宇宙世界基金会物理 AI 模型平台

NVIDIA [1]

Abstract

物理AI需要首先通过数字化进行训练。它需要一个自身的数字孪生体（即策略模型），以及一个世界模型（即世界的数字孪生体）。本文提出Cosmos World Foundation Model Platform以帮助开发者为他们的物理AI系统构建定制化的世界模型。我们将世界基础模型定位为一种通用的世界模型，可以被微调以适应下游应用的定制化需求。我们的平台涵盖了视频筛选流程、预训练的世界基础模型、预训练后生成的例子，以及视频分词器。为了帮助物理AI建设者解决我们社会面临的最关键问题，我们使我们的平台开源，并提供了开放权重的模型，可通过具有宽松许可的途径获取。 NVIDIA Cosmos .

1. Introduction

物理AI是一种配备有传感器和执行器的AI系统：传感器允许其观察世界，而执行器则允许其与世界互动并对其进行修改。它承诺可以释放人类工人从危险、繁重或乏味的物理任务中解脱出来。尽管在过去十年中，由于数据和计算能力的提升，AI的多个领域取得了显著进展，但物理AI的发展却相对缓慢。这主要是因为训练物理AI的数据扩展更具挑战性，因为所需的数据必须包含交错的观察和行动序列。这些行动会扰动物理世界，并可能导致系统和世界遭受严重损害。尤其是在AI还处于初级阶段时，探索性的行动至关重要。一种世界基础模型（World Foundation Model，WFM），即一个物理世界的安全数字双胞胎，已被长期视为解决数据扩展问题的解决方案。

在本文中，我们介绍了用于构建物理AI的Cosmos World Foundation Model (WFM)平台。我们主要关注视觉世界基础模型，其中观测数据以视频形式呈现，扰动可以以多种形式存在。如图所示： Fig. 2 我们提出了一种预训练-然后后训练的范式，将WFMs分为预训练和后训练的WFMs。为了构建一个预训练的WFM，我们利用大规模的视频训练数据集使模型接触到多样化的视觉体验，从而使其成为通才。为了构建一个后训练的WFM，我们对预训练的WFM进行微调，使用特定物理AI环境收集的数据集来达到针对特定、专门化物理AI设置的专业化WFM。 Fig. 1 显示了我们训练前和训练后的 WFM 的示例结果。

数据决定了AI模型的上限。为了构建一个高上限的预训练WFM（假设WFM为特定上下文中的术语），我们开发了一个视频数据整理管道。我们使用该管道来定位视频中动态丰富且视觉质量高的片段，这些片段有助于学习嵌入在视觉内容中的物理知识。我们从包含200万小时视频的集合中提取了大约1亿个长度在2到60秒之间的片段。对于每个片段，我们使用视觉语言模型（VLM）以每256帧生成一段视频字幕。视频处理计算密集型。我们利用现代GPU中可用的H.264视频编码器和解码器的硬件实现来进行解码和转码。我们的视频数据整理管道利用了许多预训练的图像/视频理解模型。这些模型具有不同的吞吐量。为了最大化生成可训练视频数据的整体吞吐量，我们构建了一个基于Ray的编排管道（假设Ray为特定上下文中的术语）。 莫里茨等人。 ， 2017 ）。细节在 Sec. 3 .

We explore two scalable approaches for building pre - trained WFM discussed in Sec. 5. These approaches are

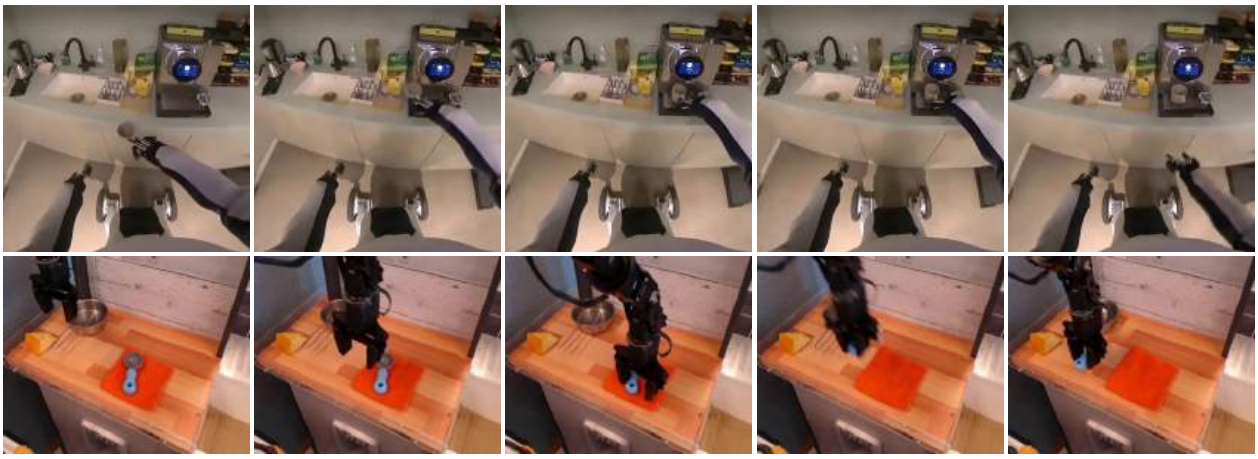
[1]贡献者和确认的详细列表可以在 App. A这篇文章。

训练前 ： 扩散 WFM

训练前 ： 自回归 WFM

培训后 ： 摄像头控制

训练后 ： 机器人操纵



培训后 ： 自动驾驶

图 1: **宇宙世界基金会模型** ．预训练的Cosmos WFMs生成高质量的3D一致视频，并具备准确的物理模拟。Cosmos模型套件包括扩散模型和自回归变换器模型，前者使用连续的潜在表示，后者使用离散的潜在表示来训练视频。通过使用专门的数据集对这些WFMs进行后训练，使其能够在广泛的物理AI设置中得到应用。具体来说，我们展示了具有摄像机可控性的模型、能够遵循指令进行机器人操作的模型以及适用于自动驾驶场景的模型。如需查看完整视频及其他更多视频示例，请访问我们的网站。 网站 ．

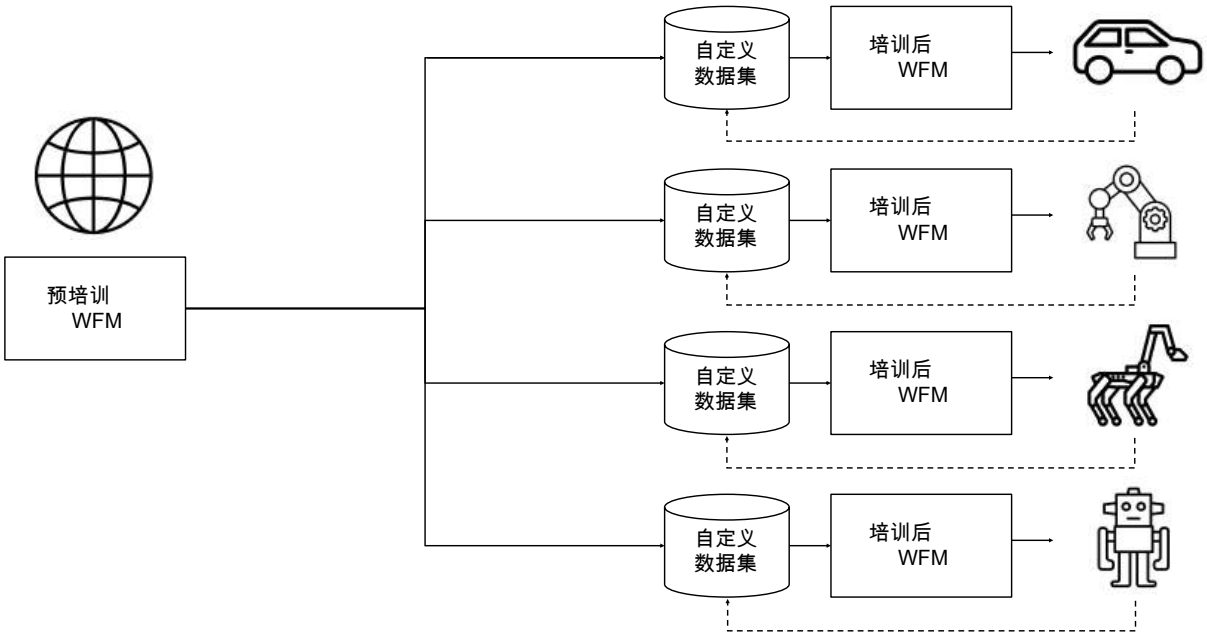


图 2 ： 预训练的 WFM 是世界模型通才 ， 使用大规模 ， 多样化的视频数据集进行训练
捕获真实世界物理的不同方面。这些预先训练的世界基础模型可以是专门的
到目标物理 AI 设置通过训练后。通常 ， 用于训练后的数据集是 “提示 ” - 视频
从目标物理 AI 设置中收集的对。提示可以是动作命令、轨迹、
说明， *etc.* 由于预训练的 WFM 提供了很好的基础 ， 用于后期训练的数据集可以是
这种训练前和训练后产生了构建物理人工智能系统的有效策略。
在该图中, 虚线表示数据循环。

基于变压器的扩散模型和基于变压器的自回归模型。扩散模型通过逐步去除高斯噪声视频中的噪声来生成视频。自回归模型则按预设顺序逐步生成视频片段，条件依赖于过去的生成结果。这两种方法都将一个复杂的视频生成问题分解为更易于处理的子问题，使其更具可操作性。我们利用最先进的变压器架构以实现其可扩展性。 Sec. 5.1 ， 我们提出了一种基于变压器的扩散模型设计 ， 该模型具有强大的世界生成能力。在 Sec. 5.2 ， 我们提出了一种基于变压器的世界发电自回归模型设计。

基于变换器的扩散模型和基于变换器的自回归模型都使用令牌来表示视频，其中前者以向量形式的连续令牌进行表示，而后者以整数形式的离散令牌进行表示。我们注意到，将视频转换为一组令牌的过程——即将视频转换为一系列令牌——是非常复杂的。视频包含了关于视觉世界的丰富信息。然而，为了使世界基础模型（WFMs）的学习得以进行，我们需要将视频压缩为一系列紧凑的令牌序列，并在计算复杂性随着令牌计数增加时最大限度地保留视频中原有的内容。在很多方面，构建一个视频编码器类似于构建一个视频编码器。我们开发了一种基于注意力的编码器解码器架构，用于学习对上述描述的连续和离散令牌的视频编码。 Sec. 4 .

我们微调预先训练的 WFM ， 以到达训练后的 WFM ， 用于在 Sec. 6 . In Sec. 6.1 我们调整预训练的扩散WFM使其成为相机姿态条件化。这一后训练过程创建了一个可导航的虚拟世界，用户可以通过移动虚拟视角来探索这个创建的世界。 Sec. 6.2 我们对各种机器人任务进行细调，这些任务包括视频-动作序列。我们表明，通过利用预训练的WFMs，我们可以更好地根据当前状态预测世界未来的状态。

机器人采取的行动。在 Sec. 6.3 ，我们演示了如何针对各种与自动驾驶相关的任务对预先训练的 WFM 进行微调。

我们开发的WFMs（世界基础模型）的预期用途是为物理AI构建者服务。为了更好地保护开发者在使用这些世界基础模型时的安全，我们开发了一个强大的防护系统，该系统包括一个预防护模块以阻止有害输入，以及一个后防护模块以阻止有害输出。详细内容将在后续部分描述。 Sec. 7 .

我们旨在构建一个全球基础模型平台，以帮助实体人工智能建设者提升其系统。为了达成这一目标，我们根据NVIDIA开放模型许可协议，在以下链接提供了预训练的世界基础模型和分词器：

[Insert link here] NVIDIA Cosmos and NVIDIA Cosmos Tokenizer 培训前脚本和培训后脚本将分别在 NVIDIA Nemo 框架 借助视频数据整理管道来帮助构建者制作微调数据集。尽管本文在世界基础模型设计方面做出了多项改进，但世界基础模型的问题仍然远未解决。还需进一步的研究以推动该领域的进步。

2. 世界基金会模型平台

Let
0: 是从时间 0 到现实世界的一系列视觉观察 . Let
是对世界的扰动。如 Fig. 3 ，WFM 是一个模型 预测未来的观测时间 + 1，
+1 ，基于过去的观察
0: 和当前的扰动
在我们的案例中，
0: 是 RGB 视频 ，而
是一种可以采取多种形式的扰动。它可以是物理AI采取的行动、一个随机扰动，或者扰动的文字描述等。

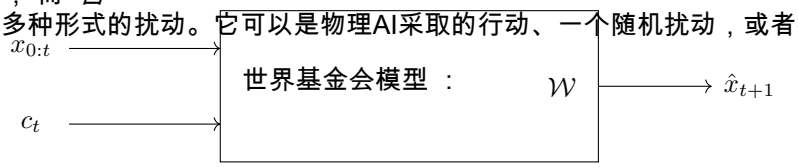


图 3 ： 世界基础模型 (WFM) 是一个生成未来世界状态的模型
基于过去的观察 和电流扰动。 +1
0:

2.1. 未来宇宙

我们认为 WFM 在许多方面对物理 AI 构建者有用 ， 包括(但不限于)

- **政策评估。** 这指的是对物理AI系统中的政策模型质量进行评估。与其将训练好的政策部署到实际运行的物理AI系统中进行评估，不如让物理AI系统的数字副本与世界基础模型互动。基于WFM的评估更加经济高效且节省时间。借助WFM，建设者可以在未见过的环境中部署政策模型，这些环境通常是不可用的。WFM可以帮助开发者快速排除无效的政策，并将物理资源集中在少数有前景的政策上。 策略模型生成要由物理 AI 系统根据

策略初始化。
当前的观察结果和给定的任务。一个well-trained WFM可以根据输入的扰动模型世界的动态模式，可以作为政策模型的良好初始化。这有助于解决物理AI中的数据稀缺问题。• 与奖励模型配对的 WFM 可以是物理世界提供的代理

政策培训。
在强化学习设置中对政策模型进行反馈。代理可以通过与工作流管理器（WFM）交互来提升解决任务的技能。

- **规划或模型预测控制。** 一种WFM可以用于模拟物理AI系统采取不同行动序列后可能出现的各种未来状态。然后可以使用成本/奖励模块根据结果量化这些不同行动序列的表现。最后，物理AI可以根据整体模拟结果执行表现最佳的行动序列，类似于规划过程。

算法或在退潮式（receding horizon）方式下，如同模型预测控制（model-predictive control）所采用的方法。世界的模型上界定义了这些决策策略的性能准确度。

- **合成数据生成。** 一个WFM可以用于生成合成数据进行训练。此外，它可以进一步调整以条件化渲染元数据，如深度图或语义图。对于模拟到现实的应用场景，可以使用条件化的WFM。

尽管列出了这些可能性，本论文并未包含将Cosmos WFMs应用于它们的实证结果。我们渴望在未来的工作中验证这些主张。

2.2. Current Cosmos

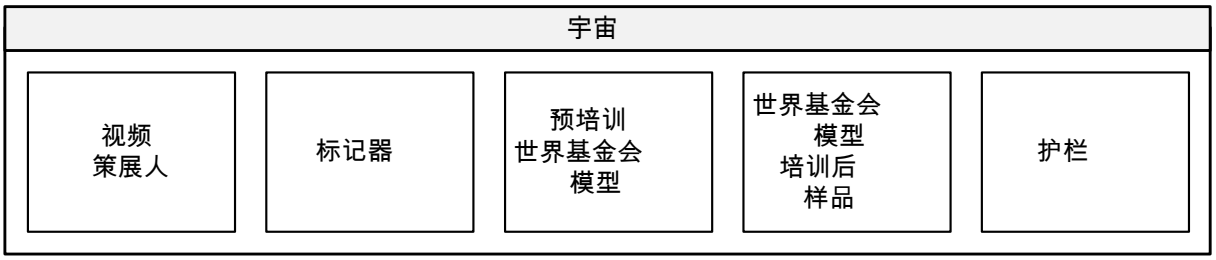


图4：Cosmos World 基金会模型平台由多个主要组件组成：视频策展人、视频分词器、预训练的世界基础模型、世界基础模型后训练样本以及边界防护。

Fig. 4 本报告可视化了Cosmos WFM平台中包含的内容，该平台包括视频策展、视频令牌化、世界基础模型预训练、世界基础模型后训练以及护栏功能。

视频策展人。 我们开发了一套可扩展的视频数据整理管道。每个视频被分割成单独的镜头，不包含场景变化。然后应用一系列筛选步骤来识别高质量且信息丰富的子集用于训练。这些高质量的镜头随后使用VLM进行标注。接着我们执行语义去重以构建一个多样化但紧凑的数据集。

视频标记化。 我们开发了一系列不同压缩比的视频分词器家族。这些分词器具有因果性特征。当前帧的分词计算并不基于未来的观察，而是基于过去和当前的信息。这种因果设计具有多个优点。在训练方面，它使得联合图像和视频的训练成为可能，因为当输入为单张图片时，因果视频分词器同时也是图像分词器。这对于视频模型利用包含丰富世界外观信息且往往更加多样的图像数据集进行训练至关重要。在应用方面，因果视频分词器与生活于因果世界的物理人工智能系统更相匹配。

WFM 预培训。 我们探索了构建预训练世界基础模型的两种可扩展方法——扩散模型和自回归模型。我们采用transformer架构，因其具有可扩展性。

对于基于扩散的WFM，预训练包含两个步骤：1) 文本到世界生成预训练；2) 视频到世界生成预训练。具体而言，我们首先训练模型根据输入的文字提示生成一个视频世界。然后，我们将模型进一步微调，使其能够基于过去的视频和输入的文字提示生成未来的一个视频世界，这一任务被称为视频到世界生成任务。

对于基于自回归的WFM，预训练包括两个步骤：1）常规的下一个token生成和2）基于文本条件的Video2World生成。我们首先训练模型根据输入的历史视频生成未来视频世界——前瞻生成。然后，我们进一步微调该模型，使其能够根据历史视频和一个文本提示生成未来视频世界。

video2world 生成模型是一个预先训练的世界模型 ， 它根据当前的

观测（过去视频）和控制输入（提示）。对于基于扩散和自回归的WFMs，我们构建了一组具有不同容量的模型，并研究了它们在各种下游应用中的有效性。

我们进一步 Fine-tune 预训练的扩散 WFM 以获得一个扩散解码器，从而提升自回归模型生成结果的质量。为了更好地控制 WFM，我们还基于大型语言模型（LLM）构建了一个提示上采样器。

世界模型培训后。 我们展示了预训练WFMs在若干下游物理AI应用中的应用场景。我们通过将相机姿态作为输入提示对预训练的WFMs进行微调，从而能够在创建的世界中自由导航。我们还展示了我们的预训练WFMs如何可能被微调以应用于类人机器人和自主驾驶任务。

护栏. 为了确保发达国家的基础模型安全使用，我们开发了一套护栏系统，以阻止有害输入和输出。

3. 数据固化

我们描述了我们的视频编目流水线，该流水线生成高质量的训练数据集，用于both tokenizers和WFMs。如所示，在 Fig. 5 我们的管道包括五个主要步骤：1）分割，2）过滤，3）标注，4）去重，和5）分片。每一步都是为了提高数据质量并适应模型训练的需求而量身定制的。我们首先介绍原始数据集，然后详细描述每一个步骤。

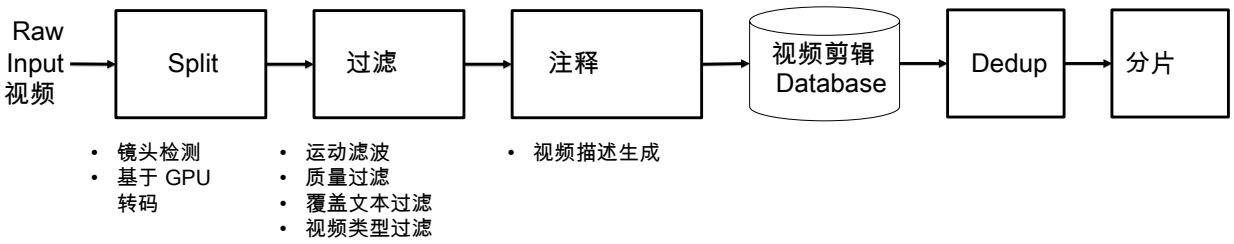


图 5: **Cosmos Video Curator 包含五个主要步骤 ： 1) 拆分 ， 2) 过滤 ， 3) 注释 ， 4) 整理和 5) 分片** 分割步骤将长视频划分为镜头，并将其转录为片段。过滤步骤移除对世界基础模型构建价值较小的片段。标注步骤为每个片段添加视频描述。然后将这些片段存储在视频片段数据库中。为了获取训练数据集，首先进行语义去重，然后根据分辨率和纵横比对视频片段进行切分。

3.1. Dataset

我们使用了自有专属的视频数据集以及公开获取的跨域互联网视频，以训练我们的模型。我们的目标是赋能物理人工智能（Physical AI）开发者。为此，我们精心筛选了视频训练数据集，旨在覆盖各种物理AI应用，并针对以下视频类别进行聚焦：

1. 驱动（11%），2. 手部动作和物体操作（16%），3. 人类运动与活动（10%），4. 空间意识与导航（16%），5. 第一人称视角（8%），6. 自然动力学（20%），7. 动态摄像机移动（8%），8. 合成渲染（4%），以及9. 其他（7%）。

这些视频提供了不同视觉对象和动作的广泛覆盖。它们的多样性提高了我们的WFMs的一般化能力，并帮助模型处理不同的下游任务。这些视频的非结构化性质及其庞大的数量从算法和基础设施两个角度来看都带来了许多挑战，以高效地处理它们。这些视频可以用各种各样的编码格式进行编码，并具有不同的宽高比、分辨率和长度。 *etc* .许多视频还经过了后处理或编辑，并应用了不同的视觉效果，如果不适当处理，这些视频中可能会产生不必要的伪影，从而影响世界模型的性能。

总共积累了约2000万小时的原始视频数据，分辨率为720p至4k不等。然而，大量的视频数据要么在语义上是冗余的，要么不包含有助于学习世界物理规律的有效信息。因此，我们设计了一系列数据处理步骤，以找出原始视频中最有价值的部分用于训练。我们还收集了图像数据，因为联合使用图像和视频进行训练已被证明可以提高生成视频的视觉质量，并加快模型训练速度。得益于我们数据整理管道的模块化设计，我们可以使用它来处理图像和视频数据，并生成用于预训练和微调的数据集。我们生成了大约10 ^{8} 用于训练前的视频剪辑和大约 10 ^{7} 微调。

3.2. Splitting

我们的视频长度不固定，现代深度学习模型无法直接处理非常长的视频。此外，许多视频包含镜头转换。这些视频可能从一个场景开始，然后过渡到一个完全不同的场景，两个场景之间可能没有任何联系。 *e.g* 从纽约市一个现代厨房里两个人的对话场景到非洲草原上狮子追逐斑马的画面。重要的是根据每个视频的镜头变化进行分段，并生成视觉上一致的视频片段，以便模型能够学习物理上合理的视觉内容过渡，而不是人工编辑的内容。

3.2. 1. Shot Detection

分割旨在将任意长度的原始视频暂时分割成没有镜头切换的片段。它以原始视频作为输入，并生成每个镜头的起始和结束帧索引。长度短于2秒的片段会被丢弃，因为它们可能是镜头过渡或视觉效果。长度超过60秒的片段将进一步分割，以确保最大长度为60秒。后续的过滤步骤可以确定一个片段是否包含用于学习世界物理信息的有用信息。

剪辑边界检测是经典计算机视觉问题。现有方法基于视觉特征空间的变化来检测剪辑边界，但它们在如何从视频帧中学习视觉特征方面有所不同。我们评估了几种用于此任务的算法。 Tab. 1 : PySceneDetect (卡斯特拉诺 ， 2024) ， Panda70M (Chen et al. ， 2024) ， TransNetV2 (Soucek 和 Lokoc ， 2024) 和自动射击 (朱等人。 ， 2023).

PySceneDetect 是一个流行的库，用于通过阈值化HSV颜色直方图在时间上的变化来检测场景变化。请注意，它也被最近的MovieGen工作采用 ( ) 。 Polyak 等人。 ， 2024). 熊猫70M通过基于CLIP嵌入的拼接和过滤增强了PySceneDetect。另一方面，TransNetV2和AutoShot是基于神经网络的方法，能够在给定100帧滚动输入窗口的情况下预测每帧成为过渡帧的概率。

它至关重要的是选择一种能够很好地处理 heavily edited 视频的算法，因为这类视频通常包含复杂的镜头转换和各种视觉效果。这促使我们构建一个专门的基准来评估方法是否能够从视频中生成具有清晰镜头剪辑的片段。我们的基准 (命名为) [2] ShotBench) 包括现有数据集 ， 如 RAI 、 BBC Planet Earth (摩德纳大学 AI 图像实验室 ， 2016) ， ClipShots (唐等人。 ， 2018) 和射击 (朱等人。 ， 2023 对于ClipShots，我们将过渡帧定义为每个片段注释开始和结束时间点的中点，以与其他数据集保持一致。

[2] ShotBench 可在 https: / / github. com / NVlabs / ShotBench .

表 1 ： 不同数据集上的拆分算法比较。

数据集度量		PySceneDetect	Panda70M	TransNetV2	AutoShot
BBC	Precision ↑	0.894	0.959	0.983	**0.984**
	召回 ↑	0.884	0.653	**0.951**	0.922
	F1 ↑	0.889	0.777	**0.967**	0.952
RAI	精度 ↑	0.856	**0.933**	0.918	0.889
	召回 ↑	0.807	0.746	0.921	**0.923**
	F1 ↑	0.831	0.829	**0.919**	0.906
射击	精度 ↑	0.769	**0.949**	0.883	0.866
	召回 ↑	0.673	0.462	0.767	**0.804**
	F1 ↑	0.718	0.622	0.821	**0.834**
ClipShots	精度 ↑	0.395	0.649	**0.685**	0.653
	召回 ↑	0.602	0.424	0.772	**0.781**
	F1 ↑	0.477	0.513	**0.726**	0.711

Tab. 1 比较了不同方法在ShotBench上的表现。我们将置信阈值设置为0.4，适用于TransNetV2和AutoShot。对于Panda70M，我们遵循其分隔实现，不包括过滤步骤，以确保公平比较。端到端学习方法（End-to-end learning-based approaches（ *e.g* 。 ， TransNetV2 和 AutoShot) 比使用手工制作的功能或启发式规则的方法（ *e.g* 尽管TransNetV2和AutoShot在现有数据集上的表现相当，我们发现TransNetV2在更具有挑战性的镜头变化场景中表现得更好。使用端到端的神经网络（ *i.e* TransNetV2) 还允许我们通过利用现代GPU进行加速来提高分割的吞吐量，而不受混合方法（如Panda70M）的限制，这些方法使用复杂的逻辑将PySceneDetect和ImageBind嵌入相结合。 Girdhar 等人。 ， 2023).

3.2. 2. 转码

我们的视频使用了多种不同的编解码器和各种设置，这给数据整理带来了挑战。我们从镜头检测开始重新编码每个视频片段为一致且高质量的mp4格式。这简化了后续的数据整理过程。采用统一的视频编解码器后，模型训练数据加载器的稳定性和效率也得到了显著提高。我们使用h264_nvenc编解码器并采用高比特率，并通过使用快速运动和高频纹理的视频进行压力测试，以确保不会出现可感知的视觉退化。

我们全面评估不同的硬件和软件配置以最大化转码 throughput ，并在各种环境下优化性能。 Tab. 2 现代GPU提供了硬件加速的视频编码和解码能力。NVIDIA L40S拥有用于解码（NVDEC）和编码（NVENC）的硬件加速器，而NVIDIA H100仅具有NVDEC。为了公平地与L40S进行比较，我们将H100的CPU核心数量补偿为最大可用值（28个，而不是默认的1个）。 Tab. 2 。 L40S 的吞吐量比 H100 高约 17% (0.0674 *vs* . 0.0574)。对于软件配置，从libx264切换到h264_nvenc，并且批量转码来自同一视频的多个片段，显著提升了吞吐量。我们观察到ffmpeg在充分利用NVDEC/NVENC加速器方面存在问题，尤其是在多GPU节点上。使用PyNvideoCodec替换ffmpeg进行视频流转码后，加速器利用率大幅提高，吞吐量提升最为显著（0.3702）。 *vs* . 0.1026)。我们仅保留ffmpeg用于音频混音，并使用PyNvideoCodec以更好地利用GPU的计算能力。我们实现了 ∼ 6 . 5 × 当将所有改进组合在一起时 ， 吞吐量会增加。

3.3. Filtering

从拆分步骤生成的视频片段存在噪声，覆盖各种主题且质量差异巨大。我们设计过滤步骤旨在：
1) 移除视觉质量未达到最低要求的视频片段，
2) 选取适合精细调整的高质量视频片段，
3) 调整数据分布以构建工作流模型（WFMs）。

通过执行运动过滤、视觉质量过滤、文本过滤等操作来实现上述目标。

表 2 ： 不同软件设置的转码性能。

方法 GPU CPU 编解码器批量 NVDEC (# 加速器) NVENC (# 加速器) 吞吐量 (视频 / 秒)
ffmpeg H100 28 libx264 1 7 0 0.0574
ffmpeg L40S 1 h264 _ nvenc 1 3 3 0.0674
ffmpeg L40S 1 h264 _ nvenc 16 3 3 0.1026
pynvc + ffmpeg L40S 1 h264 _ nvenc 1 3 3 **0.3702**

视频类型过滤。

3.3. 1. Motion Filtering

我们在运动过滤方面有两个主要目标：1）去除静态视频或具有随机突兀摄像机运动的视频（通常来自手持摄像机）；2）标记具有不同类型的摄像机运动的视频。 *e.g* 。 ， 平移 ， 缩放 ， 倾斜 ， *etc* 。）， 它可以提供额外的信息来指导模型训练。

我们构建了一个轻量级分类器用于运动滤波。分类器的输入是从视频片段中提取的一系列运动向量或光学流。该分类器基于ViT架构，并通过带有标签的视频进行训练。我们尝试使用h264编解码器的运动向量、Farneback光学流算法等数据进行实验。 Farneb ä ck ，
2003 ），并且采用了基于NVIDIA TensorRT加速的光学流估计网络。我们发现，在NVIDIA TensorRT加速的光学流估计基础上构建的分类器效果最佳，能够产生较高的分类准确率以进行运动过滤。

3.3. 2. 视觉质量过滤

我们考虑了两个指标，失真和外观质量，用于基于视觉质量的过滤。首先，我们移除了包含失真的视频片段，比如伪影、噪声、模糊、低清晰度、过曝和欠曝等。 *etc* 。我们使用在基于 DOVER 的人工评分视频上训练的视频质量评估模型（ 吴等人。 ， 2023). 这种方法给出每个片段的感知质量分数，并使用这些分数去除排名最低的 15% 的片段。其次，我们过滤掉低视觉质量的视频片段。我们应用一个图像美学模型（ 舒曼 ，
2022 ）来自输入剪辑的采样帧。我们设置了一个保守的阈值 ， *i.e* ., 3 ． 5 ， 因为美学对于物理 AI 不那么重要。

3.3. 3. Text Overlay Filtering

我们的一些视频经过后期处理 ， 以添加文本以包含查看器的其他信息。我们还

发现文本往往会与不同的视觉效果共现。我们的目标是学习世界的基本物理规律。去除此类过度包含文本的视频至关重要。请注意，我们关注的是在后期处理中添加的文本，而不是视频原始场景中的文本，例如驾驶视频中的街道名称。

我们训练一个基于MLP的二元分类器以检测此类视频。分类器的输入是从InternVideo2提取的视频嵌入。 W ang 等人。 ， 2025 我们使用专有的VLM构建训练集以标记正面和负面视频。经过训练的模型在验证集上实现了高预测准确性。

3.3. 4. 视频类型过滤

为了调整训练数据分布并过滤掉不必要的视频类型，我们设计了一个全面的分类体系，根据视频的内容类型和视觉风格对视频进行分类。我们训练一个分类器为每个视频片段标注分类体系中的类别。通过排除可能导致生成质量差或不现实动态的具体视频类型，如抽象视觉图案、电子游戏画面、动画内容等，我们进一步精炼了数据。 *etc* 。我们通过从与 WFM 更相关的类别上采样来进一步调整数据分布（ *e.g* 。 ， 人类行为 ， 人与物体的交互 ， *etc* 。）和对不太重要的类别进行降采样（ *e.g* 。 ， 自然或风景视频）。

鉴于缺乏与我们分类体系相匹配的预标注数据集，我们利用 proprietary VLM 创建分类器的训练和评估数据。对于每个视频片段，我们向 VLM 提供八个

均匀采样的帧并查询最合适的分类标签。利用标注的数据，我们使用来自文本过滤的相同InternVideo2嵌入值训练一个多层感知器（MLP）分类器。

3.4. Annotation

文本描述通常与图像和视频数据配对，以提供监督和用于世界模型训练的条件。我们使用视觉语言模型（VLM）为每个视频片段生成高质量且一致的标题。我们将VLM配置为专注于视频中的材料事实和细节。通过这种方式为视频提供描述，而不是依赖于Alt文本，也减轻了世界模型的学习负担，因为我们无需在训练过程中适应不同的文本样式或格式。

我们测试了几种 SOTA 方法（*i.e* ., VFC（Ge 等人。 ， 2024 ）， Qwen2 - VL（Wang 等人。 ， 2024 ）, VILA（Lin et al. ， 2024 ；
薛等人。 ， 2024)) 用于视频的字幕生成，我们发现VILA能够基于小型人工评估生成更准确的描述。我们使用一个内部的VILA模型，参数量为13B，专门针对视频字幕进行了微调。该模型具有扩大的上下文窗口，适用于处理长且多帧的上下文，最大输入和输出token长度分别为5904和256。为了提高推理效率，我们使用了FP8-量化TensorRT-LLM引擎，结果实现了10 × 与 PyTorch 半精度基线相比 ， 吞吐量加快 ， 如 Tab. 3 。我们向 VILA 提示“ *详细阐述视频的视觉和叙事元素* ”， 并从输入剪辑中输入 8 个均匀采样的帧。字幕的平均长度为 559 个字符或 97 个单词。

表 3: VILA 在单个 H100GPU 上的推断吞吐量比较。

吞吐量 (令牌 / 秒)			
PyTorch FP16 1 0.21 49.6			
TRT - LLM FP16 1 0.40 95.6			
TRT - LLM FP16 16 1.09 260.9			
TRT - LLM FP8 16 1.96 470.6			

