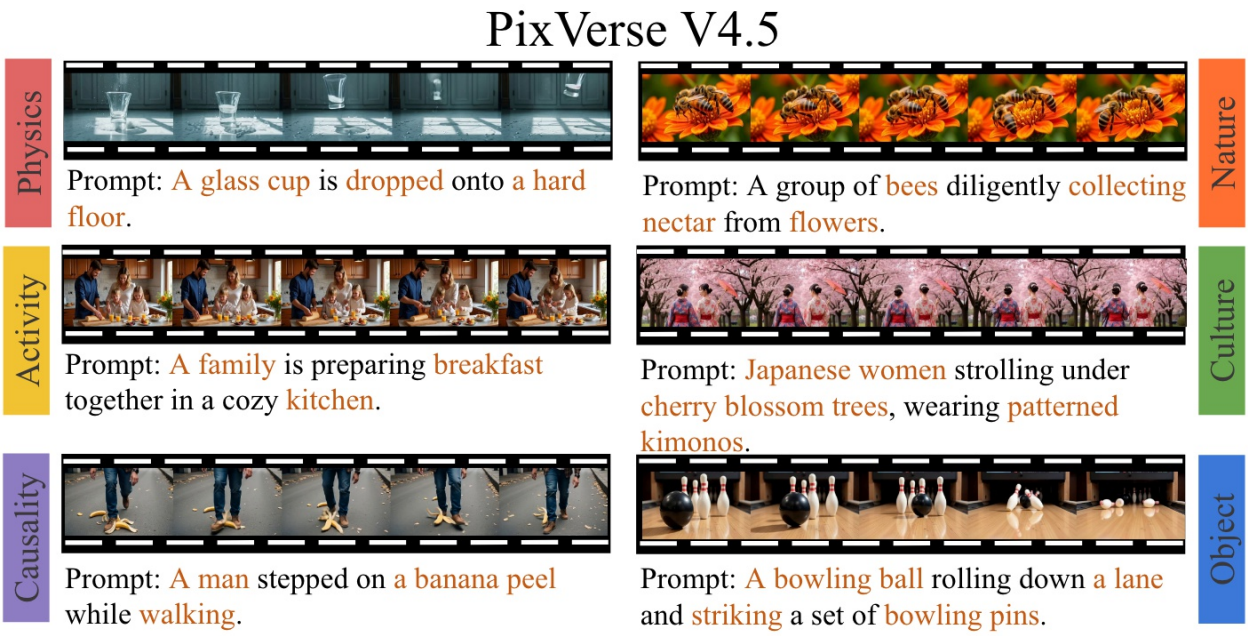

Figure 17: PixVerse V4.5 的视频生成。

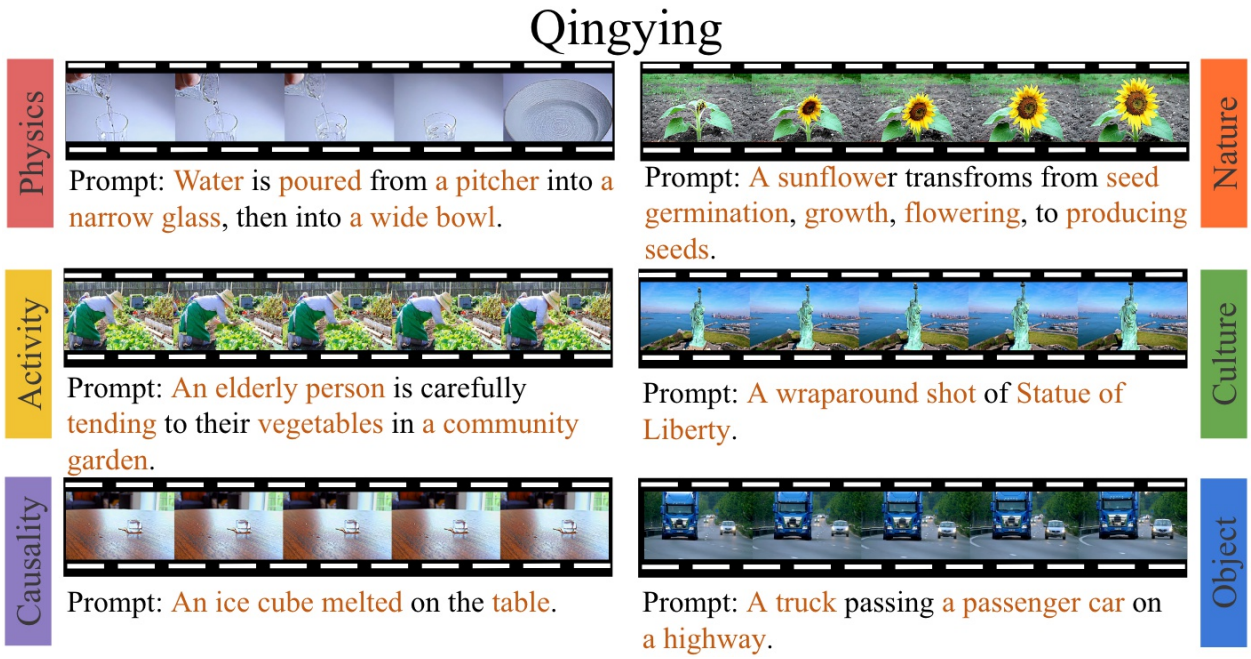

Figure 18: 青影视频生成。

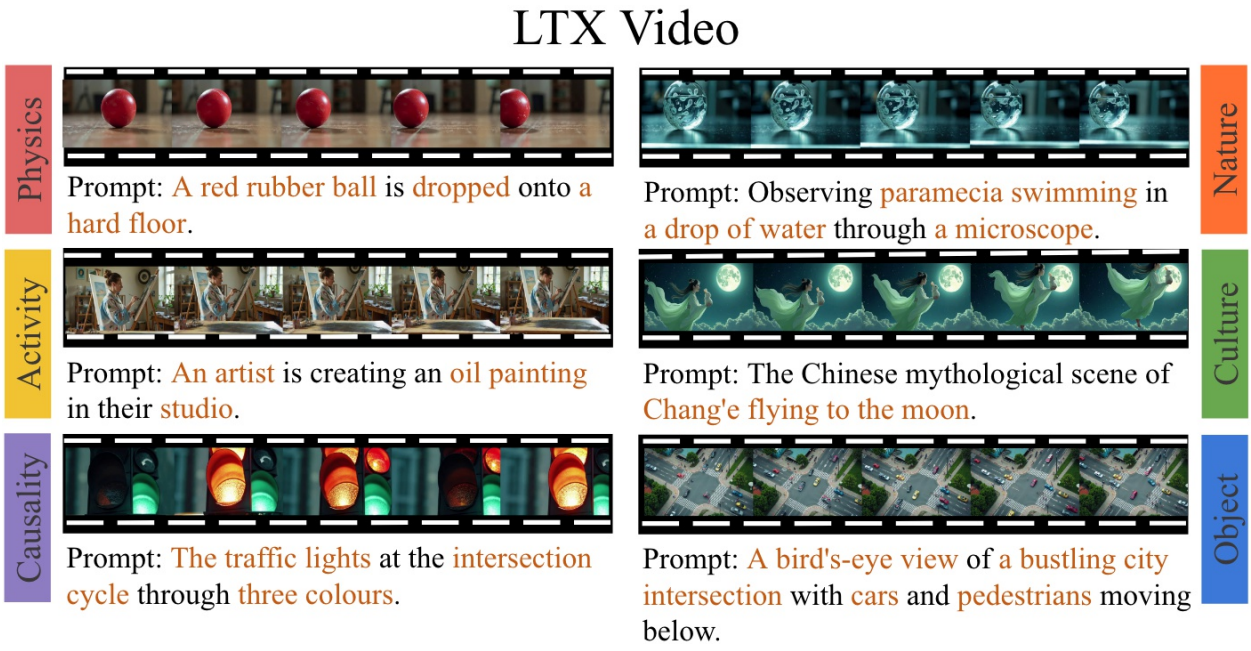

Figure 19: LTX 视频的视频生成。

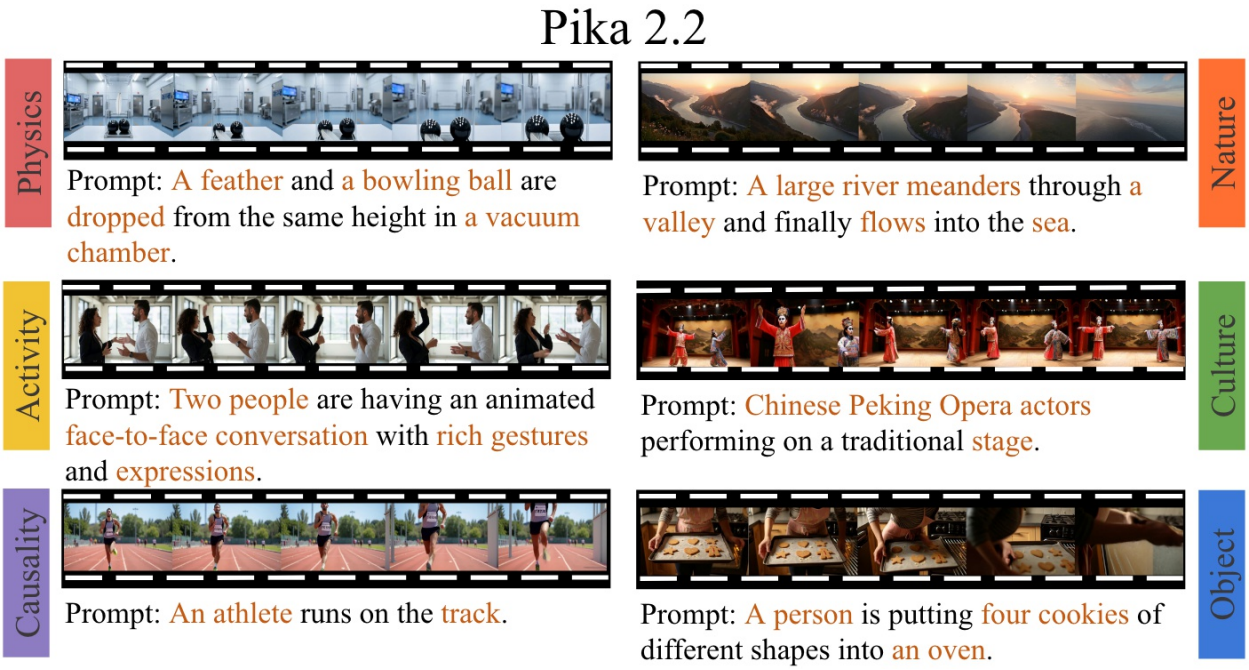

Figure 20: Pika 2.2 的视频生成。

References

[AIS25] AISphere. Pixverse v4.5, 2025.

[Ali25] Alibaba. Wan: Open and advanced large-scale video generative models, 2025.

[BRL+23] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22563–22575, 2023.

[Byt24] ByteDance. Unleash the power of ai image generator, 2024.

[CWL+24] Qinglong Cao, Ding Wang, Xirui Li, Yuntian Chen, Chao Ma, and Xiaokang Yang. Teaching video diffusion model with latent physical phenomenon knowledge. *arXiv preprint arXiv:2411.11343*, 2024.

[CXH+23] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

[CXL+24] Yi Cheng, Ziwei Xu, Dongyun Lin, Harry Cheng, Yongkang Wong, Ying Sun, Joo Hwee Lim, and Mohan Kankanhalli. Bridging the intent gap: Knowledge-enhanced visual generation. *arXiv preprint arXiv:2405.12538*, 2024.