3.5. 重复数据删除

鉴于我们视频数量庞大，训练集内可能存在重复或近似重复的样本。清除重复数据对于创建更加均衡和多样的数据分布至关重要。这还能提高训练效率并降低记忆特定训练样本的可能性。

我们采用 SemDeDup 的方法（阿巴斯等人。 ， 2023 ）和 DataComp（Gadre 等人。 ， 2024 ）用于可扩展的语义去重。我们重用在过滤过程中计算的InternVideo2嵌入，并使用多节点GPU加速的k-means实现对嵌入进行聚类。 RAPIDS ， 2023 ）with *号* = 10 ， 在每个嵌入簇内，我们计算两两之间的距离以识别重复项。当检测到重复视频时，我们选择分辨率最高的视频以确保去重过程中不会丢失任何质量。为了避免在GPU内存中存储整个两两距离矩阵，我们在块（每块256个）上实时计算必要的上三角矩阵和argmax减少操作。在去重过程中，我们大约移除了30%的训练数据。

我们还利用提取的InternVideo2嵌入和聚类结果构建了一个视觉搜索引擎，支持使用自由形式的文本和视频查询整个训练数据集。该搜索引擎对于调试数据问题以及理解预训练数据集与下游应用之间的差距非常有用。

3.6. Sharding

这个步骤旨在将处理后的视频片段打包成我们的模型训练器可以直接消费的webdatasets，用于训练。我们根据分辨率、宽高比和长度对视频进行分片，以与我们的需求保持一致。

培训课程。除了预先训练的数据集，我们还通过利用上述描述的不同过滤器创建了更高质量的微调数据集。

3.7. Infrastructure

我们的数据处理基础架构使用 AnyScale Ray（ 莫里茨等人。 ， 2017 为了实施一个流处理管道系统以支持地理上分布的集群，在大规模机器学习工作流程中解决两个关键挑战：跨同构节点的有效资源利用以及在与数据源之间存在高延迟连接的情况下保持稳健运行。通过将数据传输与计算解耦，管道能够在远程数据存储上高效运行，并且其内存需求随着管道复杂性的增加而扩展，而不是随着数据集大小的增加，从而实现无限制的流处理。

我们的架构通过并行管道阶段实现了互补硬件资源的同时利用，例如同时使用网络带宽进行数据摄入、NVDEC单元进行视频解码以及GPU进行计算密集型转换。我们扩展了Fragmentation Gradient Descent算法（ 翁等人。 ， 2023 以优化这种多资源分配为例，通过我们的调度器自动调整各个阶段，以在专用硬件加速器上维持均衡的吞吐量。

4. 令牌器

分词器是现代大规模模型的基本构建块。它们通过在无监督的方式下学习发现的一个瓶颈潜空间，将原始数据转换为更高效的表示形式。具体而言，视觉分词器将原始且冗余的视觉数据（如图像和视频）映射为紧凑的语义令牌，使其对于处理高维视觉数据至关重要。这种能力不仅使大规模变压器模型的训练变得高效，还使得在有限计算资源上进行推理变得更加普及。 Fig. 6 以示意图的方式阐述了代币化训练管道，其目标是训练编码器和解码器，使得瓶颈代币表示最大程度地保留在输入中的视觉信息。

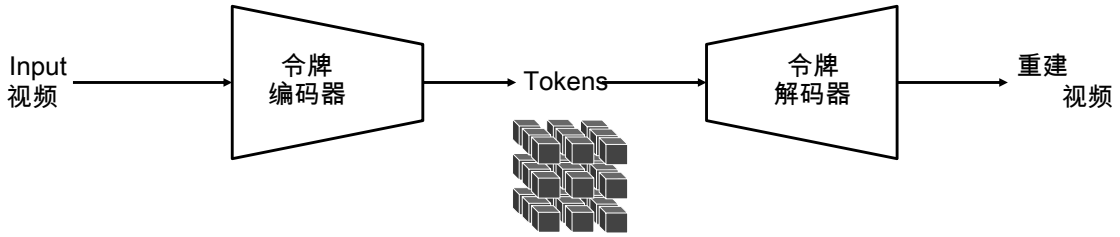
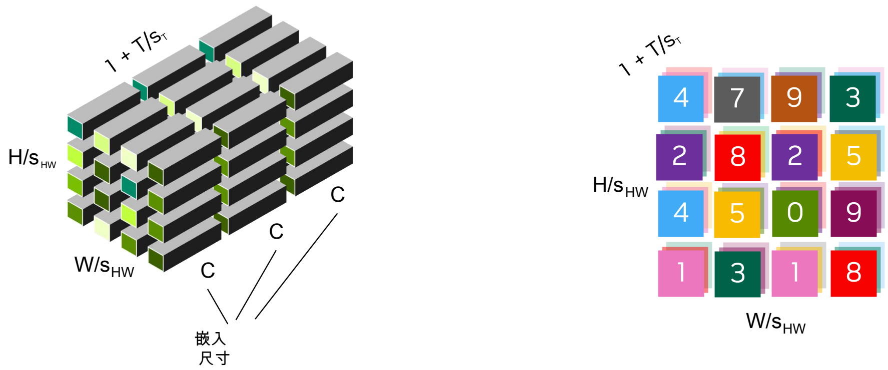


图 6: **视频标记化管道** 一段输入视频被编码成令牌，这些令牌通常比输入视频更为紧凑。解码器随后从这些令牌重构输入视频。令牌化训练旨在让编码器和解码器学习最大化保留在令牌中的视觉信息。

标记器有两种类型 ： 连续和离散 (请参见 Fig. 7 为了示例）。连续式编码器将视觉数据编码为连续的潜在嵌入，类似于潜变量扩散模型（如稳定扩散）中的处理。 Rombach 等人。 ， 2022 ）或 VideoLDM（ Blattmann 等人。 ， 2023). 这些嵌入适用于从连续分布中采样生成数据的模型。离散分词器将视觉数据编码为离散的潜在代码，并将它们映射到量化索引，类似于自回归变压器（如VideoPoet）中的做法。 Kondratyuk 等人。 ， 2024 ）。这种离散表示对于使用交叉熵损失训练的 GPT 等模型是必要的。 Fig. 7 说明了两种类型的令牌。

_tokenizer 的成功很大程度上依赖于它们能够在不牺牲后续视觉重建质量的情况下提供高压缩率的能力。一方面，高压缩率可以减少存储和计算需求，



(a) 连续代币　　　　(b) 离散令牌

图 7: **连续和离散标记器的可视化** 。沿空间的令牌 ($\frac{H}{s_{HW}} \times \frac{W}{s_{HW}}$) 和时间 ($1 + \frac{T}{s_T}$) 维度 ， 空间压缩因子为 s_{HW} 和时间压缩因子 s_T > 0 共享潜在空间中的标记化。左 ： 嵌入大小为的连续潜在嵌入 C 。第一个时间标记表示第一个输入帧 ， 启用联合图像 ($T = 0$) 和视频 (T) 。右 ： 量化索引 ，每种颜色代表一个离散的潜在代码。

在设计分词器时，这一权衡构成了一个显著的挑战，既要满足对国家需求的响应，另一方面，过度压缩可能导致关键视觉细节的丢失。

我们呈现Cosmos Tokenizer，这是一个包含图像和视频连续与离散分词器的视觉分词套件。Cosmos Tokenizer 提供卓越的视觉重构质量和推理效率，并提供多种压缩率以适应不同的计算约束和应用需求。Tab. 4 比较了不同的视觉标记器及其功能。

表 4 ： 不同视觉标记器及其功能的比较。

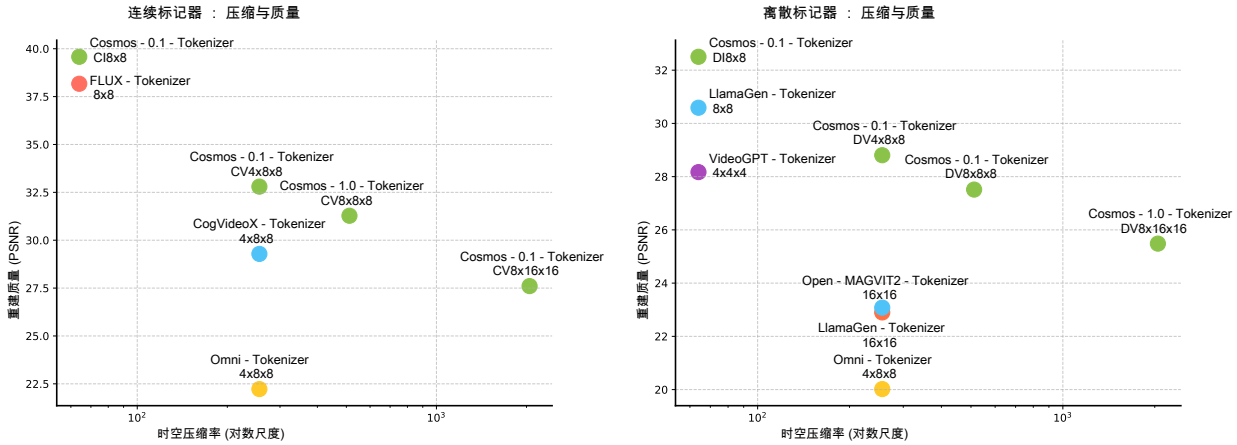
模型	因果	图像	视频	联合	离散	连续
FLUX - Tokenizer (FLUX , 2024)	-	✓	✗	✗	✗	✓
Open - MAGVIT2 - Tokenizer (Luo et al. , 2024)		✓	✗	✗	✓	✗
LlamaGen - 令牌器 (Sun 等人。, 2024)		✓	✗	✗	✓	✗
VideoGPT - Tokenizer (Yan 等人。, 2021)	✗	✗	✓	✗	✓	✗
Omni - Tokenizer (Wang 等人。, 2024)	✗	✓	✓	✓	✓	✓
CogVideoX - 令牌器 (Yang et al. , 2024)	✓	✓	✓	✓	✗	✓
Cosmos - Tokenizer	✓	✓	✓	✓	✓	✓

我们使用轻量级且计算效率高的架构设计Cosmos分词器，并采用时间因果机制。具体而言，我们运用因果时间卷积层和因果时间注意力层来保持视频帧的自然时间顺序，确保使用单一统一网络架构无缝地对图像和视频进行分词。

我们直接对高分辨率图像和长时视频进行分词器训练，而不限制类别或纵横比。与现有专注于特定数据类别和尺寸的分词器不同，Cosmos分词器可在各种纵横比（包括1:1、3:4、4:3、9:16和16:9）下运行。在推理过程中，它们对时间长度不敏感，能够对超过训练时时间长度的数据进行分词。

我们还在标准图像和视频基准数据集上评估我们的标记器 ， 包括 MS - COCO 2017 (Lin et al. ， 2014) 、 ImageNet - 1K (邓等人。 ， 2009) 和戴维斯 (佩拉齐等人。 ， 2016). 为了促进Physical AI应用中的视频分词研究，我们整理了一个涵盖多种视频类别的视频数据集，这些类别包括鱼眼视角、机器人技术、驾驶、人类活动以及空间导航。

该数据集位于 github. com / NVlabs / TokenBench 。



(a) 连续标记器　　(b) 离散标记器

图 8: **在对数尺度下，根据重建质量（PSNR）与空间-时间压缩率的比较中，连续（左）和离散（右）分词器的性能对比。** . 每个实心点代表一种分词器配置，展示了压缩率与质量之间的权衡。值得注意的是，我们的分词器在压缩率较高的情况下仍能提供卓越的质量，相较于其他方法表现出更优的质量。评估是在DAVIS数据集中进行的。我们计算了所有单独帧上图像分词器的PSNR值。

如所示 Fig. 8 我们的评估结果表明，Cosmos Tokenizer在重建质量方面显著优于现有的分词器，例如在DAVIS视频上的重建质量提高了+4 dB的PSNR。其运行速度可达12... × 可以更快地处理，单次运行可在单个配备80GB内存的NVIDIA A100 GPU上无内存耗尽地编码分辨率为1080p的视频片段最多8秒，分辨率为720p的视频片段最多10秒。一系列预训练模型还实现了空间压缩比例为8的比例。 × and 16 × ， 时间压缩因子为 4 × and 8 ×
可在 github. com / NVIDIA / Cosmos - Tokenizer .

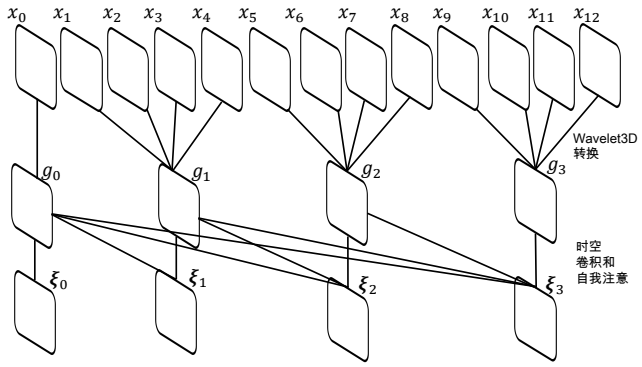
4.1. Architecture

Cosmos Tokenizer 被设计为编码器 - 解码器体系结构。给定输入视频 $\mathbf{x} \in \mathbb{R}^{(1+T)\times H\times W\times 3}$ 0: T ， with H ， W ， T 作为帧的高度、宽度和数量 ， 编码器（ $\mathcal{E}$ ）将输入标记为标记视频 $\mathbf{z}$ 0: T' $\in \mathbb{R}^{(1+T')\times H'\times W'\times D}$ ， 空间压缩因子为 $f = H/H' = W/W'$ 和一个时间解码器(), 然后从这些令牌重建输入视频 ， 结果 H' W' 的压缩因子 $f_t = T/T'$ 在重建的视频中 T' $\mathbf{x}$ $\in \mathbb{R}^{(1+T)\times H\times W\times 3}$ ， 在数学上由以下公式给出 ：

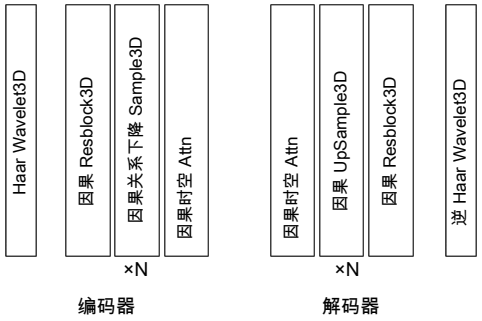
$$\hat{x}_{0:T} = \mathcal{D}\Big(\mathcal{E}\big(x_{0:T}\big)\Big). \tag{1}$$

我们的架构采用时序因果设计，确保每一阶段只处理当前和过去的帧，而不依赖于未来的帧。与常见方法不同，我们的分词器在小波空间中运行，其中输入首先经过两级小波变换。具体来说，小波变换将输入视频映射到小波空间中。 $\mathbf{x}$
0: T 以分组方式将输入向下采样 4 倍 $\mathbf{x}_0$ ， $\mathbf{x}_1$ ，and $\mathbf{x}_2$ 。这些组的组成如下 ： $\{\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\} \to \{\mathbf{z}_0, \mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_{T'}\}$
0 1:4 5:8
(T– 0 1 2 . 后续编码器阶段过程
3): T T' 4
帧在时间上的因果关系为 $\{, , ,...\} \to \{, ,...\}$
0 0:1 0:2 0 1 2 。连续的编码器阶段遵循类似的方案 ， 最终输出令牌 $\mathbf{z}$
0: T' 。因果设计有助于适应构建在顶部的模型
′
分词器处理下游物理AI应用中的时间因果设置。小波变换使我们能够在更紧凑的视频表示中操作，消除像素信息中的冗余，从而使后续层专注于更具语义性的压缩。

我们的编码阶段（波let变换之后）通过一系列残差块与下采样块交错实现。在每个块中，我们采用一种时空因子化的3D卷积，在此之前，我们首先应用一个核大小为1的2D卷积。 $\times$ 号/ $\times$ 号/ 捕获空间信息 ， 然后进行时间卷积 ， 内核大小为 号/ $\times$ 1 $\times$ 1 来捕获时间动态。我们使用左填充 号/ − 1 为了确保因果关系，在捕捉长程依赖关系时，我们利用了具有全局支持区域的空间-时间因子化因果自注意力机制——例如，1 + 差/ ′ 对于最后一个编码器块。我们使用 Swish 激活函数（ Ramachandran 等人。 ， 2017 ）用于非线性。我们利用层归一化 (LayerNorm)（ 雷巴等人。 ， 2016 ）而不是组规范化 (GroupNorm)（ 吴和他 ， 2018 ） ，从而防止在潜在空间或重建输出的特定区域中出现大的幅度（ Karras 等人。 ， 2020 ；萨达特等人。 ， 2024 ）。解码器镜像编码器，用上采样块替换下采样块。 Fig. 9 描述了 Cosmos Tokenizer 整体架构的概述。



(A) 时间因果关系 ： 时间因果关系机制的说明 ， 其中输入 号，号，. . . ，号 通过分组处理
0 1 12
中间输出 忍/，忍/，... ， 并通过时空进一步细化
0 1
卷积和注意力操作。



(b) 网络架构：编码解码网络结构包括3Dhaar 小波、因果残差、因果下采样和因果空时注意力块。解码器镜像编码器的结构，将下采样替换为上采样。

图 9: **总体而言，Cosmos Tokenizer架构展示了时间因果关系与编码器-解码器结构的集成。** 时间因果关系（左侧）处理序列输入，而编码器-解码器（右侧）利用小波变换和因果操作来捕获数据中的空间和时间依赖性。

我们采用经典的自动编码器（AE）形式来建模连续分词器的潜在空间。对于离散分词器，我们采用了有限标量量化（FSQ）方法。 Mentzer 等人。 ， 2023 作为潜在空间量化器。连续标记器的潜在维度为16，而离散标记器的潜在维度为6，这代表了FSQ级别的数量，即8。 ，8 ，8 ，5 ，5 ，5) 。此配置对应的词汇量为 64 ，000.

4.2. 培训策略

我们采用交替训练策略，按照预设的频率交替使用图像和视频的迷你批次进行训练。我们仅监督分词器解码器最终输出的结果，而不使用辅助损失，例如潜空间中的承诺损失或KL先验损失。例如，如果是一个变分自编码器（VAE） ， 金马 ， 2013 ）在使用了特定的生成方法（如VQ-VAE）而不是基础的AE（自动编码器）进行连续标记化时，就需要包含KL先验损失。 范登 · 奥德等人。 ， 2017 ）用于离散量化 ， 而不是 FSQ ，则需要有承诺损失。

我们采用两阶段训练方案。在第一阶段，我们使用L1损失来最小化输入视频与重建视频在像素级别的RGB差异（ ）。 号 ^
0: 号/ ） ， 由

$$\mathcal{L}_1 = \|\hat{x}_{0:T} - x_{0:T}\|_1 , \tag{2}$$

以及基于 VGG - 19 特征的感知损失 (Simonyan 和 Zisserman ， 2014) ，

$$\mathscr{L}_{\text{感知}} = \frac{1}{C_i H_i W_i} \left\| \text{VGG}(\hat{x}) - \text{VGG}(x) \right\|_1, \tag{3}$$

其中 $\text{VGG}(\cdot) \in \mathbb{R}^{C_i \times H_i \times W_i}$ 是来自 i 预训练的 VGG-19 网络的第 4 层是考虑的层数 ，并且 $C_i H_i W_i$ 是第 - 层的重量。

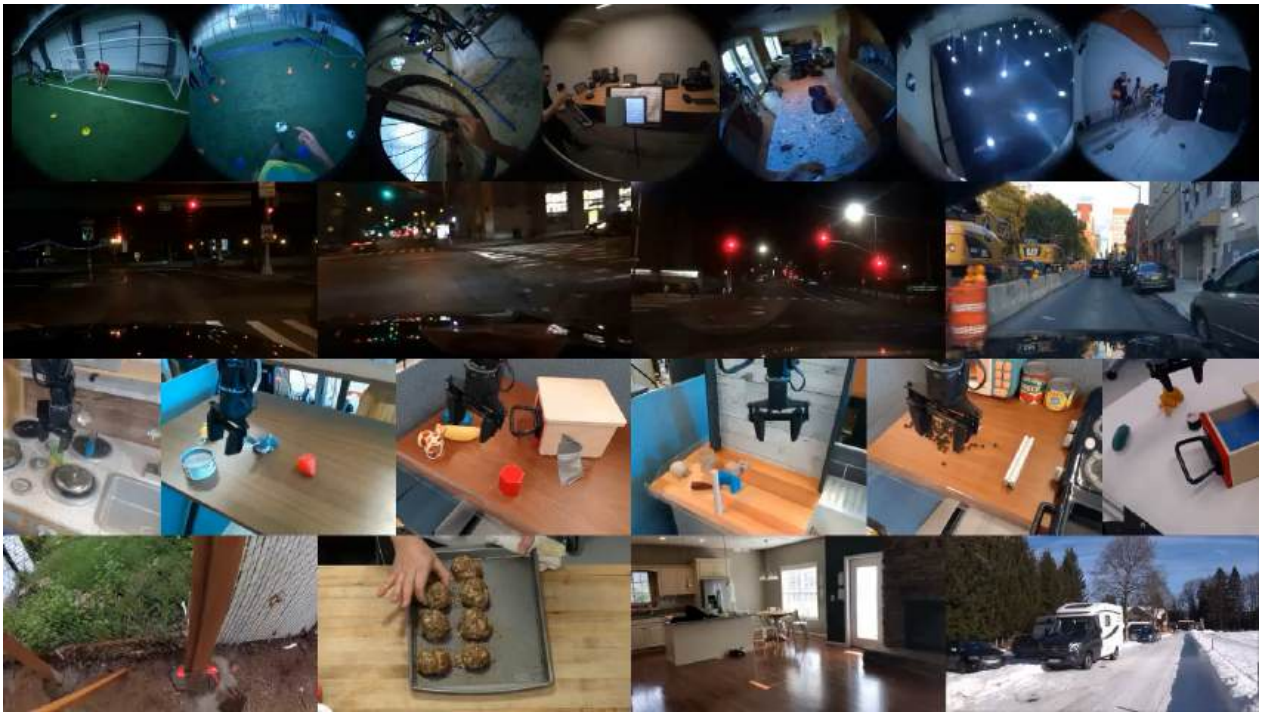


图 10: **示例视频来自** *TokenBench*。该图显示了不同的例子 ， 包括自我中心 ， 驾驶、机器人操纵和网络视频。

在第二阶段 ， 我们使用光流 (OF) 损耗 (Teed 和 Deng ， 2020) 来处理重建视频的时间平滑度 ，

$$\mathscr{L}_{\text{Flow}}(\hat{x}, x) = \frac{1}{T-1} \frac{1}{H} \frac{1}{W} \left\| \text{OF}(\hat{x}_t, \hat{x}_{t+1}) - \text{OF}(x_t, x_{t+1}) \right\|_1,$$

和革兰氏矩阵 (GM) 损失 (Gatys 等人。， 2016) 来增强重建图像的清晰度 ，

$$\mathscr{L}_{\text{Gram}} = \frac{1}{C_i C_i} \left\| \text{GM}(\hat{x}) - \text{GM}(x) \right\|_1. \tag{4}$$

此外，在微调阶段我们使用对抗损失进一步增强重建细节，特别是在高压缩率情况下。

我们训练图像分词器 (表示为 **CI** and **DI**) 在两个压缩率下 ： 8×8 and 16×16. 同样 ， 我们训练视频分词器(表示为 **CV** and **DV**) 在三个压缩率下 ： $4 \times 8 \times 8$, $8 \times 8 \times 8$, and $8 \times 16 \times 16$. 在这里 ， 压缩率表示为 $t \times h \times w$ 对于图像和 $1 + t \times h \times w$ 对于视频 ， 在哪里 1 表示时间维度 ， 并且 t and h 表示空间维度。

对于视频分词器 ， 我们创建两个变体 ：

表 5: DAVIS 和 TokenBench 上的连续视频 (CV) 标记器的评估。

令牌器	框架	公式	PSNR	SSIM	rFVD	DAVIS PSNR	DAVIS SSIM	DAVIS rFVD	TokenBench PSNR	TokenBench SSIM	TokenBench rFVD
CogVideoX - Tokenizer	$4\times8\times8$	17	VAE	29.29	0.864	19.58	32.06	0.909	6.97
Omni - Tokenizer	$4\times8\times8$	17	VAE	22.23	0.713	117.66	24.48	0.830	35.86
Cosmos - 0.1 - Tokenizer - CV4	$4\times8\times8$	49	AE	**32.80**	**0.900**	**15.93**	**35.45**	**0.928**	**6.85**
Cosmos - 0.1 - Tokenizer - CV8	$8\times8\times8$	49	AE	30.61	0.856	30.16	34.44	0.917	11.62
Cosmos - 0.1 - Tokenizer - CV8	$16\times8\times8$	49	AE	27.60	0.779	93.82	31.61	0.875	43.08
Cosmos - 1.0 - Tokenizer - CV8	$8\times8\times8$	121	AE	31.28	0.868	23.49	35.13	0.926	9.82

表 6: DAVIS 和 TokenBench 上的离散视频 (DV) 标记器的评估。

令牌器	帧	量化	PSNR	SSIM	rFVD	DAVIS PSNR	DAVIS SSIM	DAVIS rFVD	TokenBench PSNR	TokenBench SSIM	TokenBench rFVD
VideoGPT - Tokenizer	$4\times4\times4$	VQ	28.17	**0.850**	72.33	**33.66**	**0.914**	**13.85**			
Omni - Tokenizer	$4\times8\times8$	17 VQ	20.02	0.703	188.60	25.31	0.827	53.55			
Cosmos - 0.1 - Tokenizer - DV4	$4\times8\times8$	17 FSQ	**28.81**	0.818	**37.36**	31.97	0.888	19.67			
Cosmos - 0.1 - Tokenizer - DV8	$8\times8\times8$	17 FSQ	27.51	0.789	100.15	30.95	0.873	43.86			
Cosmos - 0.1 - Tokenizer - DV8	$16\times8\times8$	17 FSQ	25.09	0.714	241.52	28.91	0.829	113.48			
Cosmos - 1.0 - Tokenizer - DV8	$16\times8\times8$	49 FSQ	25.49	0.719	259.33	29.33	0.838	107.43			

1. **Cosmos - 0.1 - Tokenizer** ： 使用小型批处理对较少数量的视频帧进行采样(49 帧用于 **CV** 和 17 帧 **DV**).
2. **Cosmos - 1.0 - Tokenizer** ： 使用小型批处理对较大数量的视频帧进行采样(121 帧 ， 用于 **CV** 和 49 帧 **DV**).

这种方法确保了处理图像和视频数据的变化的时间和空间分辨率的灵活性。

4.3. Results

我们广泛评估了我们的Cosmos Tokenizer套件在各种图像和视频基准数据集中。对于图像Tokenizer的评估，我们遵循先前的研究对MS-COCO 2017进行评估（ ）。 Lin et al. ， 2014 ）和 ImageNet - 1K （ 邓等人。 ， 2009 ）。我们使用 MS - COCO 2017 验证子集 5 ， 000 个图像和 50 个 ImageNet - 1K 验证子集 ， 000 幅图象作为图象评价基准。

TokenBench. 对于视频分词器评估而言，目前尚无针对高分辨率和长时长视频的标准基准。为此，我们引入了一个名为 *TokenBench* 为了涵盖广泛的领域，包括机器人操作、驾驶、第一人称视角和网络视频，并标准化评估方法。我们采用了常用于各种任务的现有视频数据集，如BDD100K。 Yu 等人。 ， 2020 ）， EgoExo - 4D （ Grauman et al. ， 2024 ）， BridgeData V2 （ Walke 等人。 ， 2023 ）和熊猫 - 70M （ Chen et al. ， 2024 我们从每个数据集随机抽取100个视频，并通过取前10秒并调整短边大小至1080来进行预处理。对于Panda-70M，我们手动筛选掉内容质量低和运动量小的视频。对于EgoExo-4D，我们随机挑选100个场景，并从中分别选取一个主观视角视频和一个客观视角视频。这样总共得到500个视频。以下是一些示例视频的例子： *TokenBench* 可以在 Fig. 10 . We release *TokenBench* 在 github. com / NVlabs / TokenBench .

In addition to *TokenBench* ， 我们还在 1080p 分辨率的 DAVIS 数据集上评估我们的视频标记器。

基线和评估指标。 我们评估各种压缩率下的分词器以展示其在不同计算需求下的有效性，并将其与最先进的图像和视频分词器进行比较。 Tab. 4 呈现了我们在各种设置中对比的具体最新最先进（State-of-The-Art，SOTA）分词器。评估指标包括峰值信噪比（Peak Signal-to-Noise Ratio，PSNR）、结构相似性（Structural Similarity，SSIM）。

表 7: 对各种图像数据集上的连续图像 (CI) 标记器的评估。

令牌器	高度	公式	PSNR	SSIM	rFID	PSNR	SSIM	rFID
			(a) MS - COCO 2017			(b) ImageNet - 1K		
			↑	↑	↓	↑	↑	↓
FLUX - Tokenizer	8	- VAE	24.00	0.682	2.501	20.09	0.518	1.229
Cosmos - 0.1 - Tokenizer - CI	8	1024 AE	**28.66**	**0.836**	**1.760**	**28.83**	**0.837**	**0.689**
Cosmos - 0.1 - Tokenizer - CI	16	1024 AE	23.63	0.663	3.823	23.72	0.655	1.031

表 8: 各种图像数据集上的离散图像 (DI) 标记器的评估。

令牌器	高度	量化	PSNR	SSIM	rFID	PSNR	SSIM	rFID
			(a) MS - COCO 2017			(b) ImageNet - 1K		
			↑	↑	↓	↑	↑	↓
Open - MAGVIT2 - Tokenizer	16	- LFQ	19.50	0.502	6.649	17.00	0.398	2.701
LlamaGen - Tokenizer	8	- VQ	21.99	0.616	4.123	19.64	0.498	1.403
LlamaGen - Tokenizer	16	- VQ	19.11	0.491	6.077	18.38	0.448	1.657
Cosmos - 0.1 - Tokenizer - DI	8	1024 FSQ	**24.40**	**0.704**	**3.710**	**24.48**	**0.701**	**1.265**
Cosmos - 0.1 - Tokenizer - DI	16	1024 FSQ	20.45	0.529	7.234	20.49	0.518	2.518

表 9: 分词器的运行时性能比较每个图像或每个视频帧报告时间。

令牌器	类型	解析帧	参数	时间 (ms)
FLUX - Tokenizer	8	连续图像 1024 $\times$ 1024	- 84M	242
Cosmos - 0.1 - Tokenizer - CI	8	连续图像 1024 $\times$ 1024	- 77M	**62.7**
LlamaGen - Tokenizer	8	离散图像 1024 $\times$ 1024	- 70M	475
Cosmos - 0.1 - Tokenizer - DI	8	离散图像 1024 $\times$ 1024	- 79M	**64.2**
CogVideoX - Tokenizer	4 $\times$ 8 $\times$ 8	连续视频 720 $\times$ 1280	17	216M 414
Omni - Tokenizer	4 $\times$ 8 $\times$ 8	连续视频 720 $\times$ 1280	17	54M 82.9
Cosmos - 0.1 - Tokenizer - CV	4 $\times$ 8 $\times$ 8	连续视频 720 $\times$ 1280	49	105M **34.8**
Omni - Tokenizer	4 $\times$ 8 $\times$ 8	离散视频 720 $\times$ 1280	17	54M 53.2
Cosmos - 0.1 - Tokenizer - DV	4 $\times$ 8 $\times$ 8	离散视频 720 $\times$ 1280	17	105M **51.5**

重建 Fr é chet 初始距离 (rFID) (Heusel 等人。 , 2017) 用于图像 , 并重建 Fr é chet 视频距离 (rFVD) (U nterthiner 等人。 , 2019) 用于视频。

定量结果。 Tabs. 5 and 6 总结各种基准上的连续和离散视频分词器的平均定量指标。如表中所示，Cosmos Tokenizer 在所有指标上均达到了最先进的性能，相较于先前的方法，在DAVIS视频数据集上也表现优异。 *TokenBench* , 时空压缩比为 4 $\times$ 8 $\times$ 8. 此外 , 即使有 2 $\times$ and 8 $\times$ 更高的压缩比 (*i.e* ., 8 $\times$ 8 $\times$ and 8 $\times$ 16 $\times$ 16), 元宇宙Tokenizer仍优于现有技术，展示了卓越的压缩-质量trade-off。

Tabs. 7 and 8 总结各种图像基准上的连续和离散图像分词器的平均量化指标，覆盖了广泛的图像类型。如所示，与以往方法相比，Cosmos分词器始终能够以8倍的压缩比达到最先进的结果。 $\times$ 8. 更重要的是 , 在 4 $\times$ 更大的压缩比 16 $\times$ 如图 16 所示 , Cosmos Tokenizer 的图像质量通常与 8 时的现有技术相当甚至更好 $\times$ 8 压缩比 , 如 Tabs. 7 and 8 .

这些在多种图像和视频基准数据集上的定量结果证实，Cosmos Tokenizer能够以较大的空间-时间压缩更好地表示视觉内容。

运行时性能。 Tab. 9 显示了每张图像或每个视频帧的参数数量以及平均编码和解码时间，测量基于单个A 100 80GB GPU。相比之下，我们还列出了先前的最先进标记化器的参数数量和平均速度。如所示，对于图像和视频而言：

令牌器 ， 宇宙令牌器是 2 × ~ 12 × 与先前的技术相比，在保持最小模型大小的同时实现了更快的速度，这表明Cosmos Tokenizer在编码和解码视觉内容方面具有高效性。

5. 世界基金会模型预培训

预训练的物理行为模型（WFMs）是一类通用模型，能够捕捉现实世界物理和自然行为的一般知识。我们利用两种不同的可扩展深度学习范式——扩散模型和自回归模型——构建了两类WFMs。这两种扩散模型和自回归模型都将一个复杂生成问题分解为一系列更易于解决的子问题，并且极大地推动了生成模型的发展。对于扩散模型而言，复杂的生成问题被分解为一系列去噪问题；而对于自回归模型，则被分解为一系列下一个标记预测问题。我们在构建预训练WFMs的过程中，探讨了如何使用各种针对现代GPU进行并行化的技术来扩展这些深度学习范式。我们使用包含10台机器的集群对论文中报告的所有WFMs模型进行了训练。 ，000 个 NVIDIA H100 GPU ， 时间为三个月。

表10：Cosmos World基金会模型1.0发布地图。我们有两个WFMs（World Foundation Models）集合。一个基于扩散模型，另一个基于自回归模型。对于每个家庭，我们构建了两个基础模型和两个衍生模型。为了获得最佳生成质量，我们还为扩散模型构建了一个提示上采样器，并为自回归模型构建了一个扩散解码器。

Type	模型		Tokenizer	增强器
扩散	Cosmos - 1.0 - 扩散 - 7B - Text2World	→ Cosmos - 1.0 - 扩散 - 7B - Video2World	Cosmos - 1.0 - 令牌 - CV8x8x8	Cosmos - 1.0 - PromptUpsampler - 12B - Text2World
	Cosmos - 1.0 - 扩散 - 14C - Text2World	→ Cosmos - 1.0 - 扩散 - 14C - Video2World		
自回归	Cosmos - 1.0 - 自动回归 - 4B	→ Cosmos - 1.0 - 自动回归 - 5B - Video2World	Cosmos - 1.0 - 令牌 - DV8x16x16	Cosmos - 1.0 - 扩散 - 7B - 解码器 - DV8x16x16ToCV8x8x8
	Cosmos - 1.0 - 自动回归 - 12B	→ Cosmos - 1.0 - 自动回归 - 13B - Video2World		

In Tab. 10 我们展示了一张预训练的WFMs及其伴侣模型的架构图。对于基于扩散的WFM家族，我们首先构建了两个分别拥有7B和14B参数量的Text2World模型，分别命名为Cosmos-1.0-Diffusion-7B-Text2World和Cosmos-1.0-Diffusion-14B-Text2World。这些模型能够将文本提示映射到视觉世界的视频中。随后，我们将Text2World模型进行微调，使其能够接受额外的视频输入，以代表当前观察结果。最终，我们得到一个Video2World模型，在该模型中，未来的视频是基于当前观察（输入视频）和扰动（文本提示）进行预测的。这些扩散模型属于潜变量扩散模型，接受连续的标记作为输入。我们使用Cosmos-1.0-Tokenizer-CV8x8x8来生成视觉标记。WFMs的训练文本提示是由VLM通过视频描述生成的，这些描述遵循人类对视频的不同描述分布。为了缓解领域差异，我们基于Mistral-NeMo-12B-Instruct模型构建了Cosmos-1.0-PromptUpsampler-12B-Text2World。 Mistral 和 NVIDIA ， 2024 ） ， 以帮助将人类文本提示转换为我们基于扩散的 WFM 喜欢的提示。

对于基于自回归的WFM家族模型，我们首先构建两个基础模型，分别大小为4B和12B，以纯粹根据当前视频观察来预测未来视频。我们将它们分别命名为Cosmos-1.0-Autoregressive-4B和Cosmos-1.0-Autoregressive-12B。这些模型是Llama3风格的GPT模型，经过训练。

从头开始为视频预测任务进行构建，并且不涉及语言理解。为了使基于自回归的WFMs能够利用文本信息来进行下一个令牌的预测，我们通过交叉注意力层将T5嵌入输入文本提示的方式整合到WFMs中。这些自回归WFMs使用Cosmos-1.0-Tokenizer-DV8x16x16，将输入视频映射为少数整数。分词器的大量压缩有时会导致不期望的失真。为了解决这个问题，我们通过微调Cosmos-1.0-Diffusion-7B-Text2World模型建立了一个扩散解码器（Cosmos-1.0-Diffusion-7B-Decoder-DV8x16x16ToCV8x8x8），该模型将DV8x16x16空间中的离散令牌映射到CV8x8x8空间中的连续令牌。

5.1. 基于扩散的世界基金会模型

我们的基于扩散的视频生成模型（WFMs）是潜空间中的扩散模型，能够在学习到的分词器的潜空间中运行，从而提供视频的紧凑、低维表示。这种设计选择带来了多方面的优势：它在训练和推理过程中都减少了计算成本，并简化了去噪任务。 Hoogeboom 等人。 ， 2024 ； Rombach 等人。 ， 2022 ）。为了将视频标记为潜在表示 ， 我们采用了 Cosmos - 1.0 - Tokenizer - CV8x8x8 。

5.1. 1. Formulation

为了训练我们的扩散 WFM ， 我们采用 EDM 中概述的方法（ Karras 等人。 ， 2022 ， 2024 ）。去噪器的去噪分数匹配损失 $\mathcal{L}$ ， 在噪声水平下评估 σ ， 定义为

$$\mathcal{L}(D_\theta, \sigma) = \mathbb{E}_{x,n}\left[\|D_\theta(x+n;\sigma) - x\|_2^2\right] \quad (5)$$

where $x \sim p_{\text{data}}$ $n \sim \mathcal{N}(0, \sigma^2 I)$ 是从训练集中采样的干净图像或视频， 是 i. i. d. 高斯噪音 ， 以及 D_θ 是一个噪声条件神经网络 ， 其任务是对损坏的样本 x + n 进行去噪 。我们坚持 EDM 中引入的用于参数化的预处理设计 D_θ 。总体训练损失定义为加权期望 $\mathcal{L}(D_\theta; \sigma)$ 在噪声水平 ：

$$\mathcal{L}(D_\theta) = \mathbb{E}_{\sigma}\left[\lambda(\sigma)\,\mathcal{L}(D_\theta, \sigma)\right], \quad (6)$$

$$\lambda(\sigma) = \frac{\sigma^2 + \sigma_{\text{data}}^2}{(\sigma \cdot \sigma_{\text{data}})^2}, \quad (7)$$

其中噪声水平的分布 $p(\sigma)$ 由超参数控制， and 是标准 p_{mean} p_{std} . σ_{data} 训练数据的偏差和加权函数 $\lambda(\sigma)$ 确保在训练开始时每个噪声级别贡献相等。然而，随着训练的进行，这种平衡可能会恶化。为了缓解这一问题，我们将不同噪声级别的优化视为一种多任务学习的形式。我们通过引入基于不确定性加权的方法来利用这一点。 $u(\sigma)$ 作为一个连续的不确定性函数 ， 量化去噪目标的不确定性 $\mathcal{L}(D_\theta, \sigma)$ 在噪声水平 σ 我们使用一个简单的 MLP 来参数化 $u(\sigma)$ 并将整体损失降至最低 $\mathcal{L}(D_\theta)$ σ 在训练过程中，直观上来说，如果模型对任务不确定，那么噪声级别下的损失贡献会被权重降低。 *i.e* . , if $u(\sigma)$ 较高。同时 ， 该模型因这种不确定性而受到惩罚 ， 鼓励 $u(\sigma)$ 尽可能低。

与采用高斯流匹配公式的最新视频生成模型相比（ Kong 等人。 ， 2024 ； Polyak 等人。 ， 2024 ） ， 我们的工作是从扩散分数匹配的角度（ Ho 等人。 ， 2020 ； Song et al. ， 2020 ）。然而 ， 如 Gao 等人。 （ 2024 ） ，这些框架从理论上说是等价的，共享基本相似的目标和训练程序。基于EDM的方法与这些见解保持一致，主要差异在于预处理设计和超参数的选择。在实际应用中，我们尚未遇到任何基于EDM的方法的性能限制。

5.1. 2. Architecture

在本节中 ， 我们描述了去噪器网络的设计 型
刲 建立在 DiT (皮布尔斯和谢 ，
2023), 哪个最初是为标签条件下的图像生成设计的。我们调整了其架构以更好地适应我们可控视频生成的目标。我们展示了整体网络设计的可视化。 Fig. 11 .

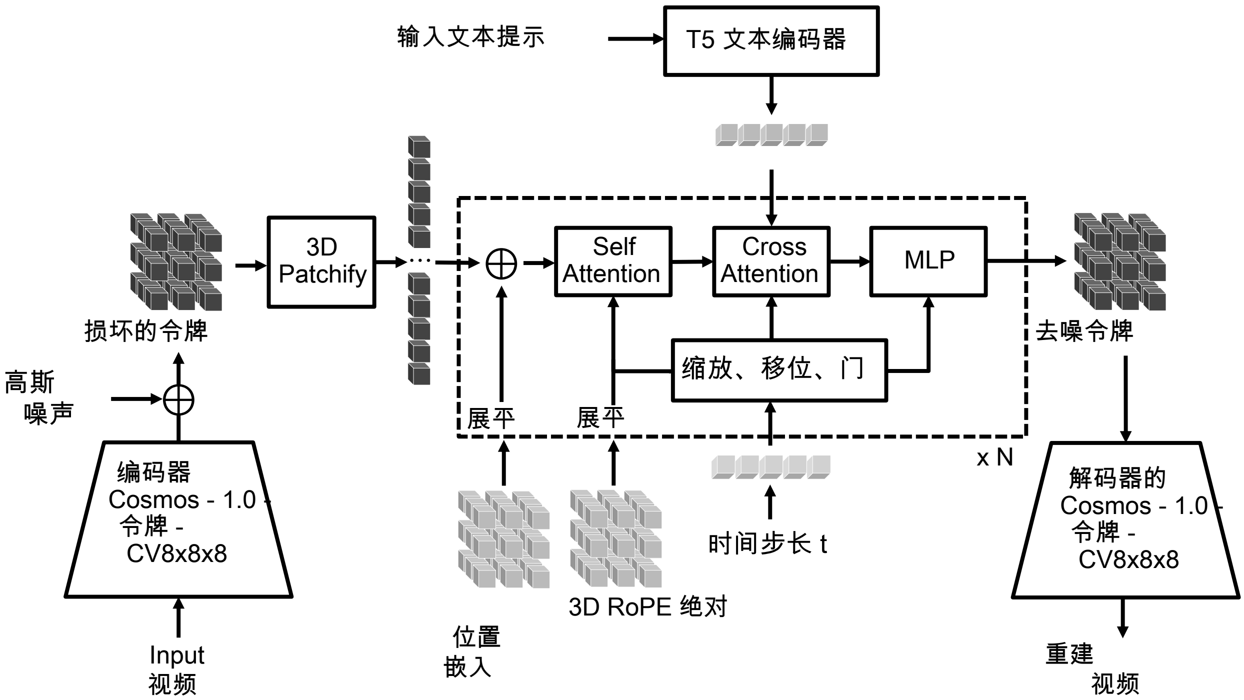


图 11: **Cosmos - 1.0 - 扩散世界基础模型的整体架构** 该模型通过Cosmos-1.0-Tokenizer-CV8x8x8的编码器处理输入视频，以获得潜在表示，随后对这些表示施加高斯噪声。然后使用3D片段化过程进行转换。在潜在空间中，架构应用重复的自注意力、跨注意力（整合输入文本）和前向MLP层，在给定的时间步长上由自适应层规范化（缩放、平移、门控）进行调节。 孬 。 Cosmos - 1.0 - Tokenizer - CV8x8x8 的解码器从细化的潜在表示中重建最终的视频输出。

3D 修补。 我们网络的输入是形状的潜在表示 刲 × 郌 × 型 × 刯 对于图像和视频数据两者，其中图像通过单帧的视频进行区分。为了准备我们的去噪网络的输入，我们首先使用线性层对状态进行“切块”，随后将其展平。这一过程涉及将形状为（ ）的非重叠立方体投影至特定空间。 孬，孬，孬 ）
孬ℎ普 转化为网络中的个体令牌输入。因此，在补丁化处理之后，图像或视频会被重塑为一个长度为*spatiotemporal序列的一维序列。* 刲 型刯/（ 孬 孬 孬 ） 孬 = 1 ，孬 = 孬 = 2
孬ℎ普 . 我们使用 孬 ℎ普 为我们的去噪网络。

具有 FPS 感知 3D RoPE 和可学习嵌入的混合位置嵌入。 我们采用 3D 因式分解的旋转位置嵌入 (RoPE) (Su et al. ， 2024 为了生成任意大小、纵横比和视频长度的内容。具体而言，我们将特征维度划分为三个大致相等的部分，分别沿时间轴、高度轴和宽度轴应用RoPE（ Rotary Positional Encoding ），以获取位置信息。在实践中，可以通过在每个块中直接拼接频率嵌入而不进行拆分和连接来高效地实现这一点，并且可以重用针对大型语言模型（LLMs）优化的RoPE内核。为了进一步支持不同帧率的视频合成，我们根据训练视频的帧率（Frames Per Second, FPS）重新缩放时间频率。由于RoPE的相对位置编码特性和我们的三维因子设计，这种帧率感知的设计与我们的联合图像-视频训练兼容。RoPE 的另一个优点是在渐进式训练过程中明显显现出来，当我们在改变分辨率或视频长度时，可以通过利用

神经切核 (NTK) - RoPE (Peng 和 Quesnelle ， 2023) ， 我们观察到快速的模型收敛 ， 即使在 5 ， 000训练步骤。此外，我们发现每个变压器块增加一个额外的学习绝对位置嵌入可以进一步增强模型性能，减少训练损失，并降低生成视频中的形态艺术效果。

文本条件的交叉注意。 我们依赖于网络中的交叉注意力层来整合语言信息。每个变压器块由顺序的自注意力、交叉注意力和前馈层组成。虽然自注意力作用于时空令牌，但交叉注意力使用T5-XXL整合语义上下文。 Raffel 等人。 ， 2020) 嵌入键和值 ， 实现有效的文本条件。

查询键规范化。 在训练的早期阶段，我们观察到注意力标量的增长存在不稳定性，导致注意力熵出现崩溃。我们遵循现有文献 (Dehghani et al. ， 2023 ；Esser 等人。 ， 2024 ；Wortsman 等人。 ， 2023) 以规范化查询 $\mathit{吾}$ 和键 $\mathit{型}$ 注意操作之前。我们使用均方根归一化 (RMSNorm) (Zhang 和 Sennrich ， 2019) 为我们网络中的所有自我关注和交叉关注层提供可学习的尺度。

AdaLN - LoRA. 我们发现 DiT 的自适应层归一化 (AdaLN) 层 (皮布尔斯和谢 ， 2023 ；徐等人。 ， 2019) 贡献了模型参数中的显著部分，但在FLOPs计算复杂度方面几乎不占任何比重。受W.A.L.T启发， Gupta et al. ， 2024) ， 我们实施低秩适应 (LoRA) (胡等人。 ， 2022 以分解这些层中的密集线性投影为低秩近似的方式进行分解。对于Cosmos-1.0-Diffusion-7B，这种架构优化实现了参数数量36%的减少（从11B减少到7B参数），同时在所有评估指标上保持了性能的一致性，这证明了我们高效参数设计的有效性。

表 11: Cosmos - 1.0 - 扩散模型的配置细节。

配置	7B - Text2World	14B - Text2World	7B - Video2World	14B - Video2World
层数	28	36	28	36
模型维度	4,096	5,120	4,096	5,120
FFN 隐藏尺寸	16,384	20,480	16,384	20,480
AdaLN - LoRA 维度	256	256	256	256
注意头数	32	40	32	40
键 / 值头的数量	32	40	32	40
MLP 激活	GELU			
位置嵌入	混合位置嵌入			
条件信息	文本; FPS	文本; FPS	文本; FPS; 框架	文本; FPS; 框架
基本学习率	2^{-15}	2^{-16}	2^{-15}	2^{-16}
重量衰减	0.1	0.2	0.1	0.2
学习率热身赛	线性调度器与 2		,500 次迭代	
AdamW 动量和 ϵ	$\beta_1, \beta_2 = 0.9, 0.99; \epsilon = 10^{-10}$			

5.1. 3. Training Strategy

本部分概述了用于训练跨越多种模态、分辨率、纵横比和条件输入的数据集模型的方法学。

联合图像和视频训练。 为了利用模型训练中大量高质量、多样化的图像数据集的丰富资源，我们实施了一种交替优化策略，该策略交错使用图像和视频数据批次。为了促进图像和视频领域之间的跨模态知识转移，我们采用了一种特定领域的归一化方案，通过独立估计图像和视频数据的足够统计量来对齐潜在分布。这一方法受到以下观察的启发：减少图像和视频潜在表示之间的分布差异可以提高生成质量。此外，我们发现在视频潜在表示的时序和通道维度上存在非平稳统计量。为了解决这种异质性，我们采用了在帧级和通道级上应用的归一化策略。

标准化到视频潜表示，从而有效地鼓励它们更好地逼近各向同性的高斯先验分布。

超越跨模态知识迁移，我们的规范化方案提供了重要的理论优势：在训练过程中信号噪声比的尺度不变性。考虑两个零均值的潜在表示，它们具有不同的尺度：一个是标准差为1的，另一个是标准差为2的。当添加高斯噪声时，这两个表示在不同尺度下仍能保持相同的信号噪声比。 $\mathcal{N}(0, \sigma^2)$ 要实现标准化表示的所需信噪比，我们必须将噪声缩放到 $\mathcal{N}(0, 4\sigma^2)$ 为了保持相同的比率，采用未归一化表示。通过标准化所有潜在表示，我们确保不同尺度下的信噪比一致，从而在训练过程中即使基础分词器更新也能促进模型适应。

为了保持计算效率，我们平衡了图像和视频批次大小，以确保在GPU上实现相似的内存利用率。然而，我们观察到视频批次去噪损失的收敛速度慢于图像批次损失。我们将这一现象归因于视频帧中固有的时间冗余性，这导致了视频批次的梯度幅度较小。受到最近多分辨率图像训练进展的启发（引用原文：Drawing inspiration from recent advances in multi-resolution image training），我们计划探索不同的优化策略来加速视频批次的收敛。 Atzmon 等人。 ， 2024 ； Chen ， 2023 ；Hoogeboom 等人。 ， 2023 我们通过将视频批次噪声水平与帧计数相对应地调整为图像批次噪声水平的平方根来解决这一收敛性差异。

表 12: 渐进式训练的阶段及其规格。

舞台	分辨率帧数上下文长度 FSDP 大小 CP 大小				
低分辨率预训练 512p (640	×512)	57	10,240 [a]	64	2
高分辨率预训练 720p (1280	×704)	121	56,320 [b]	64	8
高品质微调	720p (1280 ×704)	121	56,320 [b]	64	8

[a] 10 ， 240(上下文长度) 计算为 ： 640(宽度) ÷ 8 (标记化) ÷ 2 (patchify) × 512(高度) ÷ 8 (标记化) ÷ 2 (patchify) × [(57 − 1) ÷ 8 + 1](标记帧) 。
[b] 56 ， 320(上下文长度) 计算为 ： 1280(宽度) ÷ 8 (标记化) ÷ 2 (patchify) × 704 (高度) ÷ 8 (标记化) ÷ 2 (patchify) × [(121 − 1) ÷ 8 + 1](标记帧) 。

渐进式训练。 我们采用渐进式培训策略 ， 每个阶段的细节在 Tab. 12 初始阶段包括对分辨率为512像素的视频和图像进行训练，使用由57帧组成的视频。随后，我们转向目标分辨率720像素，并将视频长度增加至121帧。在大规模数据预训练后，我们在高质量子集上对模型进行微调，以进一步优化性能。 $\mathcal{N}(10, \sigma^2)$ 具有线性衰减学习率的迭代。与来自 Dai 等人。 （ 2023 ） ， 我们还发现微调可以提高生成视频的质量。

多方面的培训。 为了适应不同纵横比的内容，我们将数据组织成五个不同的桶，对应于1:1、3:4、4:3、9:16和16:9的比例，将每张图片或视频分配到最接近其纵横比的桶中。在训练过程中，每个数据并行处理组从一个桶中采样数据，允许不同的并行处理组使用不同的桶。我们实施最长边重缩放以最大程度地保留原始内容信息，这些信息描述在提示中。对于批量处理，我们应用反射填充缺失像素，并提供填充掩码给扩散主干，从而在推理过程中实现精确控制。

混合精度训练。 我们保留了两种模型权重的副本：一种在BF16中，另一种在FP32中。在前向和反向传播过程中，使用BF16权重以提高训练效率，导致梯度和激活也采用BF16格式。对于参数更新，权重在FP32中进行更新以确保数值稳定性。随后将更新后的FP32参数复制并转换为BF16，用于下一次迭代。为了进一步稳定训练，我们在去噪评分匹配损失上进行缩放，以优化训练过程。 Eq. (5) 通过10倍的因素。我们还发现，在AdamW中较低的β和eps系数能够显著减少损失峰值。对于我们的14B扩散模型训练，我们很少遇到损失峰值，也没有不可恢复的损失峰值。

文本条件化。 对于我们的 Text2World 模型 ， 我们采用 T5 - XXL (Raffel 等人。 ， 2020) 作为文本编码器。我们通过零填充 T5 嵌入来保持固定序列长度为 512。为了增强文本-上下文对齐，我们采用无分类器引导 (classifier-free guidance) 方法。 Ho 和 Salimans ， 2022). 不像以前的作品 (Balaji et al. ， 2022 ;Saharia 等人。 ，
2022 在随机消除文本嵌入的情况下，由于推理过程中否定提示的有效性，我们省略了这一步。值得注意的是，作为文本到图像生成器，我们的模型在无需指导的情况下生成高质量图像的能力突出，这一能力归因于高质量训练数据集的使用。虽然无分类器指导通常会促进对偏好视觉内容的模式寻找行为，但我们发现精心选择的数据可以达到类似的效果。然而，在视频生成方面，缺乏可比较的高质量数据导致在低指导设置下得到次优结果。因此，在视频生成任务中，更高的指导值被要求以产生满意的内容。

图像和视频调理。 我们将Text2World模型扩展到构建Video2World模型，这些模型支持通过将先前帧 (如果有的话) 整合到生成过程中来实现图像和视频条件处理。具体来说，条件帧与生成的帧沿时间维度进行连接。为了提高推理时输入帧变化的鲁棒性，我们在训练期间向条件帧引入增强噪声。在训练过程中，用于这种增强噪声的sigma值通过采样确定。 $\mathcal{N}_{\text{mean}} = -3.0, \mathcal{N}_{\text{std}} = 2.$ 此外，扩散模型的输入在通道维度上与一个二进制掩码连接，该掩码区分条件帧和生成帧。损失函数排除了条件帧位置的贡献，专注于生成输出。为了提高泛化能力，在训练过程中随机变化条件帧的数量。在推理过程中，模型可以根据需要灵活地使用单个条件帧 (图像) 或多个先前帧作为输入。

5.1. 4. 放大

在这里，我们概述了使我们的扩散WFMs高效扩展的技术方法。我们分析了模型的内存需求，讨论了并行策略，并将我们的训练设置与其他视频扩散模型以及最先进的LLM进行了比较。

内存要求。 消耗 GPU 内存的四个主要组件是 ：

- **模型参数** ：每个参数占用10字节。我们的混合精度训练将模型参数同时存储为FP32和BF16，并将指数移动平均 (EMA) 权重也存储为FP32。• **梯度** ： 每个参数 2 个字节。我们将梯度存储在 BF16 中。• **优化器状态** ：每个参数 8 个字节。我们使用 AdamW (Loshchilov 和 Hutter ， 2019) 作为我们的优化器 ， 并存储优化器状态 (*i.e* 。 ， 第一和第二时刻) 在 FP32 中。• **激活** : (2 $\times$ number _ of _ layers $\times$ 15 $\times$ seq _ len $\times$ batch _ size $\times$ d _ model) 字节。我们将激活存储在 BF16 中。 Tab. 13 提供了网络中主要操作的激活存储详细信息。为了优化内存使用，我们实施了选择性的激活检查点 (activations checkpointing) 。 Chen et al. ， 2016 ; Korthikanti 等人。 ， 2023) ， 重新计算内存有限层的激活 ， 如归一化函数。

例如，我们的14B模型 (Cosmos-1.0-Diffusion-14B-Text2World) 需要约280 GB用于模型参数、梯度和优化器状态，以及310 GB用于高分辨率预训练时的激活值。考虑到NVIDIA H100 GPU的80GB HBM3内存限制，我们采用全分片数据并行 (FSDP) 和上下文并行 (CP) 来跨多个GPU分布内存需求。

完全分片数据并行 (FSDP) 。 FSDP 通过在设备间拆分模型参数、梯度和优化器状态来提高内存效率。在计算过程中仅在需要时收集参数，并在其后释放它们。与标准的数据并行不同，标准数据并行会在所有设备上复制参数，而 FSDP 则将参数、梯度和优化器状态进行分布处理，每个设备只管理其自身的部分。这种方法将内存使用量降至最大临时未拆分参数集及其参数、梯度和优化器状态的部分。对于我们的实现，我们为 7B 模型使用 32 的拆分因子，为 14B 模型使用 64 的拆分因子，以平衡内存使用和通信延迟。

表13：Cosmos-Diffusion变压器FLOPs和激活内存。该表提供了每种操作的计算成本（FLOPs）和激活内存要求。对于FLOPs，我们使用2作为乘累加成本的因子进行描述。“—”表示由于数值较小而忽略其影响，或者因为在使用激活检查点重新计算值而不是存储它们的情况下忽略了该值，从而节省了内存。

图层	操作	FLOP	激活 (张量形状)
自我注意	Q, K, V 预测	$2 \times 3 \times$ seq _ len $\times$ d _ model^2	seq _ len $\times$ batch _ size $\times$ d _ model[a]
	QK Norm	—	$2 \times$ seq _ len $\times$ batch _ size $\times$ d _ model[b]
	$A = Q@K^T$	$2 \times$ seq _ len $\times$ d _ model -	[c]
	A' = Softmax (A)	—	— [d]
	$A'@V$	$2 \times$ seq _ len $\times$ d _ model seq _ len	$\times$ batch _ size $\times$ d _ model[e]
	最终投影	$2 \times$ seq _ len $\times$ d _ model^2	seq _ len $\times$ batch _ size $\times$ d _ model
交叉注意	Q, K, V Projections	$2 \times$ seq _ len $\times$ d _ model^2	seq _ len $\times$ batch _ size $\times$ d _ model
	QK 范数 - seq _ len		$\times$ batch _ size $\times$ d _ model[f]
	$A = Q@K^T$	—	— [c]
	A' = Softmax (A)	—	— [d]
	$A'@V$	—	— [g]
	最终投影	$2 \times$ seq _ len $\times$ d _ model^2	seq _ len $\times$ batch _ size $\times$ d _ model
前馈	向上投影	$4 \times$ seq _ len $\times$ d _ model^2	seq _ len $\times$ batch _ size $\times$ d _ model
	GELU	—	$4 \times$ seq _ len $\times$ batch _ size $\times$ d _ model
	向下投影	$4 \times$ seq _ len $\times$ d _ model^2	— [h]
AdaLN	LayerNorm - seq _ len		$\times$ batch _ size $\times$ d _ model[i]
	规模 - -		
	移位 - -		
	Gate - seq _ len		$\times$ batch _ size $\times$ d _ model

[a] 存储共享输入。 [b] 查询 Q 和键 K 被存储。 [c] 规范化查询 Q 和键 K are 重新计算。 [d] 注意力得分（ $A = Q@K^T$ ）被重新计算。 [e] 价值 V 被存储。规范化的注意权重（ A' = Softmax (A)）是重新计算的。 [f] 在交叉关注中，仅查询 Q 已计数; 键 K has 短得多的序列长度，因此可以忽略不计。 [g] 在交叉关注中，价值 V 序列短得多长度，因此可以忽略不计。 [h] 从 GELU 重新计算输入。 [i] 从 LayerNorm 重新计算输入。

上下文并行 (CP)。 scaling变压器在长上下文设置中引入了计算量（FLOPs）和激活记忆增加的挑战。CP通过在多个GPU上分布计算和激活解决了这些挑战。具体来说，它通过将查询、键和值分成多个部分并在多个GPU上并行处理来实现这一目标。 *芦* 和键值（ *塑, 斟* ）按照其序列维度分成 CP_SIZE 个块，其中 CP_SIZE 是 CP 组内的 GPU 数量。每个 GPU 处理一个块。 *芦* 并使用 (的块迭代地累积部分注意力输出 *塑, 斟* ）存在于相同的CP组中。不同的CP实现利用不同的通信原语，包括all-gather（ ）。 Dubey 等人。 ，2024), P2P (Liu et al. ， 2023 ） ， 以及全面（ 雅各布斯等人。 ， 2023 ）。我们采用 TransformerEngine 中的 P2P 变体（ NVIDIA ， 2024 ） ， 它通过转移来重叠计算和通信（ *塑, 斟* 在GPU之间进行数据块重叠的同时处理注意力，同时在不确定的环境因素下确保稳定的数据传输。通过精心选择数据块大小，这种重叠有效地隐藏了数据传输延迟。我们通过NVLink连接的GPU内部组织CP组，并将CP秩与FSDP秩重叠以实现最佳利用。对于具有较短上下文的图像迭代，禁用CP以提高吞吐量。交叉注意力层不使用CP，因为其序列长度较短。 *塑, 斟* ） ， 这导致计算不足以掩盖通信延迟。

使用Cosmos-1.0-Diffusion-14B作为实例，采用FSDP并设置分片因子为64，可以降低参数、梯度和优化器状态所需的内存要求，从约280 GB降至大约280。 / 64 $\approx$ 每个 GPU 4 GB 。同样 ， 采用 CP _ SIZE = 8 的 CP 将激活内存从 310 GB 减少到大约 310 GB / 8 $\approx$ 40 GB 每块GPU。需要注意的是，这些计算是低估值；实际上，分词器和未分割参数会消耗额外的内存。在CP中重叠通信和计算也需要每块GPU保留多个片段（ of ）。 *塑, 斟* ）.

与其他视频生成模型的比较。 与 Hunyuan Video 中概述的方法相比，我们的并行性策略是有意简化的（Kong 等人。，2024）和 MovieGen（Polyak 等人。，2024），其中整合了张量并行（Tensor Parallelism, TP）及其扩展序列并行（Sequence Parallelism, SP）。尽管我们的设置排除了TP/SP，但它仍能达到与模型浮点操作数利用率（Model FLOPs Utilization, MFU）相当的水平。虽然TP/SP在特定情况下，如更大规模的模型或不同的网络拓扑结构，仍然具有价值，但对于权衡分析，我们将其留作未来研究的工作内容。

提示 紧握蒸汽熨斗的把手，熟练地在皱巴巴的衬衫上滑动。每一次经过，熨斗都会释放出轻柔的蒸汽云，轻松地抚平布料，消除皱纹，展现出整洁利落的最终效果。熨斗以精准和细心的动作，随着每一次划过，都对衬衫进行着转变。淡淡的亚麻香气弥漫在空气中，增添了一份宁静的氛围。一缕柔和的光线从附近的窗户中透入，强调了布料新近光滑的质地，并在这个精细任务展开时营造出一片宁静的环境。

图 12: **从 Cosmos - 1.0 - Diffusion - 7B - Text2world 和 Cosmos - 1.0 - Diffusion - 14B - Text2world 生成的视频。** 两者生成的视频在视觉质量、运动动态和文字对齐方面都达到了高水平。值得注意的是，与7B模型相比，14B模型在捕捉更精细的视觉细节和更复杂的运动模式方面展现了增强的能力。如需查看完整视频和其他视频示例，请访问我们的网站。 网站 .

与大型语言模型的比较。 不同于预训练语言模型（LLMs）通常使用较短上下文长度进行预训练，长上下文设置会显著增加FLOPs（浮点运算次数），因为自注意力机制的成本是二次的。对于LLMs，FLOPs通常计算为$6 \times \text{seq_len} \times$ 参数，where 参数 是参数的数量（卡普兰等人。，2020），我们注意到此公式不适用于我们的扩散WFMs。我们提供了每个关键操作的前向传递FLOPs。 Tab. 13 .

5.1. 5. Prompt upsampler

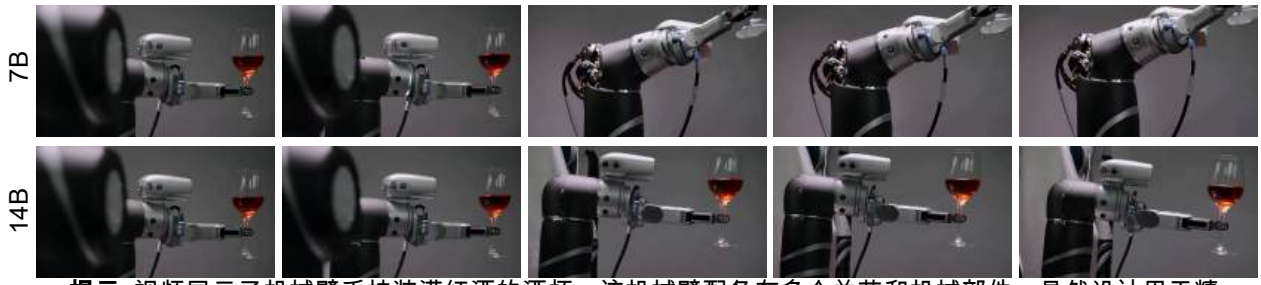
在训练过程中，我们的WFMs使用详细的视频描述作为输入文本提示以生成高质量的视频。然而，在推理过程中，用户提示可能会在长度、结构和风格上有所不同，通常会更短。为了弥合训练和推理文本提示之间的差距，我们开发了一种提示放大器，将其原始输入提示转换为更为详细和丰富的版本。它可以通过增加更多细节并保持一致的描述结构来改进提示，从而提高输出质量。

提示上采样器的主要要求是:

- **对输入提示的保真度** 上采样的提示必须忠实保留原始用户输入的关键元素，包括主要人物、动作或运动、关键属性以及整体意图。 **与培训分布保持一致** 上采样的提示应与WFMs训练提示在长度、语言结构和风格方面保持相似性。 **增强的视觉细节** ：上采样提示应设计为提示 WFM 生成更准确的图像。

提示 Text2World 模型的上采样器。 我们微调 Mistral - NeMo - 12B - Instruct（Mistral 和 NVIDIA，2024 为了构建我们的提示上采样器。为了获得配对数据，即模拟用户输入的短提示和反映训练提示分布的相应长提示，我们使用了一个视觉语言模型（VLM）来生成短提示。

条件框 0 框 29 框 59 框 89 框 120



提示 视频展示了机械臂手持装满红酒的酒杯。该机械臂配备有多个关节和机械部件，显然设计用于精确任务。它轻柔地握住酒杯，展示了其处理易碎物品的能力。背景简约，突出了机器人与酒杯之间的互动。

条件帧 0 帧 150 帧 300 帧 450 帧 680

提示 视频展示了大型工业设施内部场景，可能是工厂或仓库。空间宽敞，天花板高，金属结构明显。可以看到悬挂的起重机和各种机械设备，表明这是一个重工业制造或装配场所。地面大部分空旷，有一些散落的废弃物和标示线。安全标志和隔离带的存在强调了工业环境。照明为自然光，透过高窗洒入，照亮了工作区域。

图 13: **从 Cosmos - 1.0 - Diffusion - 7B - Video2world 和 Cosmos - 1.0 - Diffusion - 14B - Video2world 生成的视频。** 前两行展示了由模型生成的5秒视频，条件基于前9帧。后两行展示了长视频生成的结果。我们以自回归的方式生成长视频，其中第一个生成的视频基于单张输入图像，随后的五个视频则基于它们各自的前九帧。7B和14B模型均生成了具有高度视觉保真的照片级真实感视频。14B模型展示了生成更复杂场景的能力，并表现出更优的运动稳定性。如需查看完整视频和其他更多视频示例，请访问我们的网站。 网站 .