[Gen24] Team Genmo. Mochi 1. https://github.com/genmoai/models, 2024.

[GPAM+14] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[GYZ+25] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[GZH+23] Xun Guo, Mingwu Zheng, Liang Hou, Yuan Gao, Yufan Deng, Chongyang Ma, Weiming Hu, Zhengjun Zha, Haibin Huang, Pengfei Wan, et al. I2v-adapter: A general image-to-video adapter for video diffusion models. *CoRR*, 2023.

[HCB+24] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

[HDZ+23] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *The Eleventh International Conference on Learning Representations*, 2023.

[HHY+24] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.

[HMP+17] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

[JXTH24] Pengliang Ji, Chuyang Xiao, Huilin Tai, and Mingxiao Huo. T2vbench: Benchmarking temporal dynamics for text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5325–5335, June 2024.

[KLA19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[Kli24] Kling. Kling video model, 2024.

[KW14] Diederik P Kingma and Maxs Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[LBPL19] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[LCL+24] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22139–22149, June 2024.

[LCZ+23] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Notice of removal: Videofusion: Decomposed diffusion models for high-quality video generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10209–10218, 2023.

[LDF+20] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34:07, pages 11336–11344, 2020.

[LLLL24] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

[Min25] MiniMax. Hailuo ai advances cinematic storytelling with t2v-01-director and i2v-01-director, 2025.

[MLT+24] 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.

[NNZ+25] Yuwei Niu, Munan Ning, Mengren Zheng, Weiyang Jin, Bin Lin, Peng Jin, Jiaqi Liao, Chaoran Feng, Kunpeng Ning, Bin Zhu, et al. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. *arXiv preprint arXiv:2503.07265*, 2025.

[NXZ+25] 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. In *ICLR*, 2025.

[OJK+24] Gyeongrok Oh, Jaehwan Jeong, Sieun Kim, Wonmin Byeon, Jinkyu Kim, Sungwoong Kim, and Sangpil Kim. Mevg: Multi-event video generation with text-to-video models. In *European Conference on Computer Vision*, pages 401–418. Springer, 2024.

[Ope24] OpenAI. Sora system card, 2024.

[Pik24] Team Pika. Pika labs 2.2: The future of ai-driven video generation, 2024.

[RMC16] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.

[RMW14] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.

[SHL+25] Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8406–8416, 2025.

[SPH+23] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations*, 2023.

[SVC+24] Achint Soni, Sreyas Venkataraman, Abhranil Chandra, Sebastian Fischmeister, Percy Liang, Bo Dai, and Sherry Yang. Videoagent: Self-improving video generation. *arXiv preprint arXiv:2410.10076*, 2024.

[UVSK+19] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. In *ICLR*, 2019.

[Wan25] WanTeam. Wan: Open and advanced large-scale video generative models, 2025.

[WY24] Wenhao Wang and Yi Yang. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[YTZ+24] 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.

[YWL+23] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023.

[YWPH+24] Sherry Yang, Jacob Walker, Jack Parker-Holder, Yilun Du, Jake Bruce, Andre Barreto, Pieter Abbeel, and Dale Schuurmans. Video as the new language for real-world decision making. *arXiv preprint arXiv:2402.17139*, 2024.

[Zhi24] Zhipu. Cogvideox + cogsound, 2024.

[ZHL+25] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

[ZJX+25] Daoan Zhang, Che Jiang, Ruoshi Xu, Biaoxiang Chen, Zijian Jin, Yutian Lu, Jianguo Zhang, Liang Yong, Jiebo Luo, and Shengda Luo. Worldgenbench: A world-knowledge-integrated benchmark for reasoning-driven text-to-image generation. *arXiv preprint arXiv:2505.01490*, 2025.

[ZYX+23] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.