基于我们训练的长提示和相应的视频生成字幕。这种长到短的数据创建策略在以下两个方面（1）保留了由WFMs详细训练提示生成的真实视频内容和分发；以及（2）确保短提示与长提示之间的忠实性。由此产生的提示上采样器称为Cosmos-1.0-PromptUpsampler-12B-Text2World。

提示 Video2World 模型的上采样器。 对于Video2World模型，输入包括视频条件和用户文本提示。为了增强用户提示，我们利用了一个开源的视觉语言模型Pixtral-12B。 Agrawal 等人。 ， 2024 ），结合零样本提示工程，将提示放大为同时考虑视频条件和用户提示的详细描述。我们发现原生的Pixtral-12B模型表现良好，无需进行上述类似的微调。

5.1. 6. Results

In Fig. 12 我们呈现了由我们的Cosmos-1.0-Diffusion-7B-Text2World和Cosmos-1.0-Diffusion-14B-Text2World模型生成的定性结果。这两个模型都能产出高质量的视觉效果、动态运动以及文字对齐的视频。相较于7B模型，14B模型能够生成捕捉更多细节的视频。

复杂的视觉细节和复杂的动作。

我们显示从 Video2World 7B 和 14B 模型生成的视频 Fig. 13 . The Video2World模型支持图像和视频条件输入，并能够以自回归的方式生成扩展的视频。如所示： Fig. 13 我们的Video2World模型生成具有良好的运动动态和视觉保真的 PHOTOREALISTIC 视频。14B模型在场景丰富性和运动稳定性方面再次生成了更好的视频。

5.2. 基于自回归的世界基础模型

在自回归WFMs中，我们将世界模拟生成 formulize 为类似于语言建模的下一个令牌预测任务。我们首先将视频转换为离散的视频令牌序列。 $x = \{x_1, x_2, \ldots, x_N\}$
使用中引入的 Cosmos 离散令牌器 Sec. 4 。然后我们训练一个 Transformer 解码器 (Vaswani 等人。 ， 2017) 使用过去的视频令牌作为上下文来预测下一个视频令牌 ， 类似于大型语言模型 (LLM) (布朗等人 。 ， 2020 ； Dubey 等人。 ， 2024 ； Jiang 等人。 ， 2023) 。具体来说 ， 训练目标是最小化以下负对数似然(NLL) 损失 ：

$$\mathcal{L} = -\log p_\theta(x_i | x_1, x_2, \ldots, x_{i-1}; \Theta), \tag{9}$$

其中条件概率 p_θ 预测的下一个视频令牌 x_i 在 x_{i-1} 由具有参数 Θ 的变换器解码器建模。

5.2. 1. Architecture

我们基于自回归的 WFM 架构在 Fig. 14 为了适应我们的视频生成任务，我们对标准的转换器模型架构进行了多项修改，包括添加了1）三维感知的位置嵌入，2）交叉注意力以增强文本输入的控制能力，以及3）QK-归一化。 Wortsman 等人。 ， 2023).

3D 位置嵌入。 类似于我们基于扩散的 WFM (Sec. 5.1.2), 我们整合了两种互补的位置嵌入机制：用于相对位置的3D因子旋转位置嵌入（RoPE）和用于绝对坐标的位置嵌入（APE）。这些机制共同作用，为整个网络提供全面的空间和时间信息。

- **3D 旋转位置嵌入(RoPE) 。** 我们应用3D RoPE技术来编码时间、高度和宽度维度上的相对位置信息。在训练过程中，我们采用多阶段训练策略，随着训练的进行，视频序列长度逐渐增加。为了使3D RoPE适应不断变化的时间长度，我们使用YaRN（你未完整提供YaRN的具体信息，假设YaRN是一种适应性方法）。 Peng et al. ， 2023)，一种旨在扩展RoPE上下文窗口的计算效率高技术。我们仅沿时间轴应用YaRN扩展，因为随着视频序列长度仅沿时间维度增加。通过利用YaRN，我们的模型可以外推到比训练初期遇到的更长的上下文长度。• 除了 3D RoPE ， 我们还将 3D APE 纳入

3D 绝对位置嵌入 (APE) 。
每个变压器块以补充相对位置编码的方式进行。这种绝对位置编码（APE）使用周期性嵌入来编码位置信息，并且这些嵌入在时间、高度和宽度维度上被分解，确保模型了解绝对位置。嵌入直接添加到输入张量的每个阶段，从而丰富了变压器的位置上下文。我们发现结合绝对和相对位置编码能够提升模型性能、降低训练损失，并减少生成视频中的形态变形伪影。值得注意的是，在基于扩散的WFM（未完整信息，假设为某种模型或方法）中，这种结合方法尤其有效。 Sec. 5.1.2 ）采用可学习嵌入 ， 我们在基于自回归的 WFM 中对 APE 采用基于正弦的嵌入。

词汇。 标记化是大型语言模型（LLMs）中将输入文本转换为离散标记序列的关键步骤。在大型语言模型中，可能标记的词汇表由模型自身的分词器确定。

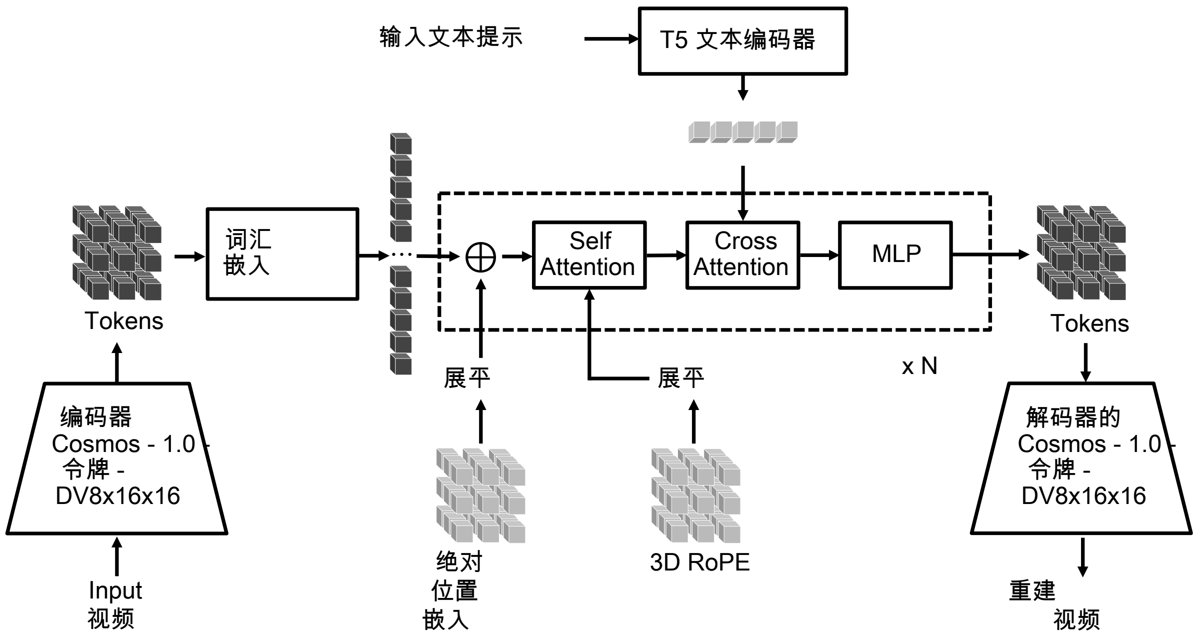


图 14 ： **Cosmos - 1.0 - 自回归 - Video2World 模型的体系结构。** 管道流程始于通过Cosmos-1.0-Tokenizer-DV8x16x16的编码器对输入视频进行编码，生成离散令牌，这些令牌随后被转换为学习嵌入。这些嵌入通过重复的变压器块处理，每个块包含绝对位置嵌入和3D RoPE组件，在进入自我注意模块之前会被展平。每个块还包含一个跨注意力模块，该模块整合了通过T5文本编码器处理的编码文本提示，随后是两层MLP。最后，Cosmos-1.0-Tokenizer-DV8x16x16的解码器从输出令牌中重构视频。

(*e.g* 。 ，tiktoken 由介绍 OpenAI (2022)) 在大型文本语料库上训练 ， 使用字节对编码 (BPE) 等算法 (Gage ， 1994).

对于我们的自回归模型 ， 我们使用 Cosmos - 1.0 - Tokenizer - DV8x16x16 作为令牌器。如在 Sec. 4 ， 我们利用有限标量量化 (FSQ) (Mentzer 等人。 ， 2023) 将 6 维潜在空间量化为 (8 ， 8 ， 8 ， 5 5) 级别。这种量化导致词汇量为 8 × 8 × 8 × 5 × 5 × 5 = 64 ， 000.

文本条件的交叉注意。 除了transformer架构中存在的自注意力块外，我们还添加了跨注意力层，以使模型能够条件于输入文本。类似于基于扩散的WFM (Sec. 5.1.2)，跨注意力机制应用于transformer模型的特征与由预训练文本编码器 (T5-XXL) 获得的文本嵌入之间。在我们的实验中，我们在每个自我注意层之后添加跨注意力块。

查询键规范化。 为了增强训练稳定性 ， 我们将查询密钥标准化 (QKNorm) (Wortsman 等人。 ， 2023). QKNorm 通过规范化查询来解决注意力机制中的不稳定性 (q) 和密钥 (k) 在计算点积之前先对向量进行归一化，从而防止softmax函数饱和，并确保更有效的学习。归一化之后，点积会被一个可学习的参数进行缩放，而不是固定值。 $1/\sqrt{d_k}$ 此可学习的缩放因子允许模型适当地控制注意力分数的大小，增强其灵活性和表达能力。

Z 损失。 为了进一步提高训练稳定性 ， 我们引入了一个被称为 z 损失的稳定项 (de Br é bisson and Vincent ， 2016) 作为我们的训练目标。z-loss 对 logits 与零的偏差进行惩罚，有效地防止模型生成过大的 logits 值，从而避免数值不稳定或梯度爆炸的问题。z-loss 定义为 logits 的平方和。
$\mathcal{L}_{z\text{-损失}} = \lambda \cdot \log^2 Z$. 我们发现
z 损失对于将梯度规范保持在健康范围内至关重要 ， 尤其是在将训练扩展到 λ

大量的 GPU 节点。根据经验 ， 我们发现 z 损失系数 $\beta = 3 \times 10^{-4}$ 取得最佳平衡 ， 有效稳定训练 ， 而不会对模型性能产生不利影响。

5.2. 2. 放大

本节描述了使我们的自回归WFMs高效扩展的技术。我们简要分析了模型的内存消耗，讨论了并行策略，并将我们的训练设置与其他自回归模型进行了比较。

内存要求。 在训练过程中 ， GPU 内存主要消耗 ：

- **模型参数** ： 每个参数 6 字节。我们将模型参数存储在 BF16 和 FP32 中。 • **梯度** ： 每个参数 2 个字节。我们将梯度存储在 BF16 中。 • **优化器状态** ： 每个参数 8 个字节。我们存储 AdamW 的第一和第二矩 (Loshchilov 和 Hutter ， 2019) 均在 FP32 中。 • **激活** ： 大约(2 × number _ of _ layers × 17 × seq _ len × batch _ size × d _ model) 字节。我们参考读者 Korthikanti 等人。 (2023) 详细分析了最新的自回归模型的激活记忆。

例如，我们的12B模型（Cosmos-1.0-Autoregressive-12B）需要大约192GB的内存来存储其参数、梯度和优化器状态。由于这超出了单个NVIDIA H100 GPU的80GB HBM3容量，我们利用张量并行性（TP）来实现这一需求。 Shoeybi 等人。 ， 2019 ）及其扩展、序列并行性 (SP) (Korthikanti 等人。 ， 2023 ） ， 以跨多个 GPU 分配内存需求和计算。

张量并行性 (TP) 。 张量平行度 (TP) (Shoeybi 等人。 ， 2019 采用TP（Tensor Parallelism）策略，权重沿输入或输出特征维度进行划分，决策目标是减小GPU间的通信量。例如，在一个两层的前向传播网络中，第一层的权重沿输出特征维度分割，而第二层的权重沿输入特征维度分割。这种布局允许中间激活值在本地处理，无需GPU间通信。最终输出通过全集通信（all-reduce communication）进行整合。通过使用TP，每个GPU仅存储其部分权重，具体而言，仅为总权重的1/N（N为GPU数量）。 ／TP_SIZE，用于线性层权重的张量尺寸。然而，TP的默认实现仍然会在像LayerNorm这样的操作中沿序列维度复制激活，导致冗余。

序列并行性 (SP) 。 SP (Korthikanti 等人。 ， 2023) 通过进一步沿序列维度分割上下文来扩展张量并行性。这种方法适用于自我注意力层中的运算符，如LayerNorm和Dropout，其中序列中的每个元素可以独立处理。启用SP后，每块GPU仅存储特定比例的数据，具体为1/ ／激活的 TP _ SIZE 。

与其他自回归模型的比较。 相比于流行的大型语言模型（LLMs），我们的模型并未采用诸如MQA或GQA等内存节省型注意力机制。否则，我们自回归模型的设计目的是尽可能接近LLMs的架构结构。 阿德勒等人。 ， 2024 ； 布朗等人。 ， 2020 ； Dubey 等人。 ， 2024 ； Jiang 等人。 ， 2023 ； Team ， 2024 ； Yang et al. ， 2024), 由于这种对齐提供了灵活性和可扩展性。未来的工作可以利用更多的并行性（如上下文并行性和管道并行性）来进一步扩大模型规模和上下文长度。

5.2. 3. 培训策略

我们分多个阶段对我们的自回归 WFM 进行预训练。

- **阶段 1** 在第一阶段，模型使用视频预测目标进行训练。给定第一个帧作为输入条件，模型被训练以预测未来的视频帧。该任务使用17帧作为上下文长度。 *i.e* 。 ， 该模型以第一帧作为输入来预测 16 个未来帧。

• **阶段 1.1** 这一阶段执行视频预测操作，但上下文长度增加至34帧。我们在时间维度上使用YaRN扩展来增加RoPE的上下文长度。 ： 在我们训练的第二阶段 ， 我们将文本条件引入到我们的模型中。文本嵌入是
阶段 2
Incorporated using newly initialized cross-attention layers。模型使用34帧上下文进行训练。为了提高文本到视频生成能力，模型使用联合图像和视频数据进行训练，如文中所述。 Sec. 5.1.3 当使用图像批次时，我们使用较大的批次大小作为上下文长度，因为图像的上下文长度远小于视频。

我们所有的模型都是用 640 × 1024 的固定空间分辨率训练的。

冷却下来。 预培训后 ， 我们进行 “冷却 ” 阶段 ， 使用高质量的数据 ， 类似于 LLM 培训实践 （ Dubey 等人。 ， 2024). 在此阶段，我们在线性衰减学习率至0的同时，在高质量的图像-视频对上进行训练。冷却期持续30个周期。 ，000 次迭代。

表 14 ： Cosmos - 1.0 - 自回归模型的配置细节。

配置	4B	5B - Video2World	12B	13B - Video2World
层数	16	16	40	40
模型维度	4,096	4,096	5,120	5,120
交叉注意层	✗	✓	✗	✓
基本学习率	1×10^{-3}	3×10^{-4}	1×10^{-3}	5×10^{-4}
重量衰减	0.01			
学习率热身赛	线性调度器与 5		,000 次迭代	
激活功能	SwiGLU			
FFN 隐藏尺寸	14,336			
注意头数	32			
键 / 值头的数量	8			
令牌数量	12,800			
词汇大小	64,000			
位置嵌入	3D RoPE ($\theta = 500{,}000$) + 3D APE			

我们训练了两组基于自回归的WFMs (Wave Function Machines) 。首先，我们构建了两个基础模型：一个容量为4B，另一个为12B。这些模型仅预测下一个视频令牌，不接受文本提示作为输入。接着，我们从每个基础模型中衍生出Video2World版本，在这些模型中添加交叉注意力层以利用文本提示来预测下一个视频令牌。

• **Cosmos - 1.0 - 自回归 - 4B** 一个用于下一视频令牌预测的4B变压器模型。该模型使用多阶段训练目标的stage 1和stage 1.1进行训练。 **Cosmos - 1.0 - 自回归 - 5B - Video2World** 一个5B参数量的变压器模型，基于我们的Cosmos-1.0-Autoregressive-4B，并额外使用了多阶段训练目标的第二阶段进行训练。 **Cosmos - 1.0 - 自回归 - 12B** 一个12B转换器模型用于下一视频令牌预测。该模型使用多阶段训练目标的第一阶段和第一点一阶段进行训练。 **Cosmos - 1.0 - 自回归 - 13B - Video2World** 一个源自Cosmos-1.0-Autoregressive-12B且经过额外的第二阶段多阶段训练目标训练得到的13B变压器模型。

5.2. 4. 对实时生成的推理优化

我们的Cosmos自回归WFMs在架构上与LLMs具有相似性，这使我们能够利用已建立的LLM推断优化技术来解决顺序解码瓶颈。我们实现了一个
[3]
结合关键值缓存、张量并行性和torch.compile，遵循PyTorch中gpt-fast的实现方式。 Paszke et al. ， 2019).

投机性解码。 为了进一步加速我们的自回归 WFM ， 我们应用了美杜莎推测解码框架 (蔡等人。 ， 2024) 。与常见的推测解码方法不同 ， 这些方法需要单独的

[3] https: / / github. com / pytorch - labs / gpt - fast

草稿模型（ 利维坦等人。 ， 2023 ）或加速有限的免培训方法（ Teng 等人. ， 2024), 美杜莎通过增加额外的解码头来并行预测多个后续标记，从而扩展了变压器骨干网络。然后，它使用拒绝采样验证这些推测出的标记。这样可以缓解逐标记处理的瓶颈，从而加速推理过程。我们展示了美杜莎技术在视觉自回归加速方面的潜力，同时不牺牲生成输出的质量。

在我们的实现中，我们通过将Medusa头部引入架构来微调预训练的自回归WFMs。这些头部在最后一个变压器隐藏状态之后战略性地插入，所有骨干参数和最终的未嵌入层在不同的头部之间共享。每个Medusa头部是一个带有SiLU激活和残差连接的单层前馈网络（FFN）。我们进一步将多个Medusa头部的权重矩阵合并为一个统一的FFN，以最大化在标记预测过程中的并行性。请注意，我们没有使用基于树的注意力机制 。 蔡等人。 （ 2024).

为了探讨我们自回归WFMs的最佳Medusa配置，我们从两个方面进行了深入研究：（1）哪些转换器层需要微调，以及（2）应添加多少Medusa头。对于第一个问题，我们比较了全量微调和选择性层冻结之间的差异。我们发现仅微调Medusa头会导致多令牌预测效果不佳，而全量微调则会引发质量下降。通过实证分析，我们确定了在保持主干冻结的同时解冻最后两个转换器层和最终去嵌入层，可以获得最佳性能。这一策略确保我们的Medusa训练在避免灾难性遗忘的情况下，实现了良好的推测性解码精度。

表 15: 美杜莎头数对平均令牌吞吐量和正向传递次数的影响。
实验是在 8 × H100 GPU 和 50 个未见过的 640 测试视频 × 1024 分辨率。

模型	美杜莎头号	0	3	6	9	12
4B	令牌吞吐量 (令牌 / 秒)	444.95	663.51	829.59	**894.67**	890.64
	# 远期通票	7680	2860	2073	1812	**1682**
5B	令牌吞吐量 (令牌 / 秒)	303.61	659.94	758.58	**982.77**	978.80
	# 远期通票	10240	2857	2382	1799	**1673**

为了探索最优的Medusa头的数量，我们计算了不同数量的Medusa头时的模型代币吞吐量和前向传递次数。消融研究在8个不同的设置下进行。 × H100 GPU ， 并在 640 的 50 个看不见的测试视频上进行了评估 × 1024 分辨率。结果 Tab. 15 建议我们的美杜莎框架可以有效地加速推理 ， 最多 2 . 0 × 令牌吞吐量和 4 . 8 B 模型的正向传球较少 ， 最多 3 . 2 × 令牌吞吐量和 6 . 1 × 5B模型的前向传递次数较少。我们表明，尽管更多的Medusa头可以减少生成所需的前提传递次数，但它可能会减慢整体令牌吞吐量。我们发现，9个Medusa头在计算效率和模型性能之间提供了最佳权衡。

In Tab. 16 我们展示了结合Medusa后的自回归WFMs的性能分析。该分析是在H100 GPU上进行的，并在640帧的测试视频上进行了评估。 × 1024分辨率下的BF16精度。结果显示，Medusa 实现始终能分别在不同的GPU配置下加速4B和5B模型的推理过程。

实时推理的低分辨率自适应。 我们通过将模型调整到 320 的较低空间分辨率来追求实时推理 × 512，这导致每个视频的token数量减少。具体而言，我们首先对离散视频分词器（Cosmos-1.0-Tokenizer-DV8）进行微调。 × 16 × 16 in Sec. 4 对于320p低分辨率视频，我们使用目标物理人工智能领域的视频进行训练。随后，我们对预先在640分辨率下训练的自回归WFM进行微调。 × 1024 分辨率 (Cosmos - 1.0 - 自回归 - 4B in Sec. 5.2.3) 在 320 的视频上使用这个低分辨率标记器 × 目标物理AI领域中的512分辨率。最后，我们添加了Medusa头到细调后的低分辨率自回归WFM中。

表 16: Cosmos 自回归模型对 640 × 1024 分辨率的测试视频的性能分析。

模型	GPU	无 DD (s)	无 DD + 美杜莎 (s)	带 DD (s)	带 DD + 美杜莎 (s)	VRAM (GB)
4B	1	31.04	23.52	61.49	53.08	29
	4	18.20	13.87	29.60	25.63	31
	8	17.62	9.91	30.30	22.83	34
5B	1	39.68	24.97	70.39	54.72	59
	4	25.59	20.96	37.29	33.35	51
	8	25.70	11.67	38.41	24.35	49
12B	1	84.78	—	116.66	—	45
	4	47.49	—	60.27	—	36
	8	45.69	—	58.81	—	37
13B	1	109.18	—	140.24	—	77
	4	67.80	—	80.76	—	55
	8	67.22	—	80.93	—	55

表格报告了在不同设置下各种Cosmos自回归WFMs的平均推理时间（单位：秒）和VRAM利用率。推理时间是在将单个条件帧作为输入生成32帧时报告的。 **No DD**：无扩散解码器的时间。 **无 DD + 美杜莎**：没有扩散解码器但有美杜莎头的时间。 **带 DD**：使用扩散解码器的时间。 **DD + 美杜莎**：带有扩散解码器和美杜莎头的时间。
VRAM：以 GB 为单位的视频 RAM 使用情况。

表 17：具有低分辨率自适应的 Cosmos - 1.0 - 自回归 - 4B 的解码吞吐量，基准 on 8 × H100 80GB GPU 使用 10 - FPS 视频 320 × 物理 AI 域的 512 分辨率。

型号 (320×512)	令牌吞吐量 (令牌 / 秒)	视频吞吐量 (帧 / 秒)
Cosmos - 1.0 - 自回归 - 4B(与美杜莎)	806.61	10.08

我们在 8 × H100 GPU 使用 `torch.compile` 的 “max-autotune” 模式并在 BF16 精度下运行，并使用目标 Physical AI 领域的 10-FPS 输入视频进行评估。 Tab. 17 我们在这种设置下报告了平均代币吞吐量和帧生成速度。我们观察到我们的模型可以在不到1秒的时间内生成10个视频帧，这表明我们可以实现每秒10帧的实时视频生成。

5.2. 5. 扩散解码器

我们的Cosmos分词器采用轻量级的编码器-解码器架构进行激进压缩，从而减少WFM训练中的token数量。由于激进压缩，这有时会导致视频生成中出现模糊和可见的伪像，尤其是在自回归WFM设置中，仅通过离散分词使用少数几个整数来表示丰富的视频。为此，我们采用了扩散解码器设计（diffusion decoder design）。 OpenAI，2024；Ramesh 等人。，2022）以解决这一限制。具体而言，我们通过微调Cosmos-1.0-Diffusion-7B-Text2Video构建了一个更为强大的Tokenizer Decoder。 Sec. 5.1 .

Fig. 15 展示了我们如何训练自回归WFMs中的扩散解码器。对于每个训练视频，我们使用Cosmos-1.0-Tokenizer-CV8x8x8和Cosmos-1.0-Tokenizer-DV8x16x16分别计算一个连续标记视频和相应的离散标记视频。我们注意到，由于更加温和的连续标记化过程和较少激进的压缩方案（8），Cosmos-1.0-Tokenizer-CV8x8x8能够生成更高质量的视频输出。 × 8 × 8 而不是 8 × 16 × 16).

离散令牌视频被视为Cosmos-1.0-Diffusion-7B模型中去噪器的条件输入。为了计算该条件输入，我们首先通过可学习词汇嵌入层将每个离散令牌的离散令牌视频嵌入到一个16维向量中。然后，我们将嵌入进行上采样2次。 ×
沿着 and 方向以确保条件输入与去噪器从连续标记视频获取的噪声输入具有相同的大小。我们将噪声连续输入与条件输入进行连接。

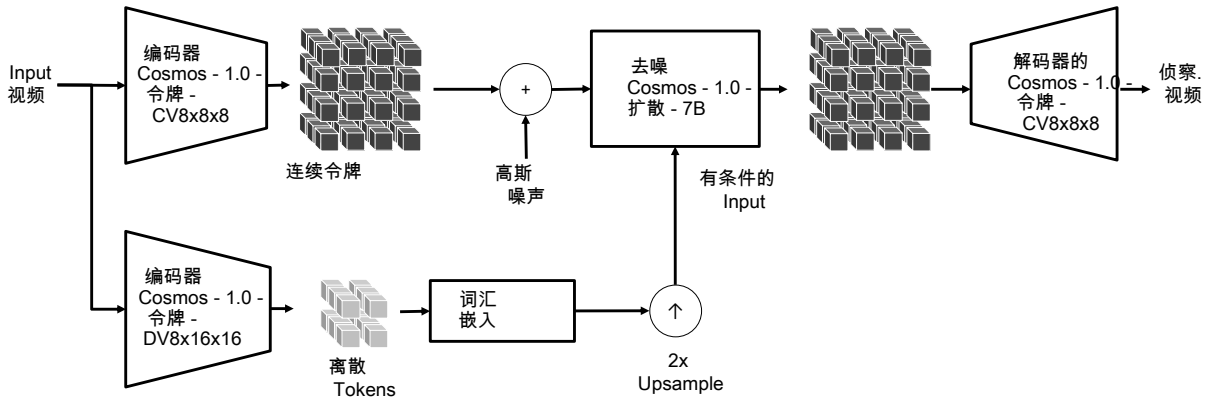


图 15: **宇宙扩散解码器训练** 在训练过程中，每个输入视频会被 tokenize 两次：一次通过目标离散 tokenizer（DV8x16x16），另一次通过一个约束较少的连续 tokenizer（CV8x8x8）。离散 token 视频被用作扩散去噪器的条件输入。

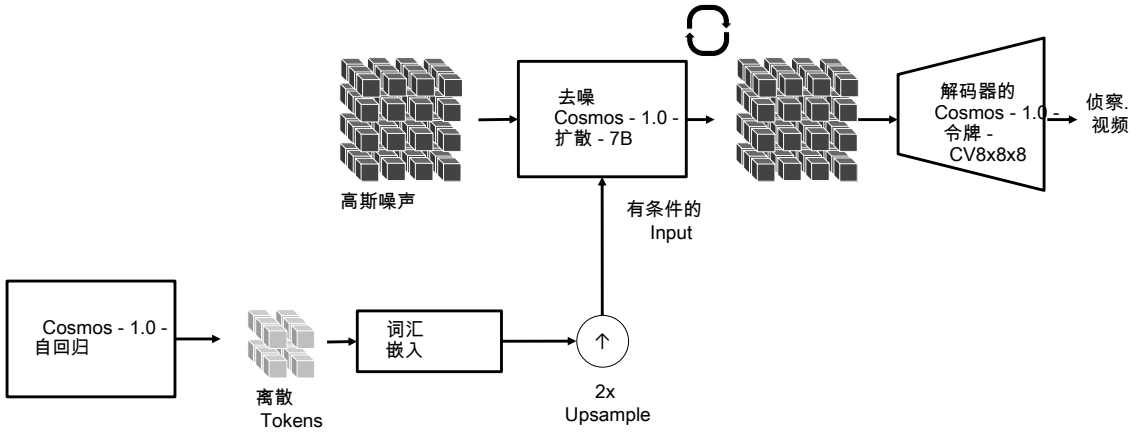


图 16: **宇宙扩散解码器推理** 在推断过程中，来自Cosmos-1.0-Autoregressive模型的输出视频token作为条件输入传递给去噪器。

沿通道维度的输入，这些成为扩散去噪器的输入。去噪器的第一层扩展了通道维度，以适应新的输入形状。我们通过移除添加的噪声来微调更新后的Cosmos-1.0-Diffusion-7B。由于离散令牌视频未受到噪声干扰，去噪器学习利用条件输入中驻留的信息进行去噪。结果是一个更高质量的解码器，该解码器通过解决储备扩散问题来解码离散令牌，从而为分词器提供更高质量的输出。

Fig. 16 说明了推理。输出离散令牌视频 (8 下 × 16 × 16个离散压缩）由我们自回归WFM解码为视频，通过两个步骤进行。首先，我们展开条件去噪器以生成一个连续标记视频（小于8） × 8 × 基于自回归WFM输出的连续压缩（8连续压缩）。接下来，连续标记视频由Cosmos-1.0-Tokenizer-CV8x8x8解码以生成最终的RGB视频。

5.2. 6. Results

In Fig. 17 我们展示了使用不同模型规模的自回归WFMs的定性结果。在未提示设置中，我们将Cosmos-1.0-Autoregressive-4B模型与Cosmos-1.0-Autoregressive-12B模型进行比较，

提示 : None.

提示 一辆汽车向前行驶，穿过一座大型高架桥。道路宽敞，远处可以看到几辆其他车辆。天气看起来是晴朗的，时间是白天。场景设定在一个繁忙的高速公路上，两侧有混凝土结构和绿化。

图 17: **从宇宙自回归世界基金会模型生成的视频。** 前两行是 4B 和 12B 型号的视频生成结果 ， 而底部两行是 Video2World 结果 带有文本提示。我们观察到 12B 和 13B 模型比 4B 显示更清晰的视频和更好的运动 和 5B 模型在提示和未提示的设置中。要检查完整的视频和更多的视频示例 ， 请访问我们的 网站 .

我们观察到，12B模型生成的视频在运动效果和细节清晰度方面更好。类似地，在受提示设置下，比较Cosmos-1.0-Autoregressive-5B-Video2World和Cosmos-1.0-Autoregressive-13B-Video2World时发现，13B模型的运动效果优于5B模型。

In Fig. 18 我们展示了使用扩散解码器时获得的改进。由于我们在离散分词器中采用了有损压缩，因此自回归模型的输出较为模糊。使用扩散解码器可以增强细节同时保留内容。

我们实证发现，基于自回归的Text2World WFMs的输出并不随着来自文中讨论的提示上采样器生成的上采样提示而改善。 Sec. 5.1.5 我们推测这可能是因为这些WFMs在大部分训练过程中都是通过纯视频生成任务预训练的。它们没有被充分强制利用文本输入。

5.2. 7. 局限性

在自回归WFMs生成的视频中观察到的一个值得注意的失败案例是物体意外地从下方出现。 Fig. 19 这示例说明了这一问题的一个实例。为了理解我们模型的失败率，我们通过创建包含100个物理AI输入到自回归模型的评估集，进行了一项系统研究来深入理解这一问题。

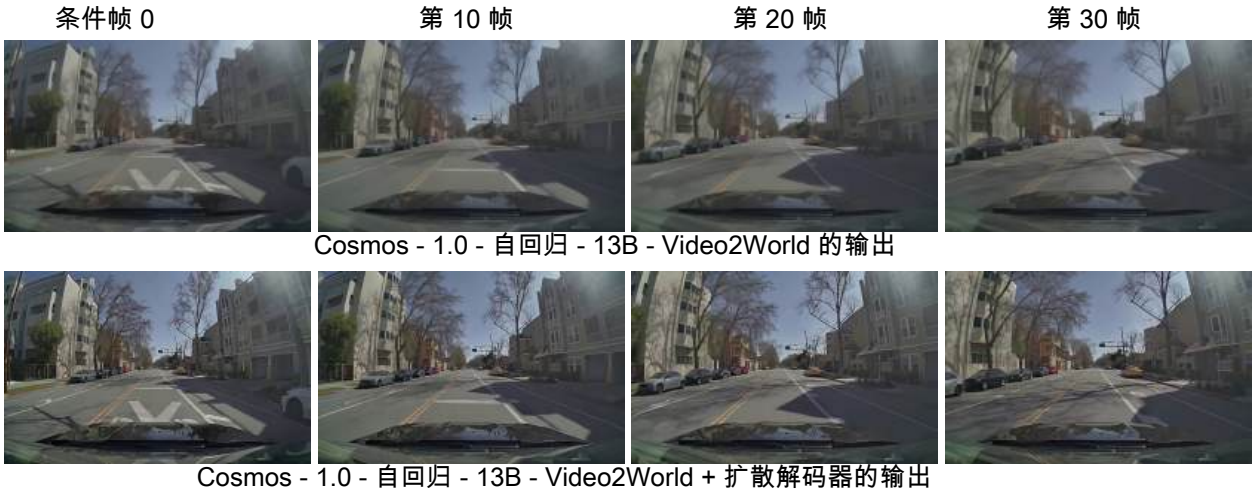


图 18: **扩散解码器比较** 在上图部分，我们展示了使用Cosmos-1.0-Autoregressive-13B-Video2World模型生成的视频结果。在下图部分，我们展示了自回归模型输出后通过扩散解码器处理后的增强视频。我们观察到，仅自回归模型产生的结果模糊不清，而扩散解码器可以在保持内容不变的情况下提升视频的清晰度。

图 19: **Cosmos 自动 WFM 的故障案例** 我们在生成的视频中观察到失败案例 ， 其中一些对象(以红色显示) 从下面意外出现。

WFMs。我们使用两种输入模式生成所有模型的视频——图像（单帧）条件和视频（9帧）条件。对于所有生成的视频，我们手动检查失败案例并报告失败率。 Tab. 18 我们观察到，在单帧条件下的较小模型Cosmos-1.0-Autoregressive-4B和Cosmos-1.0-Autoregressive-5B-Video2World表现出更高的篡改率，而较大的模型Cosmos-1.0-Autoregressive-12B和Cosmos-1.0-Autoregressive-13B-Video2World则更为稳健。使用9帧视频条件进行生成时，所有模型均表现稳定，失败率低于2%。

5.3. 评价

预训练的视觉世界模拟模型（WFMs）是通用型模型。其能力应在多个方面进行衡量。在这里，我们从两个方面评估我们的模型。首先，我们评估生成视频的三维一致性。理想的WFMs应能够从几何上合理的三维世界中生成视频模拟。其次，我们评估生成视频的物理一致性。我们计算渲染的动力学与物理定律的一致程度。对WFMs的评估是一项高度非平凡的任务。我们承认还需要考虑其他几个重要的评估方面。我们将更全面的评估留作未来的工作。

5.3. 1. 3D 一致性

WFMs 设计用于通过视频生成模拟三维世界，并且评估生成的视频与视觉世界的三维结构的一致性至关重要。除了外观真实外，

表 18 ： Cosmos 自回归模型的故障率分析。

模型	图像调节	视频调节 (9 框架)
Cosmos - 1.0 - 自回归 - 4B	15%	1%
Cosmos - 1.0 - 自回归 - 5B - Video2World	7%	2%
Cosmos - 1.0 - 自回归 - 12B	2%	1%
Cosmos - 1.0 - 自回归 - 13B - Video2World	3%	0%

生成的视频应保持与场景随时间变化的物理原理的一致性，这是下游物理AI应用的一项关键要求。

测试数据和基准模型。 我们专注于静态场景的场景，以有效地使用基于多视几何的现有工具测量视频的3D一致性。我们精选了来自RealEstate10K数据集测试集的500个随机视频。 周等人。 ， 2018). 我们还使用 proprietary VLM 对视频进行标注，以获得描述视频为静态场景的文字提示，从而无需考虑场景运动对度量计算的影响。我们将结果与 VideoLDM 进行对比 (VideoLDM) 。 Blattmann 等人。 ， 2023) 作为基线方法。

指标。 生成的视频实际上是底层3D视觉世界的有效二维投影。我们设计了以下指标来衡量生成视频的3D一致性。

1. **几何一致性** 我们通过量化共轭几何约束 (包括萨姆森误差 Sampson error) 的满足程度来评估生成世界的三维一致性。 哈特利和齐塞曼 ， 2003 ； 桑普森 ， 1982) 和摄像机姿态估计算法的成功率 (舍恩伯格等人。 ， 2016 ； 舍恩伯格和弗拉姆 ， 2016) 在生成的视频上。

2. **查看合成一致性** 我们评估世界基础模型在合成插值的新视角图像时保持与底层3D结构一致性的能力。

萨蒙森误差是一视角点到另一视图中对应极线的距离的一阶近似。给定 N 点对应 (用齐次坐标表示) $\{(\bar{x}_i, \bar{y}_i)\}_{i=1}^{N}$ 在给定的帧对中 ， 我们将 Sampson 误差定义为

$$[\ ]\ 1\ 0\ 0 \quad y\ F \quad \text{其中 } S = 0\ 1\ 0 \quad \| SFx \|^2 \quad 2 \tag{10}$$

F是从对应关系中估计的基本矩阵。我们使用平方根版本的误差函数，使像素单位中的度量更具直观性。我们采用SuperPoint的组合方法。 DeTone 等人。 ， 2018) 和 LightGlue (林登伯格等人。 ， 2023) 用于检测和匹配帧对中的关键点对应关系，并使用OpenCV的8点RANSAC算法估计变换矩阵F。我们通过与960像素长度的比例来归一化平均误差，相对于帧的对角线长度。 × 540 画布。

我们还评估生成视频与其自合成新型视角的一致性。遵循新型视角合成文献的常见做法 (Mildenhall 等人。 ， 2020) ， 我们将每 8 帧作为测试帧 ， 并拟合 3D 高斯飞溅模型 (Kerbl 等人。 ， 2023) ， 其余的训练帧使用 Nerfstudio 库的默认设置 (Tancik 等人。 ， 2023) 。我们报告了峰值信噪比 (PSNR) 、结构相似性 (SSIM) 和 LPIPS (张等人。 ， 2018) 作为量化合成测试视图质量的指标。

结果。 我们将定量评估结果呈现在 Tab. 19 宇宙WFMs在几何一致性和视图合成一致性方面相对于我们的基线模型实现了显着更好的三维一致性。不仅来自宇宙WFMs的兴趣点在三维上更加一致，而且相机姿态估计也表现出更高的准确性。

表 19: 基于 Cosmos 模型的 3D 一致性评价。

方法	成功率 (%) ↑	几何一致性 桑普森错误 ↓	视图综合一致性 姿势估计 ↓	PSNR ↑	SSIM ↑	LPIPS ↓
VideoLDM (Blattmann 等人。, 2023)	0.841	4.4%	26.23	0.783	0.135	
宇宙 - 1.0 - 扩散 - 7B - Text2World	**0.355**	62.6%	**33.02**	**0.939**	**0.070**	
宇宙 - 1.0 - 扩散 - 7B - Video2World 0.473	**68.4%**	30.66	0.929	0.085		
Cosmos - 1.0 - 自回归 - 4B 0.433 35.6% 32.56 0.933 0.090						
宇宙 - 1.0 - 自回归 - 5B - Video2World 0.392 27.0% 32.18 0.931 0.090						
真实视频(参考) 0.431 56.4% 35.38 0.962 0.054						

成功率也显著提高，这既反映了整体质量的提升，也体现了增强的3D一致性，甚至达到了真实世界视频的水平。在成功估计相机姿态的情况下，合成的留出视图在所有图像合成指标上都显示出更高的质量。这些结果突显了我们Cosmos WFMs生成3D一致视频的能力，确立了它们作为有效的世界模拟器的地位。

5.3. 2. Physics Alignment

理想的工作流管理（WFM）应具备对物理定律的深刻理解，并能够生成符合这些定律的未来观测结果。虽然我们预先训练的WFMs在一定程度上体现了物理理解并推动了现有技术水平，但仍可以很容易地生成违反物理定律的例子。我们认为，在数据整理过程中需要移除物理上不可行的视频，并改进模型设计。尽管我们将构建高度与物理相一致的WFM作为未来的工作，但我们仍然感兴趣于测量大规模数据驱动预训练中自然涌现的直观物理知识的程度。

为了探索这一点，我们使用物理仿真引擎设计了一个受控的基准数据集，灵感来自（Kang 等人。，2024）我们通过基于物理的模拟来测试预训练的WFMs对牛顿物理学和刚体动力学的遵守程度。具体来说，我们使用模拟来生成特定于感兴趣的物理定律的真实且逼真的视频，这些视频描述了测试场景下的物理现象。然后，我们将这些“参考真实”视频与给定共享上下文（过去观测和扰动）的WFMs产生的“预测”视频进行比较。

合成数据生成。 使用 PhysX（NVIDIA，2024）和艾萨克·辛（NVIDIA，2024），我们设计了八个3D 场景，旨在评估不同的物理效果：

1. **自由落体**：物体在平面上掉落(重力，碰撞，*etc*.) 2. **倾斜平面坡度**：物体在斜坡上滚动 (重力，惯性矩，*etc*.) 3. **U 形坡**：物体沿着 U 形斜坡滚动 (势能、动能、*etc*.) 4. **稳定的堆栈**：一堆处于平衡状态的物体 (平衡力) 5. **堆栈不稳定**：一堆不平衡的物体 (重力、碰撞、*etc*.) 6. **多米诺骨牌**：顺序落下的矩形砖序列 (动量传递、碰撞、*etc*.) 7. **跷跷板**：跷跷板两侧的物体 (扭矩、旋转惯性、*etc*.) 8. **陀螺仪**：平坦表面上的旋转顶部 (角动量，进动，*etc*.)

对于每个场景，我们随机化动态物体的数量和类型（包括不同的大小、纹理、形状），从中选择 Omniverse 资产。NVIDIA，2024 在这样的背景下，我们模拟了物体随时间的变化状态，并从四个静态摄像机视角生成输出视频。总共生成了800个长度为100帧的1080p视频。在每个模拟中放置的物体被定位，使得它们从第一帧开始就全部可见，以避免任何存在性混淆。

指标。 我们有兴趣通过比较模拟的地面实况来评估对物理定律的遵守情况

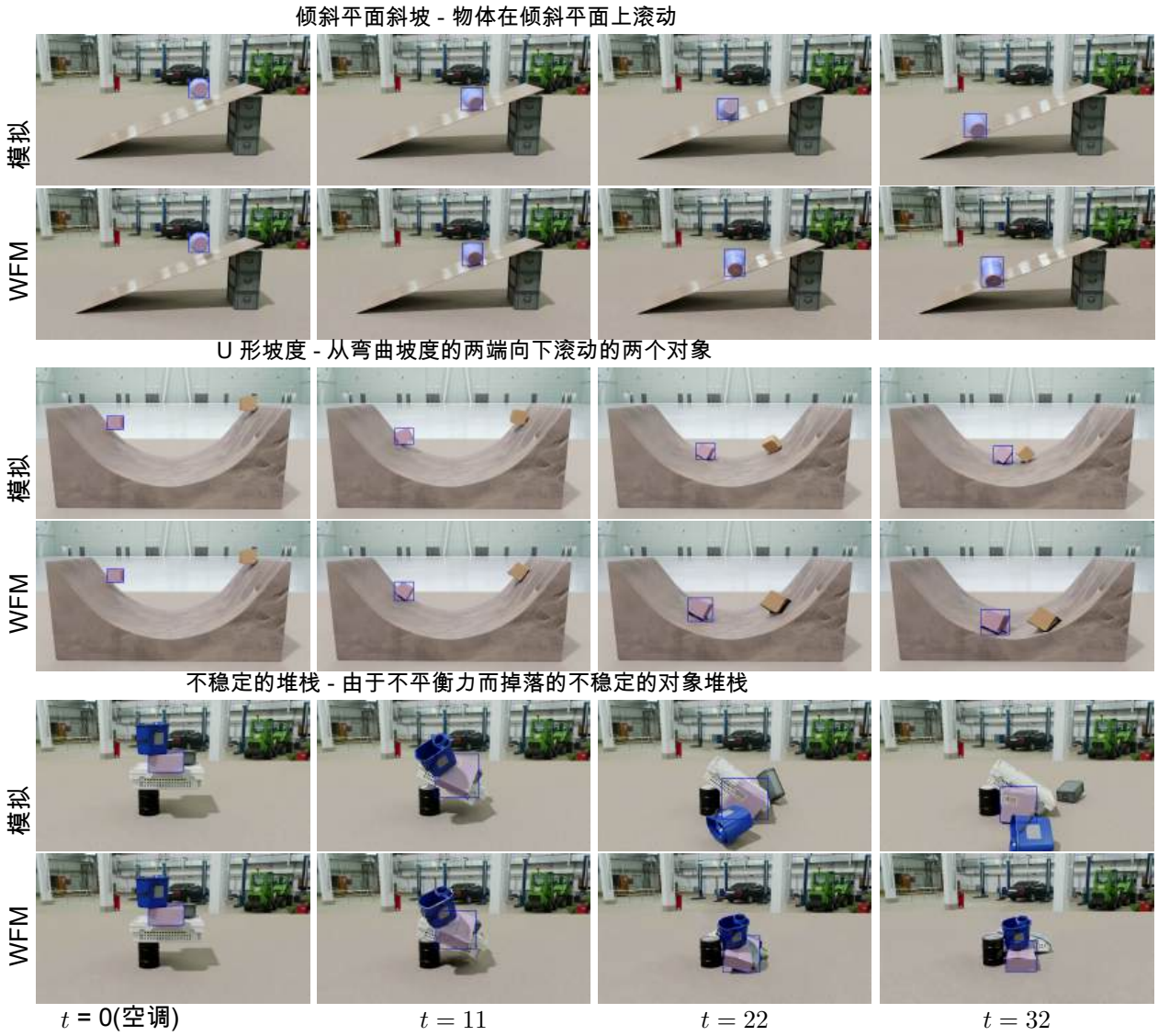


图 20: **模拟中的物理场景展开 *vs* .预先训练的 WFM 。** 我们展示了从参考（物理正确的）模拟（每组的第一行）和Cosmos-1.0-Diffusion-7B-Video2World滚动（每组的第二行）中获得的三种复杂性逐步增加的示例场景。我们基于9帧和一个关注模拟对象动力学状态的提示来条件化WFM。对于计算我们的对象级指标（平均IOU），我们每个示例展示一个跟踪对象（蓝色边界框和掩码）。

表20：物理对齐结果。我们根据像素级、特征级和对象级指标，比较了不同版本的Cosmos WFMs在预测物理场景未来状态时的准确性。这些度量是在由Cosmos WFMs自回归变体支持的最大长度为33帧的情况下计算的。

模型条件	PSNR	像素级 SSIM ↑	特征级 DreamSim ↑	对象级 平均 IoU ↑
宇宙 - 1.0 - 扩散 - 7B - Video2World 提示 + 1 帧	17.34	0.538	0.836	0.332
Cosmos - 1.0 - 扩散 - 7B - Video2World 提示 + 9 帧	**21.06**	**0.691**	0.859	0.592
宇宙 - 1.0 - 扩散 - 14B - Video2World 提示 + 1 帧	16.81	0.521	0.836	0.338
Cosmos - 1.0 - 扩散 - 14B - Video2World 提示 + 9 帧	20.21	0.635	0.860	**0.598**
宇宙 - 1.0 - 自回归 - 4B 1 帧	17.91	0.486	0.827	0.394
Cosmos - 1.0 - 自回归 - 4B 9 帧	18.13	0.482	0.859	0.481
Cosmos - 1.0 - 自回归 - 5B - Video2World 提示 + 1 帧	17.67	0.478	0.818	0.376
Cosmos - 1.0 - 自回归 - 5B - Video2World 提示 + 9 帧	18.29	0.481	0.864	0.481
宇宙 - 1.0 - 自回归 - 12B 1 帧	17.94	0.486	0.829	0.395
Cosmos - 1.0 - 自回归 - 12B 9 帧	18.22	0.487	**0.869**	0.487
Cosmos - 1.0 - 自回归 - 13B - Video2World 提示 + 1 帧	18.00	0.486	0.830	0.397
Cosmos - 1.0 - 自回归 - 13B - Video2World 提示 + 9 帧	18.26	0.482	0.865	0.482

视频输出直接生成自WFM。因此，为了产生未来的观测结果，我们将WFM条件化于真实视频的前几帧（要么是1帧，要么是9帧）。当适用时，我们还会将一个WFM进一步条件化于一个文本提示（通过在条件帧上使用专有的视觉语言模型进行配对标题获得），重点关注过去观测中模拟对象的动力学状态。请参见 Fig. 20
对于模拟场景与预测场景的一些示例。对于评估 ， 我们使用以下指标 ：

1. **像素级度量** . 对于像素级别的比较，我们计算峰值信噪比（PSNR）和结构相似性指数测量值（SSIM），以比较WFM滚动预测帧与真实视频参考帧。2. 。对于稍微高级别的语义比较 ， 我们计算 DreamSim 相似度

功能级别度量
分数（ Fu et al. ， 2023 ） ， 预测帧和参考帧之间的特征相似性度量。 3. **对象级度量** 最后，因为我们最关心的是感兴趣对象如何受到持续物理现象的影响，我们使用跟踪来计算对象级别的指标，这些指标可以消除混淆因素（背景变化、视觉质量等）。 *etc*). 由于测试条件是合成生成的，我们能够访问场景中动态对象的真实实例分割掩码。使用 SAMURAI（ Yang et al. ， 2024), 我们在第一帧中传播地面真实实例掩码，并将其应用于剩余的预测视频帧以提取轨迹，从而允许我们量化对象级别的指标。我们计算每帧和每个感兴趣对象的地面真实值与预测对象掩码之间的交并比（IoU）。

我们在视频的不同帧、评估集中的不同视频以及四组随机种子的展开中计算这些指标的平均值。PSNR和SSIM在所有帧上进行计算，但不包括用于条件处理的帧。

结果。 物理比对的定量结果概述在 Tab. 20 基于定量和定性的结果，我们得出以下观察。不出所料，随着条件输入帧数的增加，模型能够更好地预测整体物体运动学（这使得我们能够更好地推断出速度和加加速度等一阶和二阶量）。

从表格中，我们还发现，在9帧条件设置下，我们的基于扩散的WFMs在像素级预测性能上优于我们的自回归WFMs。这与我们观察到的基于扩散的WFMs生成的视频具有更高视觉质量的现象相吻合。我们也注意到，我们的结果并未表明更大的模型在物理对齐方面表现更好。尽管我们观察到更大的模型生成的视频具有更高的质量。

更高的视觉质量下，所有的WFMs都面临着物理学遵守性的挑战，需要更好的数据整理和模型设计。

一般而言，我们观察到上述刚体模拟已经测试了我们工作流模型（WFMs）的极限，这些模拟作为识别特定故障案例的重要工具。这些问题范围从低级别的对象瞬时性问题（物体的自发出现和消失）、变形（形状变化）到更复杂的问题如不合理的运动学、重力违背等。 *etc* 我们认为这类结构化的模拟提供了测试物理一致性的一个有用方法论。因此，我们计划随着时间的推移不断改进它们，通过整合更复杂的场景、增强光真实感以缩小模拟与现实之间的差距（因为WFM预训练数据由真实的视频组成），以及细化评估指标以进行全面的物理理解评估。

6. 培训后的世界基金会模型

在本节中，我们展示了如何对我们的Cosmos WFMs进行微调以支持多种Physical AI应用。我们包括了从使用摄像头控制进行后训练以实现可导航的3D视觉世界生成、使用动作控制在两个不同的机器人平台上对两种不同的机器人操作任务进行后训练、以及使用多视图支持进行自动驾驶代理训练等示例。

表 21: 第 6 节讨论的训练后 WFM 的图谱。

Section	模型	条件 (s)
Sec. 6.1	宇宙 - 1.0 - 扩散 - 7B - Video2World - 样本 - CameraCond	文本 + 图像 + 相机
Sec. 6.2	Cosmos - 1.0 - 自回归 - 7B - Video2World - 样本 - 指令	文本 + 视频
Sec. 6.2	Cosmos - 1.0 - 扩散 - 7B - Video2World - 样本 - 指令	文本 + 视频
Sec. 6.2	Cosmos - 1.0 - 自回归 - 7B - Video2World - Sample - ActionCond	动作 + 视频
Sec. 6.2	Cosmos - 1.0 - 扩散 - 7B - Video2World - 样本 - ActionCond	动作 + 视频
Sec. 6.3	宇宙 - 1.0 - 扩散 - 7B - Text2World - 样本 - 多视图	Text
Sec. 6.3	Cosmos - 1.0 - Diffusion - 7B - Text2World - Sample - MultiView - TrajectoryCond	文本 + 轨迹
Sec. 6.3	宇宙 - 1.0 - 扩散 - 7B - Video2World - 样本 - 多视图	文本 + 视频

Tab. 21 提供了本节不同子部分中讨论的后训练WFMs列表。我们还列出了条件输入以突出运行模式。请注意，对于每个模型，我们在其名称后添加“-Sample”，以强调我们的目标是提供我们预训练WFMs的示例应用。这些模型绝非完整的系统或任何实际应用场景下的生产模型。开发人员需要根据其自定义数据集对WFMs进行微调，以适应其目标应用的Physical AI设置。

6.1. 用于摄像机控制的培训后 WFM

通过相机姿态调整，我们将相机控制集成到Cosmos-1.0-Diffusion-7B-Video2World中，使其成为一个有效的三维世界模拟器。我们将此结果后训练的结果称为Cosmos-1.0-Diffusion-7B-Video2World-Sample-CameraCond。我们专注于从单个参考输入图像生成三维世界，利用相机控制从指定的相机轨迹产生时间上连贯且三维一致的视频仿真，其中视角的变化与场景的底层三维结构相一致。

6.1. 1. Dataset

我们使用 DL3DV - 10K (Ling 等人。 ， 2024), 用于此任务的大规模静态场景视频数据集。作为预处理步骤，我们将所有视频分割成每段包含256帧的片段。为了在每个片段内的所有帧中密集获取相机姿态标注，我们使用GLOMAP对分割后的片段进行结构从运动（Structure-from-Motion, SfM）计算。 Pan 等人。 ， 2025 ）。我们将第一帧的相机姿势设置为身份变换 ， 并计算所有的相对相机姿势

后续帧。我们还使用了自有的VLM（视频语言模型）来为视频添加字幕，从而获取描述视频为静态场景的文字提示。

6.1. 2. Fine - tuning

我们通过将采样的潜在嵌入与 Plucker 嵌入连接来添加相机控制调节（ Sitzmann 等人。 ， 2021), 其空间维度与潜在嵌入相同。具体来说，在给定相机姿态的情况下，我们通过计算普吕克坐标来实现这一过程。

$$r = (d \ , \ m) \in R^{6} \text{ 其中 } m = c \times d \ , \tag{11}$$

其中 \(c \) 是相机中心位置，\(d \) 是每个潜在像素（其中潜在嵌入被视为下采样图像）的单位射线方向。所有相机姿态相对于初始帧而言是相对的。Cosmos-1.0-Tokenizer-CV8x8x8 用于 Cosmos-1.0-Diffusion-7B-Video2World 模型中，具有时间压缩率 8。 × ，因此，在每8帧中，我们使用第4帧的Plücker嵌入与相应的潜在表示进行连接。

我们将训练视频的输入帧大小调整为 704 × 1252 并将它们填充到 704 × 1280 像素，带反射。我们在训练过程中采样 57 帧。训练目标和其他超参数与基础扩散 WFM 训练相同（注：括号内的内容为原文中的直接引用，保持不变）。 Sec. 5.1.3).

6.1. 3. 评价

我们假设给出了世界的单个参考图像，并从输入图像生成未来的滚动结果，将其作为视频呈现。我们将此与CamCo进行比较。 徐等人。 ， 2024), 在此设置下最先进的用于相机可控视频生成的模型。为了进行公平比较，我们使用了同样在DL3DV-10K数据集上进行微调的CamCo模型。 Ling 等人。 ， 2024 ）训练集。由于我们经过训练后的WFM生成了57帧，而CamCo只能生成14帧，因此我们将比较相同的57帧轨迹，并通过时间下采样4倍进行对比。 ×
适用于 CamCo 。 CamCo 的视频分辨率限制为 256 × 256. 我们还最大限度地集中裁剪输入图像和测试帧以进行评估。

对于测试数据 ， 我们使用来自 RealEstate10K 的相同 500 个样本（ 周等人。 ， 2018 ）先前在中描述的测试集 Sec. 5.3.1 我们使用初始框架作为参考图像，并将数据集中提供的摄像机轨迹作为摄像机控制输入，同时对这些轨迹进行重新缩放，使得两条轨迹两端之间的距离归一化为1。

指标。 以下 徐等人。（ 2024 对于视频质量，我们使用Fréchet inception distance（FID）进行评估。在两个方面评估经过训练的世界模型的相机可控性：生成视频的质量和3D一致性。 Heusel 等人。 ， 2017 ）和 Fr é chet 视频距离 (FVD) (Unterthiner 等人。 ， 2019 ）分别评估框架和视频级别的质量。我们使用与参考视频相同的测试数据来计算指标（请注意，这些指标不用于像素级比较）。

对于 3D 一致性 ， 我们通过运动结构的能力进行评估（ Pan 等人。 ， 2025 ； 舍恩伯格等人。 ， 2016 ； 舍恩伯格和弗拉姆 ， 2016 ）用于重新估算相机姿态，并将结果与输入的相机控制轨迹进行比较。给定 i 在视频帧中 ， 我们将摄像机轨迹误差量化为两个项 ： 平均旋转误差 E_{rot} 和翻译错误 E_{trans} ， 分别定义为

$$E_{rot}^{(i)} = \arccos\left(\frac{\text{轨迹}(R_i^{\top} \hat{R}_i) - 1}{2}\right), \quad E_{trans}^{(i)} = \|t_i - \hat{t}_i\| \tag{12}$$

where $\hat{R}_i$ and $\hat{t}_i$ 是输入的旋转和平移 i - th 帧 (用作地面实况) ， 和 R_i and t_i 重新估算的数量。为了考虑从相机姿态估计结果到相似变换的不确定性，我们遵循以下步骤： Lin et al. (2021 ） 并对预测的相机轨迹进行 Procrustes 分析 ， 以与地面真相对齐。

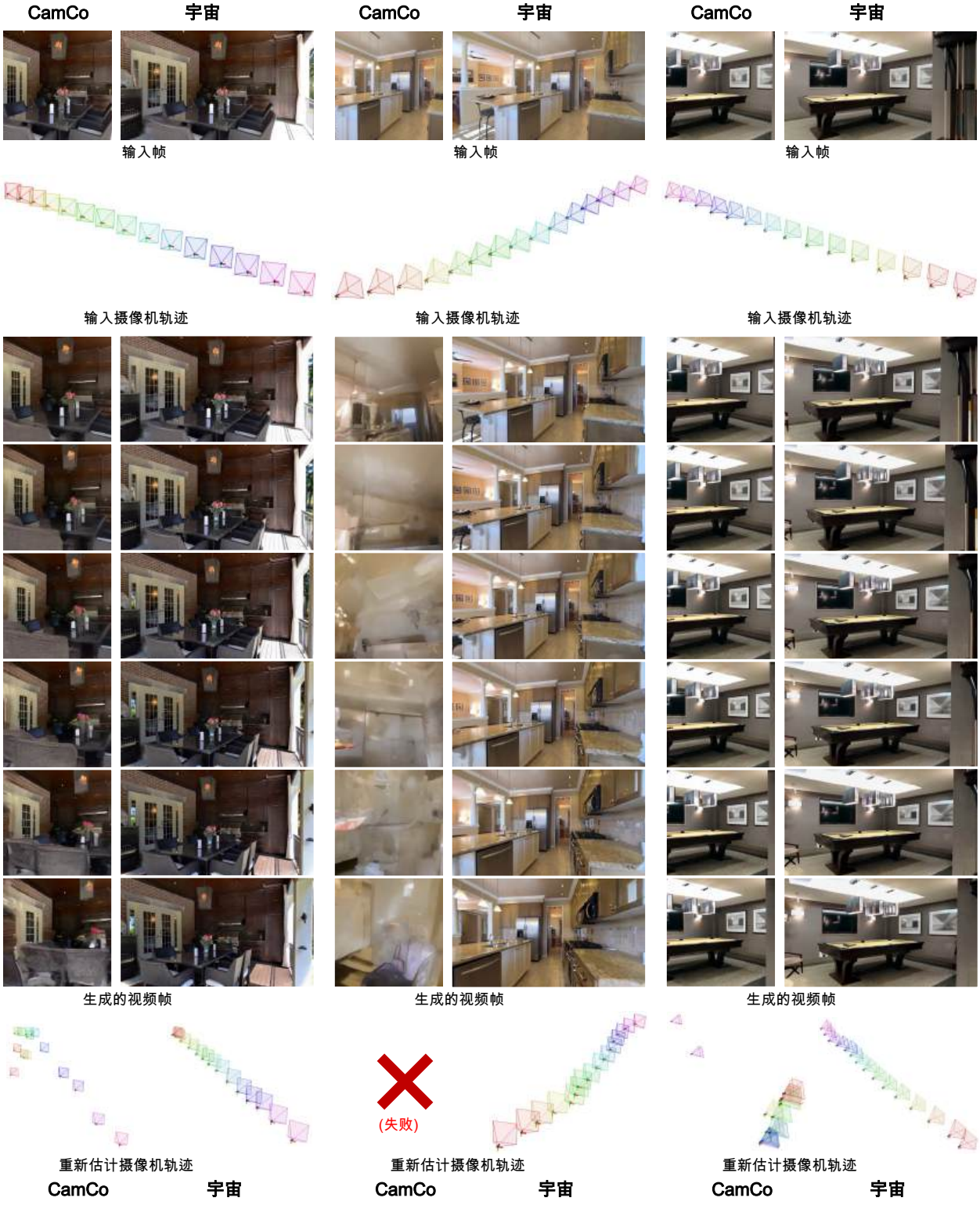


图 21: **摄像机控制模型的定性比较** . 根据输入框架和相机轨迹（按时间从红色到紫色着色），我们比较了 Cosmos-1.0-Diffusion-7B-Video2World-Sample-CameraCond 与 CamCo（CamCo）的表现。 徐等人。 ， 2024 ）生成未来帧以及重新估计的相机姿态。CamCo 受数据分布变化的影响，往往会生成不准确的轨迹 ，甚至产生离分布的图像合成，导致无法估计的相机姿态。相比之下，Cosmos 相机控制模型能够成功地 生成与相机控制输入对齐的未来帧，同时保持高质量的视频和高三维一致性。

表 22: 训练后 WFM 与相机对照的定量比较。

	摄像机轨迹对齐视频生成质量				
方法	成功率 (%) ↑	姿势估计旋转误差 (°) ↓	平移误差 (°) ↓	FID ↓	FVD ↓
CamCo (徐等人。，2024)	43.0%	8.277	0.185	57.49	433.24
宇宙 - 1.0 - 扩散 - 7B - Video2World - 样品 - CameraCond	**82.0%**	**1.646**	**0.038**	**14.30**	**120.49**

比较。 我们将结果呈现在 Tab. 22 。首先，我们后训练的WFM能够生成真实且连贯的3D世界。这体现在较低的FID/FVD分数（更高的视觉质量）和较高的摄像机姿态估计成功率上。Cosmos-1.0-Diffusion-7B-Video2World-Sample-CameraCond在摄像机控制方面表现更佳，因为摄像机轨迹重新估计与原始控制输入的接近程度显著提高。

我们还提供了视觉比较 Fig. 21 尽管CamCo在生成超出输入图像的内容方面面临挑战，Cosmos-1.0-Diffusion-7B-Video2World-Sample-CameraCond能够生成符合三维世界结构的视觉效果。值得注意的是，这两种模型均在DL3DV-10K上进行了后续训练，并在RealEstate10K数据集上进行了评估，这在训练和测试之间引入了显著的数据分布差异。Cosmos模型成功克服了这一数据分布差异，并展示了其泛化到未见过的输入相机轨迹的能力。

定性结果。 Fig. 22 显示了我们在相机上的类似操纵杆的控制输入的结果 ，包括 *向前移动* ，*向后移动* ，*向左旋转* ，and *向右旋转* 这展示了应用场景，即通过使用操纵杆控制模型生成未来的视频帧来导航模拟世界。物理人工智能代理也可以利用这种控制预测在不同情景下世界的未来。

为了展示生成的多样性，我们展示了使用相同输入图像和相机控制但在不同随机种子下生成的结果。 Fig. 23 "在当前状态下，Cosmos-1.0-Diffusion-7B-Video2World-Sample-CameraCond 能够生成不同的世界，同时保持视频中的三维空间和时间一致性。这可以用来模拟给定当前状态的不同可能未来。"

6.2. Post - training WFM for Robotic Manipulation

一个世界模型有潜力成为机器人操作的强大规划器和模拟器。在这里，我们展示了如何针对两项任务微调我们预先训练好的WFMs：（1）基于指令的视频预测；（2）基于动作的下一帧生成。 **基于指令的视频预测** 输入是机器人的当前视频帧以及一条文本指令，输出则是机器人按照指令执行后的预测视频。对于 **基于动作的下一帧预测** 输入当前机器人视频帧以及当前帧与下一帧之间的动作向量，输出是预测的下一帧，展示机器人执行指定动作的结果。给定一系列动作序列，模型可以自回归地运行以预测机器人执行这些动作的视频。

6.2. 1. Datasets

我们为上述两个任务管理两个数据集。对于 **基于指令的视频预测** 我们创建了一个内部数据集，名为Cosmos-1X数据集。该数据集包含大约200小时由1x.Tech公司的人形机器人EVE捕获的第一人称视频。 Technologies ， 2024 ）执行各种任务 ， 包括导航 ， 折叠衣服 ， 清洁桌子 ， 拾取物体 ， *etc* 。从原始视频中 ， 我们选择了大约 12 个 ， 000集时长从1秒到9秒不等的视频片段。每集配有一页句子的指令标签，后续通过 proprietary VLM 进行上采样。视频以30帧/秒的速度录制，分辨率为512。 × 512.

For **基于动作的下一代框架** ，我们使用了一个名为 Bridge 的公共数据集（ Ebert 等人。 ， 2022 ） ， 与

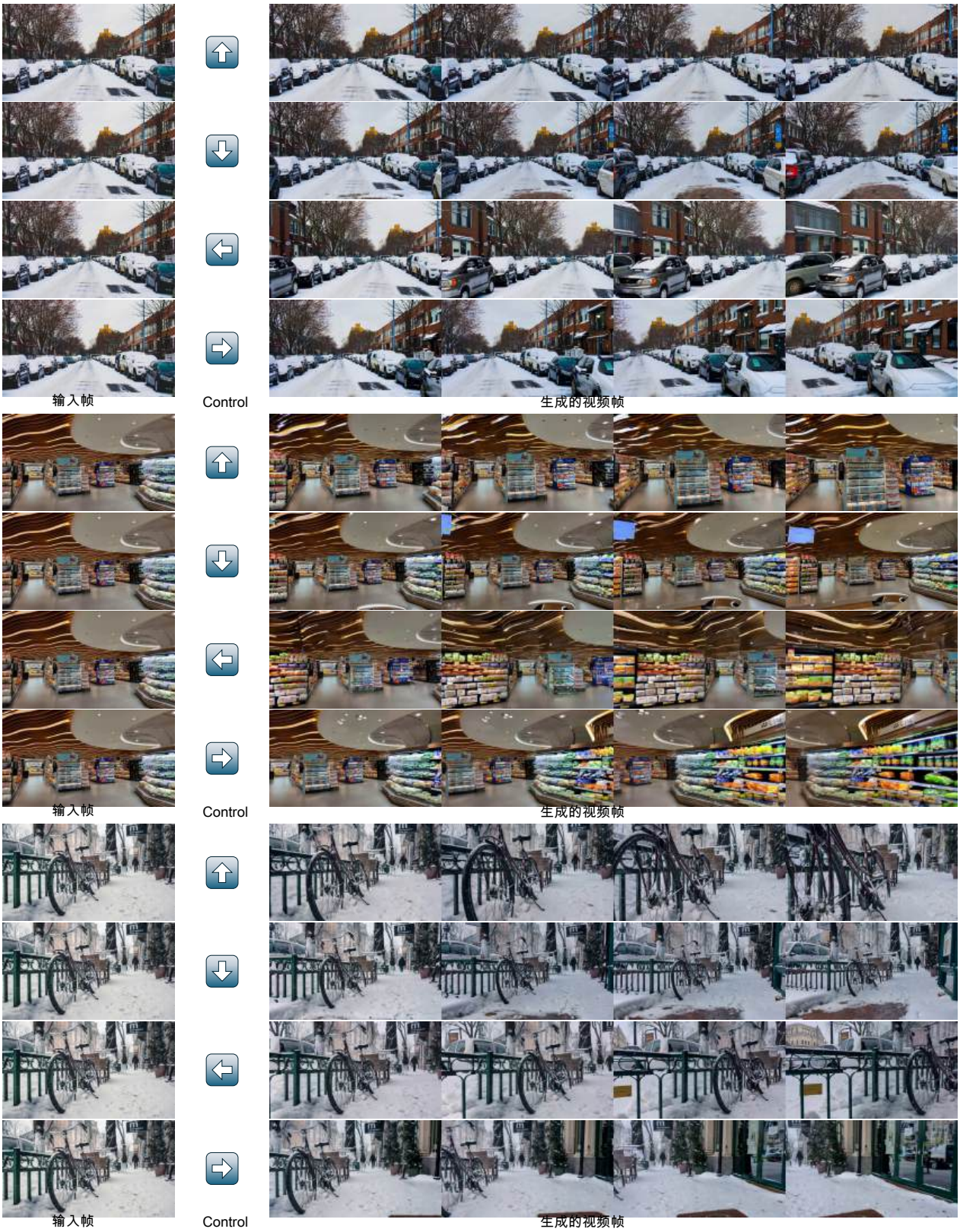


图 22: **Cosmos - 1.0 - 扩散 - 7B - Video2World - 样品 - 带操纵杆控制的 CameraCond 结果** 。对于每个输入帧(最左列) ， 我们应用使用类似操纵杆的控件创建的 4 种不同的相机轨迹 ： *向前移动* ， *向后移动* ， *向左旋转* ，and *向右旋转* 我们从生成的视频中可视化第 14 、 28 、 42 和 57 帧。

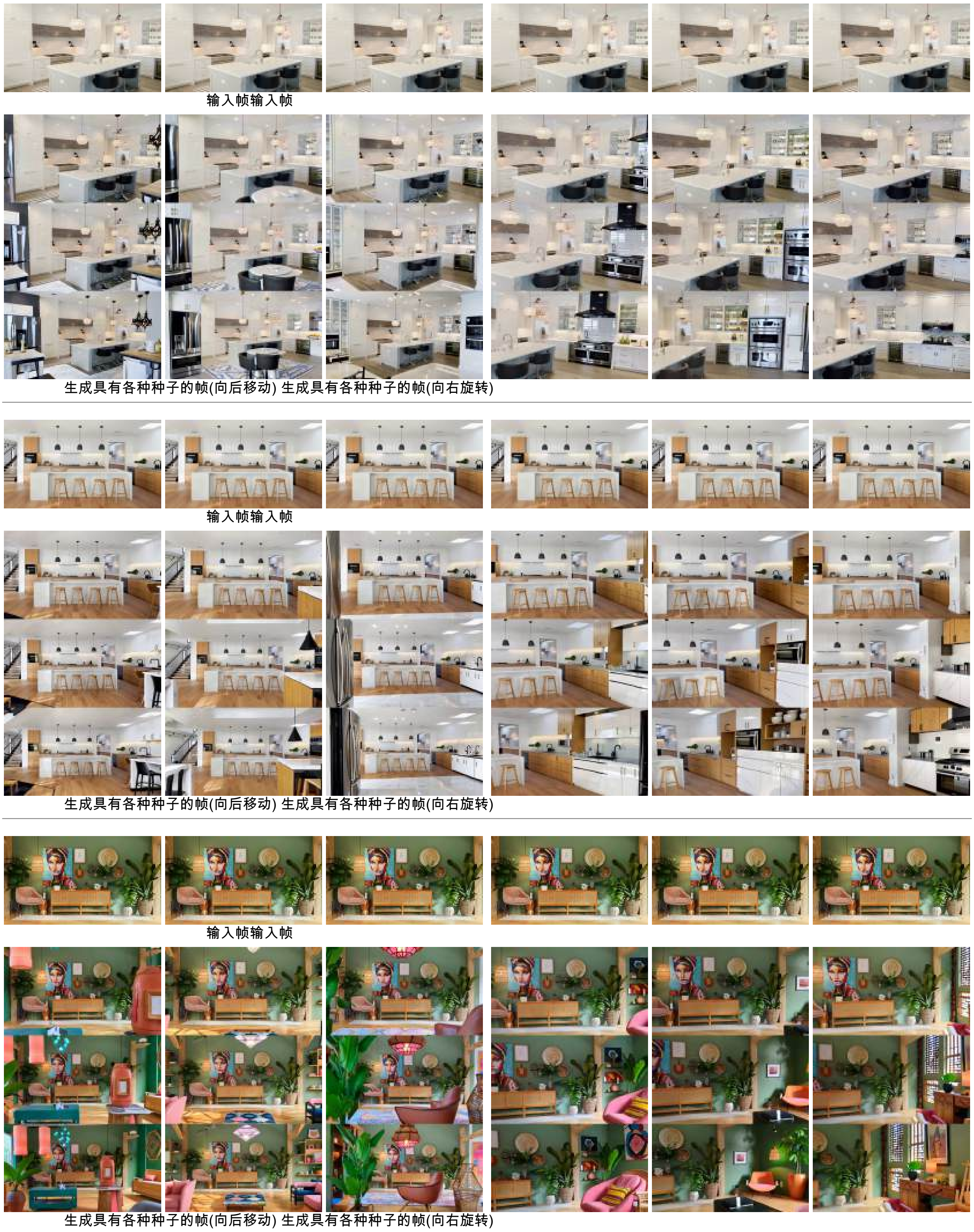


图 23: **具有不同种子的 Cosmos - 1.0 - 扩散 - 7B - Video2World - Sample - CameraCond 结果。** 我们展示了在给定相同的输入图像和相机条件的情况下，通过我们的相机控制模型模拟多种未来场景的能力。对于每个组别，我们应用相同的输入帧和使用操纵杆创建的相机条件。第一个组别显示 *向后移动* 第二组显示 *向右旋转* 在每个组中，我们展示了每列使用3个不同随机种子生成的视频。我们可视化了生成视频中的第19、38和57帧。

与以前的工作相同的配置（ 朱等人。 ， 2024 ）进行比较。 Bridge 数据集包括大约 20 ， 000 集第三人称视角的机器人手臂在厨房环境中执行不同的任务 ， 视频为 320 × 256分辨率，每秒捕获5帧。对于每一帧视频，对应的动作被定义为夹具坐标空间中的一个7维向量（Δ... ） $\mathit{舌}, \Delta\ \mathit{舌}, \Delta\ \mathit{舌}, \Delta\ \mathit{彗}, \Delta\ \mathit{彗}, \Delta\ \mathit{彗}, \Delta$ $\mathit{丞}\ \mathit{哥}\ \mathit{舌}$ 如 OpenVLA 中的夹持器（ Kim 等人。 ， 2024 ）.

6.2. 2. 微调

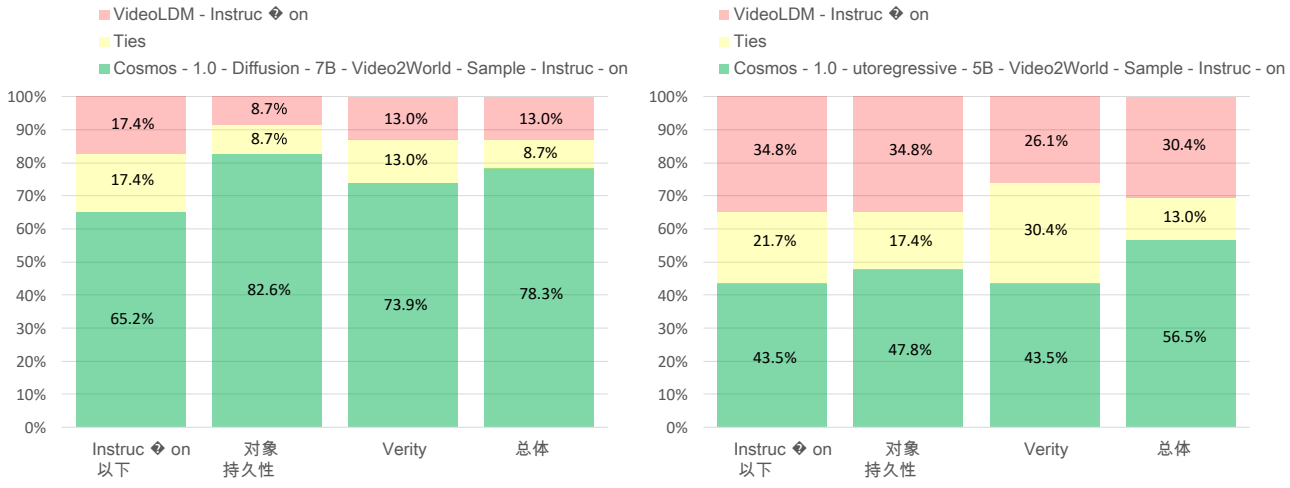
我们微调了我们的宇宙 - 1.0 - 扩散 - 7B - Video2World（ Sec. 5.1 ）和 Cosmos - 1.0 - 自回归 - 5B - Video 2World（ Sec. 5.2 ）用于基于指令的视频预测和基于动作的下一帧预测任务。

For **基于指令的视频预测** 我们构建了两个模型，基于基础WFMs。第一个模型称为Cosmos-1.0-Diffusion-7B-Video2World-Sample-Instruction，第二个模型称为Cosmos-1.0-Autoregressive-5B-Video2World-Sample-Instruction。我们计算了指令的T5嵌入，并将其添加到基础模型的微调中，通过交叉注意力机制进行。

For **基于动作的下一帧预测** 我们还基于基础WFMs构建了两个模型。第一个模型名为Cosmos-1.0-Diffusion-7B-Video2World-Sample-ActionCond，第二个模型名为Cosmos-1.0-Autoregressive-5B-Video2World-Sample-ActionCond。

由于动作是一种在预训练过程中未曾遇到的新模态，我们在模型中引入了额外模块以进行条件处理。对于Cosmos-1.0-Autoregress-5B-Video2World-Sample-ActionCond，我们增加了一个动作嵌入MLP，将动作向量投影到张量中，然后通过交叉注意力机制将其整合进模型。对于Cosmos-1.0-Diffusion-7B-Video2World-Sample-ActionCond，我们也增加了一个动作嵌入MLP，用于预测动作并将其转换为张量，但通过将其添加到DiT模块的时间戳嵌入中来整合到模型中。

6.2. 3. 评价



(a) 宇宙 - 1.0 - 扩散 - 7B - 视频 2 世界 - 样本 - 指令 *vs* . VideoLDM - Instruction.

(b) Cosmos - 1.0 - 自回归 - 5B - Video2World - Sample - 指令 *vs* . VideoLDM - Instruction.

图 24: **Cosmos - 1X 数据集上基于指令的视频预测的人工评估结果** 研究结果显示，与基线模型（ VideoLDM-instruction ）相比，我们微调后的模型（ Cosmos-1.0-Diffusion-7B-Video2World-Sample-Instruction 和 Cosmos-1.0-Autoregressive-5B-Video2World-Sample-Instruction ）在四个评估维度上的偏好度更高。

For **基于指令的视频预测** ， 我们微调 VideoLDM（ Blattmann 等人。 ， 2023 ）在Cosmos-1X数据集上进行训练，并将VideoLDM-Instruction作为基准进行比较。为了评估模型的视频生成性能，我们定义了以下维度：

输入帧指令条件生成输入帧指令条件生成

提示 ： 通过将书籍垂直放置在书架上来整理书籍。

提示 ： 从整洁的工作台上的盒子中抓住并抬起绿色物体。

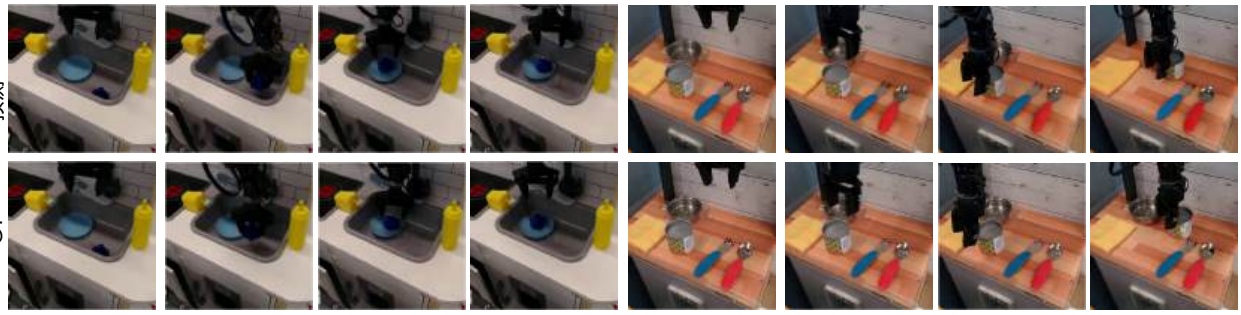
提示 ： 将绿色织物折叠在桌子上。

提示 ： 从存储货架中检索一个盒子 ， 使用其铰接式的手在仓库设置。

宇宙 - 1.0 - 扩散 - 7B - 视频 2World - 样本 - 指令宇宙 - 1.0 - 自回归 - 5B - 视频 2World - 样本 - 指令

图 25: **Cosmos - 1X 数据集上基于指令的视频预测样本** 。左边是结果 Cosmos - 1.0 - Diffusion - 7B - Video2World - Sample - 指令模型 ， 右边是 Cosmos - 1.0 - 自回归 - 5B - Video2World - 样本 - 教学模式。

输入帧预测帧输入帧预测帧



宇宙 - 1.0 - 扩散 - 7B - 视频 2World - 样本 - 行动 Cond 宇宙 - 1.0 - 自回归 - 5B - 视频 2World - 样本 - 行动 Cond

图 26: **Bridge 数据集上基于动作的下一帧预测样本** 。左边是结果 Cosmos - 1.0 - Diffusion - 7B - Video2World - Sample - ActionCond 模型的结果 ， 右边是 Cosmos - 1.0 - 自回归 - 5B - Video2World - Sample - ActionCond 模型。如图所示 ， 预测的视频帧密切匹配这两种型号的 GT 视频帧。

• **Instruction following** ：生成的视频是否与输入语言指令对齐？• **对象持久性** ：场景中存在的对象是否保留在整个生成的视频中？ **Verity** ： 生成的视频是否忠实地代表现实世界 ， 而没有意外的想象对象 ？ **总体** ： 生成的视频对于机器人进行相应的计划是否合理 ？

人类评估者被指派观察由不同模型生成但具有相同语言指令的一对匿名视频，并从上述列出的维度进行比较。一组十名人类评估者对23个测试集进行了评估。统计结果总结如下： Fig. 24 .

如所示，我们在四个评估维度上发现，Cosmos-1.0-Diffusion-7B-Video2World-Sample-Instruction 和 Cosmos-1.0-Autoregressive-5B-Video2World-Sample-Instruction 的表现均优于 VideoLDM-Instruction。Cosmos-1.0-Diffusion-7B-Video2World-Sample-Instruction 达到了 78 ． 与 13 相比 ， 总体偏好为 3% ． 0% 对于 VideoLDM-Instruction. Cosmos-1.0-Autoregressive-5B-Video2World-Sample-Instruction 在性能上也超过了基于扩散模型的 VideoLDM-Instruction。对于两种微调后的时空特征提取器（ WFMs ），部分预测视频帧的结果展示在 Fig. 25 ， 这显示了预测的质量

视频。

For **基于动作的下一帧预测** ， 我们在 Bridge 数据集上微调了我们的模型。作为基线 ， 我们微调了 IRASim (朱等人。 ， 2024 为了推导出基于行动的下一帧预测模型IRASim-Action，我们以自回归的方式进行下一帧预测来生成视频。为了评估视频生成的质量，我们将生成的视频与从官方桥梁测试集随机选取的100个集合并的真实视频进行比较。

表 23: 桥数据集上基于动作的下一帧预测的评估。

方法	PSNR ↑	SSIM ↑	潜在 L2 ↓	FVD ↓
IRASim - Action	19.13	0.64	0.38	593
Cosmos - 1.0 - 自回归 - 5B - Video2World - Sample - ActionCond	19.95	0.80	0.36	434
宇宙 - 1.0 - 扩散 - 7B - Video2World - Sample - ActionCond	**21.14**	**0.82**	**0.32**	**190**

计算的度量汇总为 Tab. 23 ， 包括 PSNR 、 SSIM 、潜在 L2 (朱等人。 ， 2024), 和 FVD。如所示，Cosmos-1.0-Autoregressive-5B-Video2World-Sample-ActionCond 和 Cosmos-1.0-Diffusion-7B-Video2World-Sample-ActionCond 模型均优于基线模型（IRASim-Action）。一些预测的视频帧被呈现。 Fig. 26 ， 它显示了预测视频的质量与实际情况的比较。

6.3. Post - training WFM for Autonomous Driving

一个适用于真实环境驾驶场景的全球模型有潜力成为训练自动驾驶代理的强大模拟引擎。由于大多数自动驾驶车辆配备了多个摄像头，分别朝向不同的方向，因此，为自动驾驶车辆理想的全球模型也应具备多视角特性，并且最好能够与目标车辆传感器的具体配置相匹配。在这里，我们展示了如何对预训练的WFM进行微调，以创建适用于自动驾驶任务的多视角全球模型。

6.3. 1. Dataset

我们整理了一个内部数据集，称为真实驾驶场景（RDS）数据集。该数据集包含约360万段20秒的全景视频片段（相当于约20 ， 000小时的数据是使用NVIDIA内部驾驶平台采集的。每段视频记录了六个摄像头视角的画面：前方、左侧、右侧、后方、后左和后右。此外，数据集还包括我们用于构建轨迹数据的自我运动信息。我们使用前方摄像头视频记录的时间戳来同步其他所有视角的画面。

这个数据集是从一个大型标注数据语料库中选择出来的，以匹配数据属性的目标分布。具体的属性标签包括：

• 竞标者车辆密度 (*e.g* 。 ， 无 ， 低 ， 中 ， 高) • 天气 (*e.g* 。 ， 晴朗 ， 下雨 ， 下雪 ， 雾) • 照明 (*e.g* 。 ， 白天 ， 黑夜) • 自我车速 (*e.g* 。 ， 站立 ， 低 ， 局部 ， 高速公路速度) • 自我车辆行为 (*e.g* .高曲率、中曲率、低曲率轨迹与加速度 • 道路类型/人口密度 (基于OpenStreetMap定义：乡村、住宅区、城市）。

此外，通过第二次数据挖掘运行扩充了数据集，以确保包含稀有道路结构的片段数量不低于一定数量。 *e.g* 。 ， 收费站 ， 桥梁 ， 隧道 ， 减速带 ， *etc*). 最后，来自每个摄像头视角的视频分别添加字幕，起始模板文本字符串为：“该视频是由安装在车辆上的摄像头捕获的。摄像头面向前方/左侧/右侧/后方/后左/后右。”

6.3. 2. Fine - tuning

我们微调我们的宇宙 - 1.0 - 扩散 - 7B - Text2World (Sec. 5.1 将RDS数据集用于构建一个多视图世界模型。为了确保多个视角下的视频生成一致性，我们对描述的架构设计进行了轻微调整。 Sec. 5.1 并微调 WFM 以同时从所有六个摄像机生成视频。

我们建立了三个多视角世界模型 ， 总结为 Tab. 21 . 第一个是名为“Cosmos-1.0-Diffusion-7B-Text2World-Sample-MultiView”的多视图世界模型，它可以基于文本提示生成六种相机视角。第二个是名为“Cosmos-1.0-Diffusion-7B-Text2World-Sample-MultiView-TrajectoryCond”的模型，该模型是在“Cosmos-1.0-Diffusion-7B-Text2World-Sample-AV-MultiView”基础上构建的，并且额外接受轨迹输入作为条件输入信号。最终的模型“Cosmos-1.0-Diffusion-7B-Video2World-Sample-MultiView”则是从“Diffusion-7B-Video2World-Sample-MultiView”模型微调而来，以支持基于视频的条件输入。它通过将先前帧纳入生成过程中实现了这一点。“Cosmos-1.0-Diffusion-7B-Video2World-Sample-MultiView”可以接受“Cosmos-1.0-Diffusion-7B-Text2World-Sample-MultiView”生成的视频输出，并生成其扩展。所有这三个模型均输出分辨率为848像素的视频，每幅视频包含57帧，共计6个视图。 × 480.

独立于视图的位置嵌入和视图嵌入。 相反，我们选择不将FPS感知的3D RoPE位置嵌入扩展到包括额外的视图维度，而是使用了在文中描述的相同的位置嵌入。 Sec. 5.1 为了表示视图差异 ， 我们修改了去噪函数 $\mathcal{F}_\theta$ 通过引入全局视图嵌入作为输入，以获得额外的视角。也就是说，摄像机视角信息通过全局视图嵌入而非位置嵌入来提供。

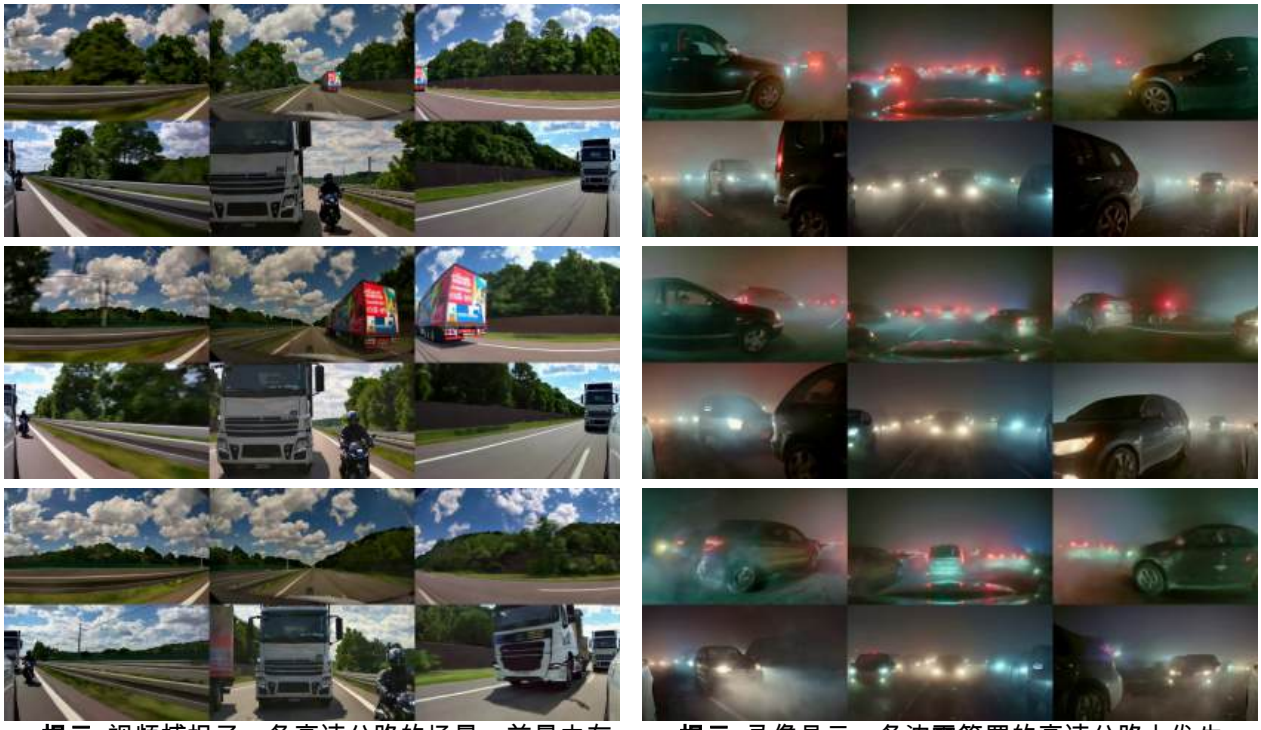
依赖于视图的交叉注意力。 在我们的多视角设置中，同一场景的六个视角会对应六种不同的视频描述。虽然我们将这六个视角作为一个整体视为扩散过程的状态，并在六个视角的所有元素之间进行自注意力操作以去除噪声，我们发现对文本输入采用视角依赖的交叉注意力是有益的。具体而言，每个视角的交叉注意力操作仅关注该特定视角的文本描述。请注意，在我们的数据集中，每个视角都有不同的视频描述。借助视角嵌入和视角依赖的交叉注意力，我们从Fine-tuning Cosmos-1.0-Diffusion-7B-Text2World 中衍生出了Cosmos-1.0-Diffusion-7B-Text2World-Sample-MultiView。

轨迹控制条件。 可选地，除了文本条件外，我们还可以微调模型生成遵循给定未来轨迹路径的视频，以实现对代理更精确的控制。这使得能够生成既符合由现实世界数据记录的驾驶轨迹，又符合输入文本描述所指定的驾驶环境的独特驾驶场景。微调后的模型为Cosmos-1.0- Diffusion-7B-Text2World-Sample-MultiView-TrajectoryCond。

我们定义一个轨迹为三维空间中的一系列64个点，表示代理从初始位置（0）出发的一系列平移。 ，0 ， 从起始点到最终目的地，每一点之间间隔0.1秒。我们计算轨迹输入的嵌入，并将其结果作为条件输入到fine-tuned Cosmos-1.0-Diffusion-7B-Video2World模型的去噪器中。我们注意到，通过提供每个间隔的动作向量 ，可以实现更精细的控制信号（遵循先前的工作）。 胡等人。 ， 2023 ； Kim 等人。 ， 2020 ， 2021 ）或在机器人操纵任务中（ Sec. 6.2 ）。我们将此类扩展留待以后的工作。

6.3. 3. 评价

我们首先在 Fig. 27 . 使用Cosmos-1.0-Diffusion-7B-Text2World-Sample-MultiView模型，我们生成了一个包含六个视角的57帧视频，然后使用Cosmos-1.0-Diffusion-7B-Video2World-Sample-MultiView模型将其扩展至201帧。 Fig. 28 我们展示如何预训练的世界模型提升了泛化能力，使得从RDS数据集生成罕见或域外场景成为可能，例如在河流上驾驶。最后， Fig. 29 展示了来自Cosmos-1.0-Diffusion-7B-Text2World-Sample-MultiView-TrajectoryCond的结果，其中ego车辆准确遵循输入指令。



提示 视频捕捉了一条高速公路的场景，前景中有一辆白色卡车正朝摄像机方向行驶。卡车有一个大的货箱区域，并且后面跟着一名佩戴全脸头盔的摩托车手。路面标有白色线条，并且右侧设有金属护栏。天空部分阴天，路边可见绿色的树木和灌木丛。该视频是从移动的车辆拍摄的，这从运动模糊和卡车及摩托车手视角的变化中可以看出来。

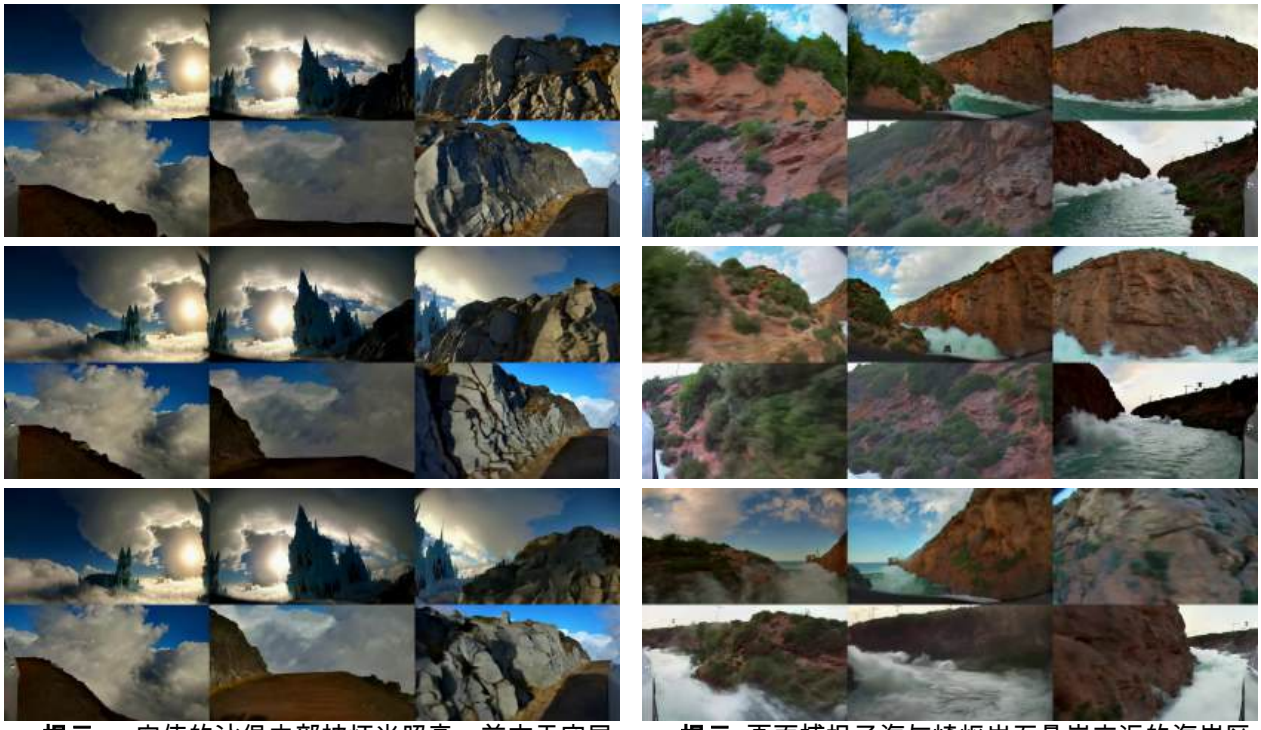
提示 录像显示一条浓雾笼罩的高速公路上发生多车连环相撞事故。由于浓雾导致能见度严重降低，只能看到前方车辆的尾灯。突然，刹车灯闪烁，车辆开始swerve并紧急停下。公路上布满了停驶和事故车辆。四周被浓雾遮蔽，进一步增加了现场的混乱和混乱。

图 27: **由Cosmos-1.0-Diffusion-7B-Text2World-Sample-MultiView生成，并通过Cosmos-1.0-Diffusion-7B-Video2World-Sample-MultiView扩展至8秒的条件文本样本** 这个图表可视化了所有六个摄像头视角，并将每个时间戳对应到一行中。左侧示例展示了高速公路场景，其中一辆摩托车正行驶在一辆大型卡车旁边。右侧示例则展示了ego车辆在大雪天跟随一辆轿车右转的场景。

轨迹。

对于定量结果 ， 作为基准 ， 我们遵循相同的微调配方来微调 VideoLDM (Blattmann 等人。 ， 2023 为了构建一个名为VideoLDM-MultiView的多视角世界模型，我们使用了一系列评估指标来衡量视频生成质量、多视角一致性以及轨迹跟随精度。为了评估视频生成质量，我们使用了1000个样本来计算评分。对于与一致性相关的指标，为了更好地理解不同模型在不同场景下的行为，我们将地面真实轨迹分为四种类型：向前移动、向左转、向右转和其他（包括静止或复杂运动）。对于每种类别，我们收集了200个样本及其相应的提示和条件，总计800个样本。以下是对这些指标的详细描述以及结果。

发电质量。 我们利用 Fr é chet 初始距离 (FID) (Heusel 等人。 ， 2017) 和 Fr é chet 视频距离 (FVD) (U nterthiner 等人。 ， 2019) 为了衡量生成视频的质量与真实视频的差距，我们首先从每段视频中提取16帧以计算每个观看者的分数。然后，我们报告各种方法下的平均分数。如所示， Tab. 24 ， 我们发现了 Co smos - 1.0 - Diffusion - 7B - Text2World -



提示 ：宏伟的冰堡内部被灯光照亮，前方天空展示着绚丽的日落和位于城堡左侧的明亮月亮，月光在场景中洒下淡淡的蓝光。快速移动的深色戏剧性云彩进一步增添了超现实的氛围。三维写实艺术风格专注于光影和纹理，创造出引人注目的视觉效果。

提示 画面捕捉了海与崎岖岩石悬崖交汇的海岸区域。前方，活力四射的蓝色波浪撞击岩石，形成白色的泡沫。悬崖呈现出棕色和绿色的色调，表明有植被以及可能的苔藓或海藻。

图 28: **由 Cosmos - 1.0 - Diffusion - 7B - Text2World - Sample 生成的文本条件样本 - 通过 Cosmos - 1.0 - Diffusion - 7B - Video2World - Sample - MultiView 将 MultiView 扩展到 8 秒** . The post - trained world model effectively preserves its generalization ability. In the left example, the ego car is 驶向冰堡 ， 而在正确的例子中 ， 自我汽车显示在河上行驶。

在两者指标上，Sample-MultiView 和 Cosmos-1.0-Diffusion-7B-Text2World-Sample-MultiView-Trajectory Cond 显著超越了 VideoLDM-MultiView，这证明了我们预训练的基于 7B 扩散的 WFM 在质量上优于 VideoLDM-MultiView 的基准。

多视图一致性。 我们使用 Sampson 错误的扩展版本 (哈特利和齐塞曼 ， 2003 ； 桑普森 ， 1982) 制定于 Sec. 5.3.1 为了量化生成的多视图视频的几何一致性。由于我们RDS数据集中真实视频共享相似的鱼眼相机内参，我们使用中值校正将关键点去畸变为一个均匀大小为960的常规针孔相机模型。 × 540 和 120 度的水平 FoV 。在此设置下 ， 将为生成的多视图视频计算两个度量 ：

1. *暂时性 Sampson 错误 (TSE)* 衡量每台相机生成的内容是否随时间保持一致。它是每个视图的相邻帧之间的Sampson误差的中值。2.
衡量多视图一致性是否随时间保持不变。它是
交叉视图 Sampson 错误 (CSE)
在不同的生成视图中平均的 Sampson 错误。中使用的基本矩阵
使用在所有时间帧上累积的关键点来估计 CSE 。

如 Tab. 24 所示 ， 我们发现 Cosmos - 1.0 - Diffusion - 7B - Text2World - Sample - MultiView 和 Cosmos -

表 24: 用于多视图驱动视频生成的训练后多视图世界模型的评估。

	生成质量多视图考虑。			
方法 FID	↓	FVD ↓	TSE ↓	CSE ↓
VideoLDM - MultiView 60.84 884.46 1.24 6.48				
宇宙 - 1.0 - 扩散 - 7B - Text2World - Sample - MultiView	**32.16**	**210.23**	0.68	2.11
宇宙 - 1.0 - 扩散 - 7B - Text2World - Sample - MultiView - TrajectoryCond	-	-	**0.59**	**2.02**
真实视频(参考) - 0.69 1.71				

表 25 ： 用于多视图驾驶的训练后多视图世界模型的轨迹一致性评估
视频生成。 TAE 的数量按 10 缩放 ^{2} 为方便起见, TFE 的单位为 cm 。

方法 TAE - ATE ↓		TAE - RPE ↓ R	TAE - RPE ↓ - t	TFE ↓
VideoLDM - MultiView 0.88 22.94 0.77 -				
宇宙 - 1.0 - 扩散 - 7B - Text2World - Sample - MultiView	0.77	**4.25**	0.29	-
宇宙 - 1.0 - 扩散 - 7B - Text2World - Sample - MultiView - TrajectoryCond	**0.54**	4.31	**0.18**	**20.20**
真实视频(参考) 0.49 4.60 0.14 13.49				

1.0-Diffusion-7B-Text2World-Sample-MultiView-TrajectoryCond 在VideoLDM-MultiView的基础上更好地实现了多视图几何一致性。生成的视频整体几何合理性显著提高，特别是对于从我们WFM微调的世界模型而言。我们还注意到，在轨迹控制条件下，这种一致性进一步提升，得益于明确的3D指导。因此，Cosmos-1.0-Diffusion-7B-Text2World-Sample-MultiView-TrajectoryCond 被评为最佳方案。

轨迹一致性: 轨迹协议错误 (TAE) 。 我们设计了一种鲁棒的多视角相机姿态估计管道 ， 类似于 Liang 等人。 （ 2024 ）基于 Teed 和 Deng （ 2021 在不确定的环境中实现稳定的市场份额增长。这样的姿态估计算法管道包含了一个在线动态遮罩生成模块和一个高度高效的密集束调整模块，能够实现对多视图相机姿态的稳健且实时的估计。我们使用此管道分别对前摄像头的姿态进行了估计，考虑了“前+左前”摄像头配置和“前+右前”摄像头配置。然后通过计算它们的轨迹误差来展示其一致性，反映多视角生成的一致性。具体地，我们计算了平移部分（RPE-t）和旋转部分（RPE-R）的绝对轨迹误差（Absolute Trajectory Error, ATE）和相对姿态误差。为了公平比较，我们将轨迹长度标准化为1.0，并排除了摄像头移动较小的情况。 *e.g.* 。 ， 一辆汽车在红灯时停下来) 。

如所示 Tab. 25 ，结果与多视图几何一致性研究的发现一致，表明从Cosmos WFM微调的世界模型轨迹一致性明显优于从VideoLDM-MultiView微调的世界模型。我们注意到，经过后处理训练的Cosmos世界模型的轨迹一致性接近真实世界的视频。

轨迹一致性 ： 跟踪错误 (TFE) 。 此外，对于将轨迹控制条件输入模型的模型，我们使用上述相同的相机姿态估计管道来利用多视图信息计算前向摄像头的姿态，并将预测的轨迹与真实轨迹条件进行比较。这衡量了模型跟踪真实轨迹条件的能力。

可视化轨迹输入框 25 框 50 框 75 框 100

图 29: **来自 Cosmos - 1.0 - Diffusion - 7B - Text2World - Sample - MultiView - TrajectoryCond 的轨迹条件生成样本** . 根据左最列的轨迹输入，我们生成遵循给定轨迹的多视图视频。我们在本图中可视化了前摄像头视角。

给定的轨迹路径。如 Tab. 25 ，使用我们的 Cosmos 训练后世界模型生成的视频估计的轨迹误差仅为 < 7 厘米不如真实情况下的参考标准精确。这种显著细微的差距表明我们的模型能够准确跟随给定的轨迹路径，这对于训练自主驾驶代理至关重要。

对象跟踪一致性。 最后 ， 我们使用 YOLOv11x 应用了对象检测和跟踪 (Khanam 和 Hussain ， 2024) 生成的8秒视频上。人类注释员被要求识别跟踪算法错误解释物理上不可能场景的实例，例如两个独立对象 (*e.g* ，一个人和一辆车错误地合并成一个跟踪实体。为了评估这一点，我们为注释员提供了包含157个对象的20个随机生成视频样本。令人惊讶的是，这157个对象中没有出现任何物理上不可能的场景，这证明了我们生成的驾驶视频的物理一致性和物体持久性。

7. 护栏

为了确保我们的WFMs的安全使用，我们开发了一个全面的防护系统。该系统分为两个阶段：预防护阶段和后防护阶段。预防护阶段利用Aegis (Ghosh 等人。 ， 2024 使用关键词列表来屏蔽有害提示，并在后守护阶段通过视频内容安全性分类器和人脸模糊过滤器阻止有害视觉输出。该流程图示如下： Fig. 30 .

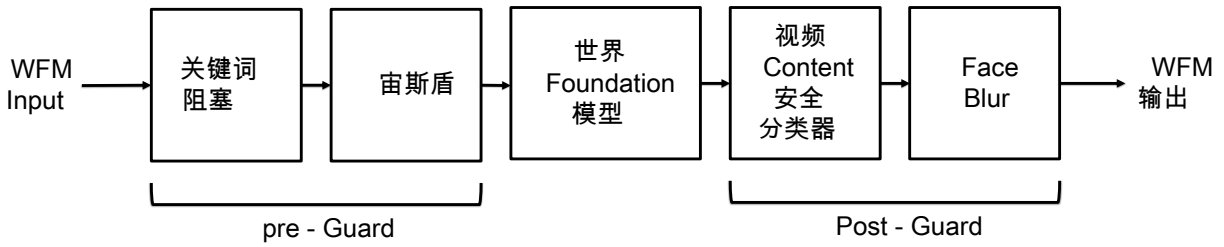


图 30: **宇宙护栏概述** 。 Cosmos Guardrail 包含 pre - Guard 和 post - Guard ， 其中 pre Guard 基于 Aegis 阻止输入（ Ghosh 等人。 ， 2024 ）和关键字 ， 而 Post - Guard 基于视频内容安全分类器阻止输出并模糊输出面。

7.1. Pre - Guard

我们的预防护是一种文本域护栏，包含基于LLM的护栏用于语义复杂的提示，以及一个基于简单黑名单的检查器用于明确不安全的关键字。

7.1. 1. 关键字阻塞

黑名单启发式方法将作为第一道防线来减轻生成不安全内容的风险。这设计了通过在提示中进行关键词搜索来阻止一个大型语料库中明确有害词汇的生成。输入词汇会使用词形还原（lemmatization）进行处理。*WordNetLemmatizer* ， 一种使用英语词法数据库的工具（ 米勒 ， 1995 ） 从其变体中提取根词。例如 ，“ *abacii* ”is“ *算盘* 这些词素化后的词语随后与硬编码的黑名单中的词语进行比较，如果发现任何亵渎之词，则整个提示将被拒绝。我们使用了一套全面的关键字集以最大程度地保护我们的用户。

7.1. 2. 宙斯盾护栏

作为第二道防线 ， 我们使用 *宙斯盾 - AI - 内容 - 安全 - LlamaGuard - LLM - 防御 - 1.0* （ Ghosh 等人。 ， 2024 ） ， 这是一个微调版本的 *Llama - Guard* （ Inan 等人。 ， 2023 ）基于NVIDIA的Aegis内容安全数据集训练，该数据集涵盖了NVIDIA广泛的安全风险分类中的13个关键类别。AEGIS 1.0有两种版本：防御性版本和许可性版本。防御性版本采用了比许可性版本更严格的权限边界。Cosmos使用Aegis的防御性版本来阻止那些试图生成有害内容的潜在有害用户提示。如果输入提示被此提示过滤器分类为不安全，则不会生成视频，并显示错误消息。

为了将Aegis用作提示过滤器，我们如果将提示归类为不安全，则将其划分为以下类别：暴力、性内容、犯罪策划、武器、滥用物质、自杀、儿童性虐待材料、仇恨言论、骚扰、威胁和污言秽语。从提示过滤的角度来看，任何未落入上述类别的提示均被视为安全。

7.2. Post - Guard

我们的后Guard是一款视野域护轨系统，包含一个生成输出的安全视频内容过滤器和一个面部模糊滤镜。

7.2. 1. 视频内容安全过滤器

视频内容安全过滤器是一个针对我们视频数据集和生成结果进行训练的帧级多类分类器。其中，某些类别被认为是安全的，而其他类别则被认为是不安全的。训练分类器的主要挑战在于平衡误报（safe内容被错误地标记为unsafe）和漏报（unsafe内容被错误地分类为safe）的问题。为了最小化分类错误，我们在训练过程中仔细平衡了数据。

我们收集了三种类型的地面真实标注数据。首先，从我们的数据集中采样一大批视频，提取帧，并使用视觉语言模型（VLM）确定其类别。接着，我们使用一组提示生成合成视频，以确保涵盖边缘案例和最少代表的内容类别。最后，人类注释员为部分数据集提供“金标准”标签，增加了验证的重要层次，并帮助我们不断 refinement 分类器的准确性。我们提取了 SigLIP（ ）。 翟等人。 ， 2023 ）为每个视频帧嵌入 ， 并在嵌入上训练一个简单的 MLP 分类器。

在推断过程中，我们为每一帧生成一个SigLIP嵌入，并随后应用分类器。如果任何一帧被分类为不安全，则整个视频将被标记为不安全。

7.2. 2. Face Blur Filter

我们使用 RetinaFace （ 邓等人。 ， 2020 使用先进的面部检测模型来识别具有高置信分数的面部区域。对于任何检测到且尺寸大于20的面部区域进行处理。 × 20像素时，我们应用像素化处理以模糊掉特定区域，同时保持整体场景布局不变，适用于物理AI应用。

7.3. Red Team Effort

我们 employs 一支专门的红队，利用标准样本和对抗样本积极探测系统，并将这些样本收集到内部攻击提示数据集中。这些视频输出由一组经过特别培训的专家注释员进行标注，以根据与分类学相关的多个危害类别对生成的视频进行分级评估。 Sec. 7.1.2 这些注释还指定了检测到不安全内容的起始和结束帧，从而生成高质量的注释。红队还独立测试了每个护栏组件，并使用有针对性的例子来识别弱点并提高边缘情况下的性能。截至出版日期，红队已经测试和标注了超过10 ， 000 个不同的提示视频对 ， 经过精心制作 ， 涵盖了广泛的不安全内容。

8. 相关工作

世界模型。 “世界模型 ”的概念起源于 Ha 和 Schmidhuber （ 2018), 哪些提出了使用神经网络模型学习现实世界的表示，以便根据当前状态和输入预测未来状态。准确的物理世界模型表示不仅能够实现对未来状态的可靠预测，还能支持更加明智的决策制定。这一通过建模来模拟物理世界的概念并不是新的；传统的自动化和机器人技术行业早已利用基于物理定律和系统辨识的数学模型来进行规划和控制算法（ Murray 等人。 ， 2017 然而，这些特定于系统的模型通常局限于低维状态空间，这限制了它们在不同系统之间的泛化能力和知识迁移能力，从而在应用于新任务或环境时限制了模型的复用性。近期深度学习领域的进展，特别是生成式人工智能的发展，使得直接从视觉观察中学习世界模型成为可能。

现代世界模型管道可以根据其骨干体系结构进行分类。大多数作品（ Hafner 等人。 ， 2019 ， 2021 ， 2023 ；汉森等人。 ， 2024 ；Kim 等人。 ， 2020 ， 2021 ）， 包括原始纸张（ Ha 和 Schmidhuber ， 2018 ）由Ha和Schmidhuber提出，采用循环神经网络在通过自编码器学习的潜在空间中建模系统状态演变。近年来的趋势则将世界模型视为视觉观察空间中的生成模型，通常表现为条件视频生成模型（ ）。 *e.g* 。 ， 动作到视频 ， 文本到视频) 。这些模型可以是自回归的（ 布鲁斯等人。 ， 2024 ；Liu et al. ， 2024 ；Micheli 等人。 ， 2023 ；罗宾等人。 ， 2023 ；Yang et al. ， 2023 ）或基于扩散的（ 阿隆索等人。 ， 2024 ；丁等人。 ， 2024 ；Valevski 等人。 ， 2024 ）， 如这项工作所考虑的。另一种有前途的方法是生成模拟（ 华等人。 ， 2024 ；Nasiriany 等人。 ， 2024 ） ， 它结合了生成式 AI 和物理模拟器来对现实世界进行建模。

训练有素的世界模型可以以各种方式应用 ， 包括验证（ 胡等人。 ， 2023 ） ， 基于规划的模型预测控制（ Bar et al. ， 2024 ；汉森等人。 ， 2024 ） ， 以及基于模型的加固

学习（阿隆索等人。，2024；丁等人。，2024；罗宾等人。，2023；Yang et al.，2023；张等人。，2024）。世界模型的有效性已经在计算机游戏等领域得到了证明（阿隆索等人。，2024；布鲁斯等人。，2024；Hafner 等人。，2021；Kim 等人。，2020；Valevski 等人。，2024），真实世界的机器人（吴等人。，2023；Yang et al.，2023），以及自动驾驶（Blattmann 等人。，2023；胡等人。，2023；Kim 等人。，2021；赵等人。，2024）。我们设想，基础世界模型将对这些行业产生变革性影响。

视频生成模型。 视频生成模型这一领域近年来经历了快速发展。从最初生成短时、低分辨率视频的初始模型，该领域已经取得了显著进步，目前视频生成模型处于生成式人工智能研究的前沿。Ho 等人。，2022；Huang 等人。，2024 近年来，出现了令人瞩目的视频生成模型，如Sora、Dream Machine、Gen 3和Kling，这些模型能够生成逼真、高分辨率的视频（快手，2024；Luma，2024；OpenAI，2024；跑道，2024）。自第一个视频生成模型发布以来，在短短几年内就取得了这些进步。

大多数现有的工作集中在视频生成模型上，这些模型专注于文本到视频的任务，基于文本提示生成视频（Blattmann 等人。，2023；Ge 等人。，2023；Girdhar 等人。，2024；Lin et al.，2024；Ma et al.，2024；Yang et al.，2024）. 这些模型使用户能够使用精心设计的文字提示创建令人印象深刻的视频。其他流行的任务包括从给定图像帧生成视频的 Image-to-Video（）。Blattmann 等人。，2023；Gururani 等人。，2023；Mallya 等人。，2022；Ren et al.，2024；Wang 等人。，2021，2024），在给定参考视频的情况下生成新视频的视频到视频（Ku 等人。，2024；Liu et al.，2024；Mallya 等人。，2020；Wang 等人。，2018，2019）和基于世界模型和具体人工智能发展驱动的动作生成视频的动作对视频（阿隆索等人。，2024；布鲁斯等人。，2024；Tulyakov 等人。，2018；Valevski 等人。，2024）.

大多数视频生成模型采用扩散模型框架（Blattmann 等人。，2023；Ge 等人。，2023；Lin et al.，2024；Ma et al.，2024；Yang et al.，2024）逐步将噪声转化为视频序列。自回归模型也被用于视频生成，提供了以统一方式处理视频及其他模态的优势（邓等人。，2024；Kondratyuk 等人。，2024；Liu et al.，2024）尽管自回归模型显示出了潜力，基于扩散的视频模型在视觉质量方面仍然表现出色。我们的目标是帮助物理人工智能开发者推进其应用。我们认为基于扩散和自回归的模型各自有其优缺点。基于扩散的模型能够生成具有更好视觉质量的视频。而基于自回归的模型则能更好地利用LLM社区开发的各种技术。我们构建了基于扩散（Cosmos-Diffusion）和基于自回归（Cosmos-Autoregressive）的物理场模型，并将它们提供给物理人工智能建设者使用。

视频生成与摄像机控制。 3D一致视频生成可追溯到视图合成和3D重建的早期工作，当时的研究社区致力于使用神经渲染技术对各种3D表示进行操作，以创建3D一致的视频（where the community sought to create 3D-consistent videos using neural rendering applied to various 3D representations）。Kerbl 等人。，2023；李等人。，2023；Mildenhall 等人。，2020；Wang 等人。，2021；周等人。，2018）。在这一行研究中，单图像 3D 视图合成（Charatan 等人。，2024；Lin et al.，2023；塔克和 Snavely，2020；Wiles 等人。，2020；Yu 等人。，2021 在特别具有挑战性的情况下，通常需要从多视图图像数据集学习强大的三维先验模型。由于三维先验模型往往难以很好地扩展，基于学习的方法也通过使用可扩展的Transformer架构的纯数据驱动方法来探索视图综合的学习过程。Dosovitskiy 等人。，2021；Vaswani 等人。，2017）。这绕过了对显式 3D 先验知识的需求（Rombach 等人。，2021；Sajjadi 等人。，2022）: 而不是依赖于应用于3D表示的神经渲染，新的视图是直接通过条件依赖于相机输入的神经网络合成的（Tatarchenko 等人。，2016）。这种范式已经成功地通过扩散模型（Liu et al.，2023），在 3D 资产生成中找到广泛的应用（李等人。，2024；Lin et al.，2023；NVIDIA，2024；普尔等人。，2023；钱等人。，2024；Shi et al.，2023）. 近期在视频生成质量方面的进展表明，通过扩大训练视频数据的规模有可能实现完全的3D一致性（Full 3D一致性）。布鲁克斯等人。，2024）。此后，此类模型的相机可控性已成为

调查（ 他等。 ， 2024 ； Wang 等人。 ， 2024 ； 徐等人。 ， 2024 ）在机器人和自主导航方面的巨大潜在应用。

机器人控制的生成模型。 最近深度生成模型的进展激发了将其应用于机器人控制领域的显著兴趣。多种方法相继出现，其中一项研究方向直接将扩散模型用于视觉运动策略，展示了在各种机器人任务中模仿学习方面取得的重大进步。 Chi et al. ， 2023 ； Ke 等人。 ， 2024 ； 普拉萨德等人。 ， 2024 ； Wang 等人。 ， 2024). 与其他两个更相关的主题是使用预训练的图像和视频生成模型作为运动规划器以及使用图像和视频数据进行生成预训练。生成式运动规划方法 (generative motion planning approach) 布莱克等人。 ， 2023 ； 杜等人。 ， 2024 ； 芬恩和莱文 ， 2017 ； Ko 等人。 ， 2024 ； 周等人。 ， 2024)旨在通过生成中间视觉子目标而非显式的动作序列来增强对未见环境的一般化能力。这种视觉表示策略更为稳健，因为图像和视频子目标可以在多种不同的环境配置中泛化，而传统的动作序列通常针对特定的环境和任务。生成预训练方法（ Cheang 等人. ， 2024 ； Gupta et al. ， 2024 ； 他等。 ， 2024 ）利用大规模图像和视频数据集进行预训练。而 Gupta et al. （ 2024 ）从预先训练的文本到图像扩散模型中提取和利用特征 ， 以指导后续的策略学习 ， Cheang 等人. （ 2024 ）and 他等。 （ 2024 采用两阶段框架：首先对模型进行预训练以预测未来的帧，然后进一步微调以同时预测动作和未来帧。

自动驾驶的生成模型。 视频生成模型有望通过基于多种输入模态（如文本、图像、轨迹、3D数据或地图）生成真实的驾驶视频来革新自主驾驶仿真。 Blattmann 等人。 ， 2023 ； Gao 等人。 ， 2024 ,,; 胡等人。 ， 2023 ； 贾等人。 ， 2023 ； Kim 等人。 ， 2021 ； Lu 等人。 ， 2025 ； Wang 等人。 ， 2023 ， 2024 ； Yang et al. ， 2024 ）。尽管它们具有潜力 ， 但现有的方法受到数据规模限制（ Gao 等人。 ， 2024 ,; 贾等人。 ， 2023 ； Lu 等人。 ， 2025 ； Wang 等人。 ， 2023 ， 2024 ）、决议（ 胡等人。 ， 2023 ； Yang et al. ， 2024 ） ， 以及摄像机视图的数量（ Blattmann 等人。 ， 2023 ； Gao 等人。 ， 2024), 限制了其作为全面驾驶模拟器的有效性。为了克服这些局限性，我们利用预训练的强大WFM的能力，开发出一个灵活且可扩展的驾驶模拟器。我们的模型实现了高分辨率、提升的帧率和多视角一致性。

Tokenizer. 学习再现输入视觉数据的潜在特征已经有相当长的历史（ 他等。 ， 2022 ； Hinton 等人。 ， 1995 ； 金马 ， 2013 ； 范登 · 奥德等人。 ， 2017). 近期，这类模型（也被称为分词器）已被广泛纳入作为提高大规模生成模型训练效率的重要组件之一（。 Esser 等人。 ， 2021 ； Rombach 等人。 ， 2022).

连续视觉分词器，通常包括自动编码器（AE）和变分自动编码器（VAE），将视觉数据压缩到一个连续的潜在空间中，在该空间中基于扩散的模型可以高效地进行训练。 Ho 等人。 ， 2020 ； Lipman 等人。 ， 2022 ； Song et al. ， 2020). 在推理阶段，生成的潜在变量通过分词解码器解码回RGB空间。各种扩散模型就是以这种方式训练用于图像（ ）处理的。 Betker 等人。 ， 2023 ； Dai 等人。 ， 2023 ； FLUX ， 2024 ； Gafni 等人。 ， 2022 ； Podell 等人。 ， 2024 ； Ramesh 等人。 ， 2022 ； Rombach 等人。 ， 2022 ）和视频生成（ 等人。 ， 2023 ； Blattmann 等人。 ， 2023 ,; 布鲁克斯等人。 ， 2024 ； Ge 等人。 ， 2023 ； Girdhar 等人。 ， 2024 ； Wang 等人。 ， 2023 ； Yu 等人。 ， 2023 ； Zeng 等人。 ， 2024).
离散视觉分词器还涉及量化器（ Lee 等人。 ， 2022 ； Mentzer 等人。 ， 2023 ； 范登 · 奥德等人。 ， 2017 ； Yu 等人。 ， 2024 ,; 赵等人。 ， 2024 将连续的潜在变量进一步离散化到离散空间中，从而便于与其他模态（如文本和音频）一起集成到大型语言模型（LLMs）和视觉语言模型（VLMs）中。因此，离散分词器被应用于各种视觉理解场景中（vis-à-vis various visual understanding tasks）。 Sun 等人。 ， 2024 ； Team ， 2024 ； Wang 等人。 ， 2024 ； 吴等人。 ， 2024 ）以及图像（ Chang et al. ， 2022 ； Esser 等人。 ， 2021 ； Ramesh 等人。 ， 2021 ； Sun 等人。 ， 2024 ； Yu 等人。 ， 2022 ）和视频生成任务（ Ge 等人。 ， 2022 ； Hong et al. ， 2023 ； Kondratyuk 等人。 ， 2024 ； Luo et al. ， 2024 ； Villegas 等人。 ， 2023 ； 吴等人。 ， 2022 ； Yan 等人。 ， 2021 ； Yu 等人。 ， 2023).

宇宙标记器是基于先前的研究而广泛构建的 ， *e.g* ., FSQ (Mentzer 等人。 ， 2023) 和因果架构 (Yu 等人。 ， 2023) ， 目标是创建一套高效、高质量的分词器。

9. 结论和讨论

Cosmos世界基金会模型标志着朝着构建通用物理世界模拟器迈出的重要一步。本研究概述了我们全面的方法，包括数据curate流程、连续和离散标记器的设计、扩散和自回归世界基础模型的架构，以及针对多样化下游物理AI任务的微调过程。值得注意的是，我们展示了预训练世界模型对关键应用的适应性，包括3D世界导航、机器人操作和自动驾驶车辆系统，这些系统既需要3D一致性又需要动作可控性。

局限性。 尽管取得了进展，世界基础模型的发展仍处于早期阶段。当前的模型，包括我们的模型，在作为物理世界的可靠模拟器方面仍存在不足。我们观察到，我们的模型仍然存在一些问题，包括物体持久性缺失、接触动力学不准确以及指令执行的一致性差。此外，生成的视频的逼真度并不总是符合基本的物理原理，如重力、光的相互作用和流体动力学。

评估提出了另一个重大挑战。人为定义能够有效评估物理真实性的稳健评价标准困难重重，因为此类评估往往受到个人偏见、背景以及其它主观因素的影响。此外，这些评估可能与下游物理人工智能任务所使用的指标并不完全一致。为了应对这些挑战，值得探索的方向包括利用多模态大型语言模型开发自动评估工具，并利用现有的物理模拟器来实现可重复性和交互式的评估，以此减少对人工评估的依赖。

自回归 *vs* **. 扩散 WFM 。** 我们的评估结果是 3D 一致性 (Sec. 5.3.1) 和机器人技术的视频生成 (Sec. 6 .2) 表明基于扩散的方法当前在生成质量上表现更佳。通过微调，基于扩散的方法能够整合多样化的控制信号，包括相机姿态、末端执行器位置或自主车辆轨迹，并生成如多视角视频等新颖格式的输出。然而，基于自回归的方法仍具有巨大的未开发潜力。它们可以 (1) 利用大型语言模型 (LLMs) 预训练的权重来继承广泛的世界知识；(2) 通过使用专为因果注意力设计的高级推理优化技术加快生成速度。如果这些能力得以充分实现，基于自回归的方法可能特别适合需要交互控制或实时处理的应用场景，例如机器人领域的规划和仿真。重要的是，基于扩散和自回归模型之间的界限并非固定不变。最近的研究表明，双向注意力的扩散变换器可以通过蒸馏成具有因果注意力的学生变换器来支持关键值缓存，在推理过程中提供更好的灵活性和效率 (Yin 等人。 ， 2024 类似地，自回归模型可以通过扩散头整合局部双向注意力来生成图像。 周等人。 ，
2024). 探索这些混合方法及其权衡仍然是一个活跃且有前景的研究领域。我们计划进一步研究这些表达形式，并在未来的工作中提供全面的分析。

A. 贡献者和致谢

A.1. 核心贡献者

- **数据固化**

  Jacob Huffman, Francesco Ferroni, Alice Luo, Niket Agarwal, Hao Wang, Jing Zhang, David Page, Va santh Rao Naik Sabavat, Sriharsha Niverty, Erik Barker, Lindsey Pavao, Stella Shi, Prithvijit Chattopad hyay, Shitao Tang, Yin Cui, Yunhao Ge, Qianli Ma, Yifan Ding, Seungjun Nah, Siddharth Gururani, Jia shu Xu, Grace Lam, Tiffany Cai, Jibin Varghese, Pooya Jannaty, Jay Zhangjie Wu, Yuxuan Zhang, Hu an Ling, Hanzi Mao, Heng Wang

- **令牌器**

  顾金伟、刘贤、葛松伟、王廷春、王浩翔、菲思姆瑞达

- **基于扩散的世界基金会模型预培训**

  Qinsheng Zhang, Lin Yen-Chen, Xiaohui Zeng, Huan Ling, Shitao Tang, Maciej Bala, Ting-Chun Wan g, Yu Zeng, Seungjun Nah, Qianli Ma, Hanzi Mao

- **基于自回归的世界基金会模型预训练**

  王浩翔 ， 丁一凡 ， 刘贤 ， 范娇娇 ， 曾晓辉 ， Yogesh Balaji

- **提示上采样器**

  葛云浩, 王浩翔, 徐家树, 尹翠

- **扩散解码器**

  环玲 ， 范娇娇 ， Fitsum Reda ， Yogesh Balaji ， 毛汉子 ， 张钦生

- **3D 一致性培训前评估**

  黄佳慧 ， 林陈轩

- **物理校准训练前评估**

  Francesco Ferroni, Prthvijit Chattopadhyay, Xinyue Wei, Yuo Ma 千里, Gergely Kl á r, Chen - Hu ssian Lin

- **摄像机控制培训后评估**

  曾晓辉、林宗义、金景义、林陈轩

- **机器人培训后评估**

  林彦蓁, 温正诚, 顾云豪, 刘显, 唐石涛, 魏方尹, 查米, 曾宇, 赵晴晴, 达英, 李昭朔, 顾金维

- **自动驾驶培训后评估**

  金承沃, 吴张杰, 黄嘉惠, 弗朗西斯科·费罗尼, 米歇尔·芬齐, 丹尼尔·杜瓦尔科夫斯基, 德斯潘娜·帕萨利多乌, 艾德·施梅林, 潘怡诗, 劳拉·莱尔-泰克塞,桑扎·菲德勒, 润涵

- **护栏**

  Jibin Varghese, Arslan Ali, Grace Lam, Pooya Jannaty

- **平台架构师**

  刘明宇

A.2. 贡献者

安琪 李, 阿萨尔 安萨维安, 阿图尔 罗兹科夫斯基, 巴托什 斯特凡尼亚克, 迪特 德福, 埃than 何, 凯奇unch 摩, 莫特泽拉 罗梅扎尼, 兹梅克 特德卡克, 韦任 阳, 乔薇薇 仁, 约恩辛 陈, 泽舒安 佩特尔

A.3. Acknowledgments

我们感谢1X Technologies慷慨提供类人机器人数据，并在本技术报告中的机器人操作后处理阶段提供了 invaluable 支持。

我们感谢Aarti Basant、Akan Huang、Alex Qi、Alexis Bjorlin、Amanda Moran、Amol Fasale、Ankit Patel、Arash Vahdat、Aryaman Gupta、Ashna Khetan、Ashwath Aithal、Bor-Yiing Su、Bryan Catanzaro、Charles Hsu、Chris Pruett、Christopher Horvath、Clark Doan、Coulten Holt、Dane Aconfora、Deepak Narayanan、Dennis Chang、Dheeraj Kapur、Dong Ahn、Ebrar Erdem、Elmar Haussmann、Fuzhao Xue、Gandhi Vaithilingam、Henry Estela、Henry Vera、Herb Woodruff、Imad El Hanafi、Jashojit Mukherjee、Jason Sewall、Jensen Huang、John Dickinson、Jonah Alben、Jonah Philion、Josh Abbott、Jun Gao、Kumar Anik、Lee Ditiangkin、Ligeng Zhu、Linxi Fan、Luke Alonso、Madison Huang、Marek Dabek、Mark Arnold、Max Ehrlich、Michele Ferretti、Misbah Mubarak、Misha Smelyanskiy、Mohamed Fawzy、Mohammad Harrim、Mohammad Shoeybi、Omkar Mehta、Pallab Bhattacharya、Paniz Karbasi、Pasha Shamis、Raju Wagwani、Rick Izzo、Robert Hero、Sharon Clay、Song Han、Songyan Tang、Sophia Huang、Sridhar Bhuvanapalli、TJ Galda、Thomas Volk、Tobias Lasser、Vaibhav Ranglani、Vijay Anand Korthikanti、Yao Lu、Yazdan Aghaghiri、Yugi Guvvala、Yuke Zhu和Zekun Hao提供的反馈和支持。

我们感谢Iain Cunningham、Jim Fan、Marco Pavone、Meredith Price、Nikki Pope和Scott Reed对这份技术报告初稿提供的反馈。

参考文献

[1] 阿莫·阿巴斯、库沙尔·蒂鲁马拉、丹尼尔·西米格、苏亚·冈古利和艾里·S·莫科斯. Semdedup：通过语义去重实现大规模网页数据高效学习。 *arXiv 预印本 arXiv ： 2303.09540* , 2023. 10

Bo Adler, Niket Agarwal, Ashwath Aithal, Dong H Anh, Pallab Bhattacharya, Annika Brundyn, Jared
[2] 卡斯珀 ， 布莱恩 · 卡坦扎罗 ， 莎朗 · 克莱 ， 乔纳森 · 科恩等。 Nemotron - 4 340b 技术报告。 *arXiv 预印本 arXiv ： 2406.11704* , 2024. 29

[3] 帕韦什· Agrawal, 萨米恩· Antoniak, 伊玛· Bou Hanna, 布特· Bout, 德文德拉· Chaplot, 杰西卡· Chudnovsky, 迪格奥· Costa, 帕杜安· De Monicault, 苏拉布· Garg, 托菲勒· Gervet, 等. Pixtral 12b。 *arXiv 预印本 arXiv ： 2410.07073* , 2024. 26

[4] 摩德纳大学 AI Image Lab. Bbc 行星地球数据集 ， 2016. URL https: / / aimagelab... unimore. it / imagelab / researchActivity. asp? idActivity = 19 。访问时间 ： 2024 - 10 - 17 。 7

[5] 埃洛伊 · 阿隆索、亚当 · 杰利、文森特 · 米谢利、安西 · 卡纳维斯托、阿莫斯 · 斯托基、蒂姆 · 皮尔斯和弗朗索瓦 Fleuret. Diffusion for world modeling: visual details matter in atari. In *NeurIPS* 2024. 55, 56

贾安, 张宋阳, 杨浩瑞, 哈罗尼, 黄佳斌, 罗洁波, 颜希. 潜在变化: 时间转移的潜在扩散, 用于高效的文字转视频生成. *arXiv 预印本 arXiv ： 2304.08477* , 2023. 57

俞瓦尔·阿特兹蒙，马茨·巴拉，约什·巴拉吉，蒂芙尼·蔡，尹 cui，焦乔乔·fan，云浩·ge，西迪·古鲁拉尼，雅各布·霍夫曼，罗纳德·伊萨克等. Edify图像：基于像素空间拉普拉斯扩散模型的高质量图像生成。 *arXiv 预印本 arXiv ： 2411.07126* , 2024. 22

[Y8] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: 基于专家去噪器集合的文字到图像扩散模型。 *arXiv 预印本 arXiv ： 2211.01324* , 2022. 23

[9] 阿米尔 · 巴尔、周高月、陈丹尼、特雷弗 · 达雷尔、延恩 · 勒村。导航世界模型。 *arXiv 预印本 arXiv ： 2412.03572* , 2024. 55

詹姆斯·贝特克、高加博、李靖、蒂姆·布鲁克斯、王建峰、李林杰、欧阳龙、庄君棠、 Joyce Lee、郭宇飞 等. 通过更好的描述提高图像生成。 *计算机科学. https: / / cdn. openai. com / papers / dall - e - 3. pdf* , 2023. 57

Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. 无预设的机器人操作与预先训练的图像编辑扩散模型。在 *NeurIPS 研讨会* , 2023. 57

稳定视频扩散：将潜在视频扩散模型扩展到大型数据集。 *arXiv 预印本 arXiv ： 2311.15127* , 2023. 56 , 57

[13] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. 使潜在变量对齐：使用潜在扩散模型进行高分辨率视频合成。 *CVPR* , 2023. 11 , 36 , 37 , 46 , 50 , 56 , 57

[14] 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. 视频生成模型作为世界模拟器, 2024. URL https: / / openai. com / research / video - generation - models - as - world - simulators . 56 , 57

汤姆·布朗、本杰明·曼恩、尼古拉斯·苏比亚、杰里德·D·凯普兰、普拉凡拉·达利瓦尔、阿文德·尼拉卡坦、普兰亚·夏米、吉里什·萨斯特里、阿曼达·阿塞尔等（2023年）。语言模型是少样本学习者。在... *NeurIPS* ，2020. 27 ， 29

[16] 杰克·布鲁斯, 迈克尔·D·丹尼斯, 艾希利·爱德华兹, 杰克·帕克-霍尔德, 时宇格, 埃德蒙·休斯, 马修·莱, 阿迪蒂·马瓦兰卡, 鲁迪·施泰格瓦尔德, 克里斯·阿pps, 等. Genie: 生成性交互环境. 在 *ICML* ，2024. 55 ， 56

[17] 天乐蔡, 李宇红, 耿正阳, 彭洪武, 李杰森·D·李, 陈德明, 和 陶_tri. Medusa: 多解码头的简单大语言模型推理加速框架. 在 *ICML* ，2024.
30 ， 31

[18] Brandon Castellano. Pyscenetect, 2024. URL https: / / www. scenedetect. com 视频剪切检测和分析工具。 7

[19] 张慧文、张翰、陆江、刘策、威廉 · T · 弗里曼。 Maskgit: 蒙面生成图像变压器。在 *CVPR* ，2022. 57

[20] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. Pixelsplat: 从图像对生成可扩展且通用的3D重建的3D高斯斑点。In *CVPR* ，2024. 56

张志林, 陈光增, 亚静, 孔涛, 李航, 李一峰, 刘宇晓, 吴洪涛,
贾峰、杨一楚等人，《Gr-2：基于大规模知识的生成视频-语言-行动模型在机器人操作中的应用》。 *arXiv 预印本 arXiv ： 2410.06158* ，2024. 57

[22] 陈天琪、徐冰、张赤元、卡洛斯 · 盖斯特林。用亚线性记忆成本训练深网。 *arXiv 预印本 arXiv ： 1604.06174* ，2016. 23

[23] 陈婷. 论噪声调度对扩散模型的重要性. *arXiv 预印本 arXiv ： 2301.10972* ，2023. 22

[24] 延世大学蔡善仪、白俄罗斯国立大学亚历山大·斯亚罗欣、威利·梅纳帕奇、Ekaterina Deyneka、萧晖纬、李欣颖、詹炳勋、方宇威、任 Jian、杨明轩等. Panda-70m: 使用多种跨模态教师标注 7000 万视频的字幕生成. *CVPR* ，2024. 7 ， 16

[25] 陈池, 邢思远, 杜一伦, 徐振佳, 埃里克·科森内, 贝内迪克特·伯彻菲尔, 和 宋睿然. 扩散策略: 通过动作扩散进行视觉运动政策学习. *RSS* ，2023. 57

戴晓亮 ， 继厚 ， 马志瑶 ， 蔡三伟 ， 王嘉良 ， 王锐 ， 张培钊 ， 西蒙
[26] 梵登亨德，肖芳王，阿比曼尤·杜贝等。Emu: 在草堆中使用迷人针眼增强图像生成模型。 *arXiv 预印本 arXiv: 2309.15807* ，2023. 22 ， 57

你：[27] 阿列克桑德·德·布雷比松和帕斯卡尔·维尼翁. Z-loss：一种属于球面家族的平移和比例不变分类损失函数. *arXiv 预印本 arXiv ： 1604.08859* ，2016. 28

[28] 德赫甘尼莫萨塔法、乔西普·杜霍尔加、巴希尔·穆斯塔法、皮otr·帕德勒夫斯基、杨·希克、贾斯汀·吉尔默、安德烈亚斯·佩特尔·施泰纳、马蒂尔德·卡龙、罗伯特·盖罗霍斯、伊布拉欣·阿卜杜拉辛等. 将视觉变换器扩展至220亿参数. 在 *ICML* ，2023. 21

[Deng H, Pan T, Diao H, Luo Z, Cui Y, Lu H, Shan S, Qi Y, and Wang X. 无需矢量量化的时间序列视频生成.] *arXiv 预印本 arXiv ： 2412.14169* ，2024. 56

[30] 贾登、魏东、理查德 · 索彻、李佳、李凯、李飞飞。 Imagenet: 一个大规模的分层图像数据库。在 *CVPR* ，2009. 12 ， 16

[31] 凤 Jiankang Deng, 郭 Jia Guo, 周 Yuxiang Zhou, 于 Jinke Yu, 费 Irene Kotsia, 和 赞 Stefanos Zafeiriou. Retinaface: 在野外的一阶段密集人脸定位。 *CVPR* , 2020. 55

[32] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: 自监督兴趣点检测与描述。在 *CVPR 研讨会* , 2018. 36

[33]丁泽涵, 张雅敏, 天远东, 郑勤清. 扩散世界模型：离线强化学习中超越逐步展开的未来建模方法。 *arXiv 预印本 arXiv: 2402.03570* , 2024.
55 , 56

[34] 阿列克谢 · 多索维茨基、卢卡斯 · 拜尔、亚历山大 · 科列斯尼科夫、德克 · 魏森伯恩、翟晓华、托马斯 Unterthiner , Mostafa Dehghani , Matthias Minderer , Georg Heigold , Sylvain Gelly 等。图像是 worth 16x16 字 ： 用于按比例进行图像识别的变压器。在 *ICLR* , 2021. 56

[35] 杜一伦，杨诗悦，戴博，戴汉jun，诺姆·纳赫姆，乔希·特尼曼，达尾·舒尔曼，皮特尔·阿贝耳. 通过文本引导的视频生成学习通用策略. 在 *NeurIPS* , 2024. 57

[36] 阿比曼尤·杜贝、阿比纳夫·贾乌里、阿比纳夫·潘迪、阿比谢克·卡德安、阿德姆·阿勒达莱、艾莎·莱特曼、阿克希尔·马图尔、艾伦·舍尔滕、艾米·杨、安吉拉·ファン等. 哈马3模型群。 *arXiv 预印本 arXiv ： 2407.21783* , 2024. 24 , 27 , 29 , 30

弗雷德里克·埃伯特、杨燕来、卡尔·施梅克佩珀、贝拉迪特·布赫、乔治奥斯·乔治阿基斯、科斯塔斯·达尼-伊迪斯、切尔西·芬恩和塞里扬·莱文。桥梁数据：利用跨域数据集提升机器人技能的一般化能力。在 *RSS* , 2022. 43

[38] Patrick Esser 、 Robin Rombach 和 Bjorn Ommer 。用于高分辨率图像合成的驯服变压器。在 *CVPR* , 2021. 57

[39] 张奕斯, 康立勋, 布拉特曼安德烈亚斯, 伊兹阿里拉希姆, 柯纳穆勒约纳斯, 西尼哈里, 李维亚丹弥克, 赖恩多利克洛恩茨, 苏埃尔阿克塞尔, 波施菲尔德弗雷德里克, 等. 扩大规模化的归一化流变换器以实现高分辨率图像合成. 在 *ICML* , 2024. 21

[40] Gunnar Farneb ä ck 。基于多项式展开的两帧运动估计。在 *斯堪的纳维亚图像分析会议* , 2003. 9

[41] Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *ICRA* , 2017. 57

[42] FLUX. FLUX.1 ： 图像生成 ， 2024. URL https: / / huggingface. co / black - forest - labs / FLUX. 1 - dev . 12 , 57

[43] 阮舒悦，雅克尔·亚基尔·塔米尔，苏博希塔·桑德拉姆，柴露，张锐，塔利·德凯尔，以及菲利普·伊索拉。DreamSim：使用合成数据学习人类视觉相似性的新维度。在 *NeurIPS* , 2023. 39

[44] 萨米尔·伊扎克·加德雷，加布里埃尔·伊尔哈罗，亚历克斯·方，乔纳森·海亚塞，乔治奥斯·斯米伦尼斯，陶·阮，瑞安·马特恩，米切尔·沃兹曼，杜布拉·古什，张洁宇，等. 数据综合：寻找下一代多模态数据集. 在 *NeurIPS* , 2024. 10

Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. "Make-a-scene: Scene-based Text-to-Image Generation with Human Priors." In *ECCV* , 2022. 57

[46] 菲利普 · 盖奇。一种新的数据压缩算法。 *C 用户日志* , 1994. 28

[47] 高瑞琪、埃米尔 · 胡格布姆、乔纳森 · 希克、瓦伦丁 · 德 · 博尔托利、凯文 · P · 墨菲和蒂姆 · 萨利曼斯。扩散满足流量匹配 ： 同一枚硬币的两面 ， 2024 。 URL https: / / diffusionflow. github. io / . 19

刘远高, 陈凯, 肖波, ♦Actor Hong, 李正国, 徐强. MagicDriveEdit: 自适应控制下高分辨率长视频生成用于自动驾驶。 *arXiv 预印本 arXiv ： 2411.13807* , 2024. 57

[49] 高瑞元、陈凯、谢恩泽、洪兰青、李振国、杨迪燕、徐强。魔术:
街景生成与不同的 3d 几何控制。在 *ICLR*, 2024. 57

[50] 高慎远, 杨嘉智, 陈黎, 加希普·奇塔, 倪一航, 安德烈亚斯·盖杰, 张jun, 和 李鸿阳. Vista: 具有高保真度和多功能可控性的通用驾驶世界模型. 在 *NeurIPS* , 2024. 57

[51] Leon A Gatys 、 Alexander S Ecker 和 Matthias Bethge 。使用卷积神经网络的图像风格转移。在 *CVPR* , 2016. 15

宋伟、托马斯·海斯、哈里·杨、希寅、关芳、大卫·雅各布斯、贾斌黄和迪维·帕里希。基于无时间感知的vqgan与有时间敏感性的变换器的长视频生成。在 *ECCV* , 2022.
57

[53] Ge宋伟, Nah申骏, Liu桂林, PoonTyler, Tao元道, Catanzaro贝南, Jacobs大卫, Huang黄佳彬, Liu刘明宇, 和 Balaji耶戈什. 保持自己的相关性：视频扩散模型的一种噪声先验. 在 *ICCV* , 2023. 56 , 57

[54] 顾云浩，曾小惠，雅各布·萨缪尔·霍夫曼，林宗毅，刘明宇，崔音. 视觉事实核查器：实现高保真详细图注生成. 在 *CVPR* , 2024. 10

[55] 薛娜·古什、普拉索恩·瓦什尼、埃里克·加林基和克里斯托弗·帕里斯内。Aegis：在线自适应AI内容安全审核，结合多种大语言模型专家意见。 *arXiv 预印本 arXiv ： 2404.05993* , 2024. 53 , 54

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. "Imagebind: One embedding space to bind them all." In [具体会议或期刊名称], [年份]. *CVPR* , 2023. 8
[57] 罗希特·吉德哈、曼纳特·辛格、安德鲁·布朗、昆廷·杜瓦尔、萨曼海·阿扎迪、 Sai Saketh Rambhatla 、阿卡布·沙、Xi Yin、迪维·帕里克和伊shan·米什拉。Emu视频：通过显式的图像条件化进行文本到视频生成的因子分解。在 *ECCV* , 2024. 56 , 57

克里斯滕·格拉曼, 安德鲁·韦斯特伯里, 贝尔纳多·托雷斯阿尼, 克里斯·基塔尼, 吉特纳达·马利克, 特里安菲洛斯·阿弗拉斯, 库玛·阿什图什, 维杰亚·贝雅亚, 涅提汉特·班萨尔, 比克拉姆·布特等。《第一人称与第三人称视角下理解熟练人类活动的Ego-exo4D》. 在 *CVPR* , 2024. 16

[59] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei - Fei, Irfan Essa, Lu Jiang, Jos é Lezama 。使用扩散模型生成逼真的视频。在 *ECCV*, 2024. 21

Gunshi Gupta, Karmesh Yadav, Yarin Gal, Dhruv Batra, Zsolt Kira, Cong Lu, and Tim GJ Rudner. 预训练的文本到图像扩散模型是用于控制的通用表示学习器。在 *ICLR 讲习班* , 2024. 57

[61] Siddharth Gururani, Arun Mallya, Ting-Chun Wang, Rafael Valle, and Ming-Yu Liu. SPACE: 语音驱动的人物动画与可控表情。 *ICCV* , 2023. 56

[62] David Ha 和 J ü rgen Schmidhuber. World models. *arXiv 预印本 arXiv ： 1803.10122*, 2018. 55

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning [63] 潜在想象力的行为。 *arXiv 预印本 arXiv ： 1912.01603*, 2019. 55

[64] Danijar Hafner ， Timothy Lillicrap ， Mohammad Norouzi 和 Jimmy Ba 。用离散掌握 atari world models. in *ICLR*, 2021. 55, 56

[65] Danijar Hafner ， Jurgis Pasukonis ， Jimmy Ba 和 Timothy Lillicrap 。通过世界模型掌握各种领域 。 *arXiv 预印本 arXiv ： 2301.04104* ，2023. 55

[66] Nicklas Hansen, Hao Su, and Xiarony Wang. Td - mpc2: Scalable, robust world models for continuo us control. In *ICLR* , 2024. 55

[67] 理查德 · 哈特利和安德鲁 · 齐瑟曼。 *计算机视觉中的多视图几何* 剑桥大学出版社 ， 2003 年。 36 ， 51

[68] 郝贺, 徐颖浩, 郭宇威, 吴特文·韦茨金, 代波, 李洪生, 和 杨泽远. Cameractrl: 使能相机控制以生成文本到视频。 *arXiv 预印本 arXiv: 2404.02101* ，2024. 57

[69] 何浩然、白晨佳、潘玲、张渭南、赵斌、李雪龙。学习一个可操作的 通过大规模无动作视频预训练的离散扩散策略。在 *NeurIPS*，2024. 57

[70] reinterpret masked autoencoders as scalable vision learners. 在Proceedings of the European Co nference on Computer Vision (ECCV) 2020. *CVPR* ，2022. 57

[71] 马丁·赫塞尔、舒伯特·拉姆萨uer、托马斯·昂特希尔纳、伯恩哈德·内塞尔和塞普·霍希雷特。基于两时间尺度更新规则训练的GANS收敛至局部纳什均衡。在 *NeurIPS* ，2017. 17 ， 41 ， 50

[72] 吉福德·E·亨特、彼得·戴安、布伦丹·J·弗伊和拉德福德·M·尼尔。无监督神经网络的“清醒-睡眠”算法。 *Science* ，1995. 57

[73] 乔纳森 · 何和蒂姆 · 萨利曼斯。无分类器扩散指导。 *arXiv 预印本 arXiv ： 2207.12598* ，2022. 23

[74] Jonathan Ho ， Ajay Jain 和 Pieter Abbeel 。去噪扩散概率模型。在 *NeurIPS* ，2020. 19 ， 57

[75] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fle et. Video diffusion models. In *NeurIPS* , 2022. 56

洪文义、丁明、郑文迪、刘兴汉、唐杰。 Cogvideo ： 大规模预培训
[76] 用于通过变压器生成文本到视频。在 *ICLR* ，2023. 57

[77] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. Simple diffusion: End - to - end diffusion for high resolution images. In *ICML* , 2023. 22

[78] Emiel Hoogeboom, Thomas Mensink, Jonathan Heek, Kay Lamerigts, Ruiqi Gao, and Tim Salimans . Simpler Diffusion (SID2): 在ImageNet512上实现1.5 FID（Frechet Inception Distance）的像素空间扩散方法。 *arXiv 预印本 arXiv ： 2410.19324* ，2024. 19

安东尼·胡、劳德·鲁塞尔、亨德森·叶、扎克·默雷兹、乔治·费多塞夫、亚历克斯·肯德尔、杰米·肖顿和加布里埃尔·科拉多。盖亚-1：一种用于自动驾驶的生成世界模型。 *arXiv 预印本 arXiv ： 2309.17080* ，2023 . 49 ， 55 ， 56 ， 57

[80] 爱德华 · J · 胡、沈业龙、菲利普 · 沃利斯、泽源 · 朱艾伦 · 朱、李元志、王谢安、王鲁东、 陈伟珠。 LoRA ： 大型语言模型的低秩适应。在 *ICLR*，2022. 21

《gensim2：利用多模态和推理大型语言模型规模化生成机器人数据》 *arXiv 预印本 arXiv ： 2410.03645* ，2024. 55

[82] 黄子祺, 何一 nan, 于嘉硕, 张fan, 史晨阳, 江玉明, 张远涵, 吴天行, 金晴阳, 竹纳塔波尔, 等. Vbench: 视频生成模型的综合基准套件. 在 *CVPR* , 2024. 56

在不确定的环境中实现稳定的市场份额增长。

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. Llama guard: Llm-based input-output safeguard for human-ai conversations.

[84] 哈桑·阿德·雅各布斯, 高山太朗, 张成明, 张敏佳, 宋修文·莱昂·索恩, 萨米扬·拉jbhandari, 和 何旭松. Deepspeed ulysses: 用于训练极端长序列变换器模型的系统优化。 *arXiv 预印本 arXiv ：2309.14509* , 2023. 24

Llama守护：基于LLM的人机AI对话输入输出保护。 *arXiv 预印本 arXiv ：2312.06674* , 2023. 54

[85] 付嘉, 威信 Mao, 应飞 Liu, 于成 Zhao, 文清 Wen, 张池 Zhang, 祁宇张 和 汪天材. Adriver-i: 一种自动驾驶通用世界模型. *arXiv 预印本 arXiv ： 2311.13549* , 2023. 57

[86] 江Albert Q.、萨布莱罗Alexandre、曼舍尔Arthur、班福德Chris、查普乐Devendra Singh、德勒斯Casas Diego、布兰Sandand Florian、林耶尔Lengyel Gianna、兰姆Guillaume、萨诺瓦Saulnier Lucile、拉瓦德Lavaud Lélio Renard、拉克胡埃Lachaux Marie-Anne、斯托克Pierre、塞科Le Scao Teven、拉维尔Thibaut、王Thomas、拉克罗伊Timothée和赛德William El。Mistral 7b。 *arXiv 预印本 arXiv ： 2310.06825* , 2023. 27 , 29

[87] 康 Bingyi，岳 Yang，卢 Rui，林 Zhijie，赵 Yang，王 Kaixin，黄 Gao，冯 Jiashi。从物理定律的角度看，视频生成距离世界模型还有多远？ *arXiv 预印本 arXiv ： 2411.02385* , 2024. 37

[88] 贾里德·卡普兰、萨姆·麦坎德利什、汤姆·亨尼汉、汤姆·B·布朗、本杰明·Chess、瑞万·查尔德、斯科特·格雷、亚历克·拉德福德、杰弗里·吴和达里奥·阿莫迪。神经语言模型的扩展定律。 *arXiv 预印本 arXiv ： 2001.08361* , 2020. 25

托罗·卡拉斯,萨穆里·莱因,米卡·艾塔拉,扬内·海尔斯坦,雅克科·莱廷宁,提莫·艾拉。分析与提升stylegan图像质量。在 *CVPR* , 2020. 14

[90] 蒂罗·卡拉斯、米ika·艾塔拉、蒂莫·艾拉和萨姆利·莱因。阐明基于扩散的生成模型的设计空间。在 *NeurIPS* , 2022. 19

[91] 蒂罗·卡拉斯，米卡·艾塔拉，雅克科·莱赫蒂宁，贾内·赫尔森滕，蒂莫·艾拉和萨姆uli·莱恩。分析并改进扩散模型的训练动态。在 *CVPR* , 2024. 19

Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragkiadaki. "3D Diffuser Actor: Policy Diffusion with 3D Scene Representations." In *CoRL* , 2024. 57

[93] 贝尔哈德·克尔布、乔治奥斯·科帕纳斯、托马斯·莱姆基勒和乔治斯·德雷塔基斯. 实时辐射场渲染的3D高斯插值方法。 *ACM 图形事务处理 (TOG)* , 2023. 36 , 56

[94] Rahima Khanam 和 Muhammad Hussain 。 Yolov11 ： 关键架构增强的概述。 *arXiv 预印本 arXiv ： 2410.17725* , 2024. 53

莫锦金, 帕尔·佩尔斯特, 西迪·卡马契蒂, 蒲霄, 阿希win·巴拉克里希纳, 尼尔·萨鲁吉, 拉斐尔·拉菲洛夫, 埃then·福斯特, 格蕾丝·兰姆, 帕纳格·桑凯蒂, 卡文· Vuong, 托马斯·科尔, 本杰明·伯奇菲尔, 刘斯戴克, 德萨·撒迪格, 塞尔盖·列文, 彭西·李昂和查尔莎·芬恩. OpenVLA：一个开源的视觉-语言-行动模型。 *arXiv 预印本 arXiv ： 2406.09246* , 2024. 46

[96] 金承沃、周宇浩、菲利翁·乔纳、托拉拉巴·安东尼奥和费德勒·桑贾. 使用GameGAN模拟动态环境的学习. 在 *CVPR* , 2020. 49 , 55 , 56

[97] 崔承沃、乔丹·菲隆、安东尼奥·托拉尔巴和桑贾·费德勒. DriveGAN：迈向可控的高质量神经模拟. *CVPR* ，2021. 49 ， 55 ， 56 ， 57

[98] 迪德里克 · P · 金马。自动编码变分贝叶斯。 *arXiv 预印本 arXiv ： 1312.6114*，2013. 14, 57

[99] 科泊恩（Po-Chen Ko）、毛嘉远（Jiayuan Mao）、杜一伦（Yilun Du）、孙少华（Shao-Hua Sun）和约书亚·贝·特尼曼（Joshua B Tenenbaum）。通过密集对应关系从无声视频中学习行动。 *ICLR* ，2024. 57

[100] Dan Kondratyuk, Lijun Yu, Shiye Gu, Jos é Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming - Chang Chiu, et al. Videopoet: A large language model for zero - shot 视频生成。在 *ICML*, 2024. 11, 56, 57

[101] Min 罗克, 代佐卓, 周晋, 邢锋江, 李欣, 吴波, 张建卫, 等. 混元视频: 大规模视 ,جان孤佳乐, 田齐, 张子 频生成模型的系统框架. *arXiv 预印本 arXiv ： 2412.03603* ，2024. 19 ， 25

[102] 奥南德·科斯蒂卡尼（Vijay Anand Korthikanti）、贾里德·卡斯珀（Jared Casper）、桑古克·雷姆（Sangkug Lym）、劳伦斯·麦凯夫（Lawrence McAfee）、迈克尔·安德许（Michael Andersch）、穆罕默德·肖伊比（Moham-mad Shoeybi）和布莱恩·卡坦扎罗（Bryan Catanzaro.）：在大型变压器模型中减少激活重计算。

[103] 马克斯·库、丛伟、任伟明、杨欢、陈文虎. Anyv2v ： 用于任何视频到视频编辑任务的免调优框架。 *TMLR* ，2024. 56

[104] 快手。克林 ， 2024 。网址 https: / / klingai. com / . 56

徐裕普, 金智衡, 金在洪, 張敏洙, 韓 Woo-Ki Shin. 使用残差量化进行自回归图像生成. 在 *CVPR* ，2022. 57

[106] 吉米 · 雷巴、杰米 · 瑞安 · 基罗斯和杰弗里 · E · 欣顿。层正常化。 *arXiv 预印本 arXiv ： 1607.06450* ，2016. 14

[107] Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformer via speculative decoding. In *ICML* ，2023. 31

李佳豪、谭浩、张凯、徐泽祥、栾福军、徐英浩、洪一聪、卡利安 · 孙卡瓦利、

[108] Greg Shakhnarovich 和 Sai Bi 。 Instant3d ： 具有稀疏视图生成和大型重建模型的快速文本到 3d 。在 *ICLR* ，2024. 56

[109] 罗钊说，托马斯·穆勒，亚历克斯·埃文斯，刘明宇，林晨轩。Neuralangelo：高保真神经表面重建。在 *CVPR* ，2023. 56

李寒雪, 任嘉薇, 阿什坎·米尔扎伊, 安东尼奥·托拉尔巴, 刘子威, 伊戈尔·吉利切斯金, 桑贾·费德勒, 希南·奥尔提里, 黎欢, 贞戈奇奇等。从单目视频中构建动态场景的前馈子弹时间重建。 *arXiv 预印本 arXiv ： 2412.03526* ，2024. 52

[111] 林斌，葛云阳，程新华，李宗健，朱斌，王少东，何贤毅，叶阳，袁胜海，陈留涵，等. 开放-SORA 计划：开源大型视频生成模型。 *arXiv 预印本 arXiv ： 2412.00131* ，2024. 56

[112] Chen - Husan Lin, Wei - Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle - adjusting neural radiance field. In *ICCV* ，2021. 41

[113] 林晨轩，高俊，唐珑铭，高尾和，曾小惠，黄欣，克尔斯滕·克雷is，桑贾·费德勒，刘明宇，林宗仪。Magic3D：高分辨率文本转3D内容创作。在 *CVPR* ，2023. 56

[114] 季林、尹洪旭、魏平、帕夫洛 · 莫尔恰诺夫、穆罕默德 · 肖伊比、宋寒。维拉: 关于前期训练用于视觉语言模型。在 *CVPR*, 2024. 10

[115] 林启恩、林颜辰、赖伟生、林宗义、石义昌、拉维 · 拉莫莫正西。视觉用于从单个输入图像进行基于 nerf 的视图合成的转换器。在 *WACV*, 2023. 56

[116] 林宗义 ， 迈克尔 · 梅尔 ， 塞尔日 · 贝隆基 ， 詹姆斯 · 海斯 ， 彼得罗 · 佩罗纳 ， 德瓦 · 拉马南 ， 皮奥特 · 多尔尔 and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 12, 16

Philipp Lindenberger, Paul - Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at [117] 光速。在 *ICCV*, 2023. 36

陆玲, 一辰盛, 支图, 赵闻天, 程欣, 万坤, 于蓝桃, 郭千宇, 于子勋, [118] Yawen Lu, et al. Dl3dv - 10k: 一个基于深度学习的三维视觉的大规模场景数据集。在 *CVPR*, 2024. 40, 41

Yaron Lipman ， Ricky TQ Chen ， Heli Ben - Hamu ， Maximilian Nickel 和 Matt Le 。流量匹配 [119] 生成建模。 *arXiv 预印本 arXiv ： 2210.02747* , 2022. 57

[120] 刘畅、李锐、张开东、蓝云伟、刘东。 Stablev2v ： 稳定的形状一致性在视频到视频编辑中。 *arXiv 预印本 arXiv ： 2411.11045*, 2024. 56

[121] Hao Liu, Matei Zaharia, and Pieter Abbeel. Ring attention with blockwise transformer for near - infini te context. *arXiv 预印本 arXiv ： 2310.01889* , 2023. 24

[122] 刘浩、阎维、马泰伊·扎哈利亚和皮特·阿贝尔。用于百万长度视频和语言的世界模型与分块环注意力机制。 *CoRR* , 2024. 55

刘浩哲, 刘世坤, 周子健, 许梦梦, 肖艳萍, 韩潇, 佩雷斯·胡安C, 刘鼎, 卡玛拉·卡塔皮蒂亚, 贾梦琳, 等。Mardi ni: 视频生成大规模场景下的掩码自回归扩散模型。 *arXiv 预印本 arXiv ： 2410.20280* , 2024. 56

刘若希、吴润迪、巴西勒 · 范 · 霍里克、帕维尔 · 托克马科夫、谢尔盖 · 扎哈罗夫和卡尔 · 冯德里克。Zero - 1 - to - 3: Zero - shot one image to 3d object. In *ICCV*, 2023. 56

[125] Ilya Loshchilov 和 Frank Hutter 。解耦权重衰减正则化。在 *ICLR*, 2019. 23, 29

[126] 陆佳辰、黄泽宇、杨泽宇、张佳慧、李章。沃沃根 ： 世界量感知扩散用于可控制的多摄像机驾驶场景生成。在 *ECCV*, 2025. 57

[127] 卢马。梦想机器 ， 2024 。网址 https: / / lumalabs. ai / dream - machine. 56

[128] 罗卓言、石凤元、葛一笑、杨宇九、王丽敏、应山。 open - magvit2: 一个 open - 面向民主化自动回归视觉生成的源项目。 *arXiv 预印本 arXiv: 2409.04410*, 2024. 12, 57

[129] 马新, 王耀辉, 贾更云, 陈新远, 刘紫薇, 李元芳, 陈存健, 于乔. 拿铁 ： 用于视频生成的潜在扩散变压器。 *arXiv 预印本 arXiv ： 2401.03048*, 2024. 56

[130] 阿伦 · 马利亚、王廷春、卡兰 · 萨普拉和刘明玉。世界一致的视频到视频合成。In *ECCV*, 2020. 56

[131] Arun Mallya, Ting - Chun Wang, and Ming - Yu Liu. implicit warping for animation with Image Sets. In *NeurIPS*, 2022. 56

[132] Fabian Mentzer ， David Minnen ， Eirikur Agustsson 和 Michael Tschanen 。有限标量量化 ： Vq - vae 制作简单。 *arXiv 预印本 arXiv ： 2309.15505*, 2023. 14, 28, 57, 58

[133] Vincent Micheli, Eloi Alonso, and Fran ç ois Fleuret. Transformers are sample - effective world models. In *ICLR* , 2023. 55

[134] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: 使用神经辐射场表示场景进行视图合成。 *ECCV* , 2020. 36 , 56

[135] 乔治 · 米勒。 Wordnet ： 英语词汇数据库。 *ACM 的通讯* , 1995. 54

[136] Mistral 和 NVIDIA 。 Mistral - nemo - 12b - 指示 ： 一个 12b 参数大语言模型 ， 2024 。 URL https: / / mistral. ai / news / mistral - nemo / . 18, 25

[137] 菲利普 · 莫里茨、罗伯特 · 西原、斯蒂芬妮 · 王、阿列克谢 · 图马诺夫、理查德 · 肖、埃里克 · 梁、威廉 Paul 、 Michael I. Jordan 和 Ion Stoica. Ray ： 新兴 AI 应用程序的分布式框架。 *CoRR* abs / 1712.05889, 2017. URL http: / / arxiv. org / abs / 1712.05889 . 1, 11

[138] 理查德 · M · 默里 ， 李泽祥和 S · 尚卡 · 萨斯特里。 *机器人操纵的数学介绍* . 儿童权利委员会出版社 ， 2017 年。 35

[139] Sorosh Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar 和 Yuke Zhu 。 Robocasa ： 通用机器人日常任务的大规模模拟。 *arXiv 预印本 arXiv ： 2406.02523* , 2024. 55

[140] NVIDIA. Isaac sim, 2024. URL https: / / developer. nvidia. com / isaac / sim . 37

[141] 英伟达。 Omniverse ， 2024 。 URL https: / / www. nvidia. com / en - us / ominverse . 37

[142] NVIDIA. Physx, 2024. URL https: / / github. com / NVIDIA - Omniverse / PhysX . 37

NVIDIA 。 Edify 3d ： 可扩展的高质量 3d 资产生成。 [143] *arXiv 预印本 arXiv ： 2411.07135* , 2024. 56

[144] 英伟达。变压器引擎 ， 2024 。 URL https: / / github. com / NVIDIA / TransformerEngine. 24

[145] OpenAI. Tiktoken ， 2022. URL https: / / github. com / openai / tiktoken . 28

[146] OpenAI. Dall · e 3, 2024. URL https: / / openai. com / dall - e 。 32 已访问: [在此处插入访问日期] 。

[147] OpenAI. Sora, 2024. URL https: / / openai. com / sora /. 56

潘林飞 ， 达尼尔 · 巴拉斯 ， 马克 · 波勒费斯和约翰内斯 · 舍恩伯格。全球结构 - 运动 [148] revisioned. In *ECCV* , 2025. 40, 41

[149] 亚当 · 帕斯克 ， 萨姆 · 格罗斯 ， 弗朗西斯科 · 马萨 ， 亚当 · 莱勒 ， 詹姆斯 · 布拉德伯里 ， 格雷戈里 · 查南 ， 特雷 Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An 命令式, high - performance 深度学习图书馆 *神经信息处理系统的进展* , 32, 2019. 30

[150] William Peebles 和 Saining Xie 。带变压器的可扩展扩散模型。在 *ICCV* , 2023. 20, 21

[151] Bowen Peng and Jeffrey Quesnelle. Ntk - aware scaled rope allows lama models to have extended (8k +) 上下文大小 ， 没有任何微调和最小的困惑退化 ， 2023 年。 21

[152] 匡文鹏, 奎尔莱·杰弗里, 范鸿禄, 和 舒普利·恩里科. Yarn: 大型语言模型高效上下文窗口扩展方法。 *arXiv 预印本 arXiv ： 2309.00071* , 2023. 27

费德里科·佩拉齐,约迪·庞特-图塞特,布莱恩·麦克威廉斯,卢克·范高尔,马库斯·格罗斯,和亚历山大·索尔金-霍恩宁。一段视频对象分割的基准数据集及评估方法。在 *CVPR* , 2016. 12

[154] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penn a, and Robin Rombach. SDXL：提高潜在扩散模型以实现高分辨率图像合成。 *ICLR* ，2024. 57

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, [155] 史博文 ， 马志耀 ， 庄庆耀 ， 等。电影创 ： 媒体基础模型的演员阵容。 *arXiv 预印本 arXiv ： 241 0.13720* ，2024. 7 ， 19 ， 25

[156] 本 · 普尔、阿贾伊 · 贾恩、乔纳森 · T · 巴伦和本 · 米尔登霍尔。梦想融合 ： 使用 2d 的文本到 3d 扩散。在 *ICLR*，2023. 56

Aaditya Prasad ， Kevin Lin ， Jimmy Wu ， Linqi Zhou 和 Jeannette Bohg 。一致性政策 ： 加速 [157] 通过一致性蒸馏实现视觉运动策略。 *arXiv 预印本 arXiv ： 2405.07503* ，2024. 57

国成钱, 金洁梅, 阿卜杜拉·哈姆迪, 建任, 阿列克桑德·西罗欣, 彭力, 荀英莉, 艾凡·斯科罗霍多夫, 彼得·翁卡, 谢尔盖·图柳亚科夫, 等。Magic123: 使用二维和三维扩散先验从一张图片生成高质量3D对象。在 *ICLR* ，2 024. 56

[159] 科林·拉夫尔，诺姆·沙伊尔，亚当·罗伯茨，凯瑟琳·李，沙伦·纳朗，迈克尔·马特纳，周yanqi，李wei 和刘彼得·J·利乌。探索统一的文本到文本变换器在迁移学习极限方面的应用。 *JMLR* ，2020. 21 ， 23

[160] Prajit Ramachandran ， Barret Zoph 和 Quoc V Le 。搜索激活功能。 *arXiv 预印本 arXiv ： 1710.0 5941* ，2017. 14

[161] 阿迪蒂亚 · 拉梅什、米哈伊尔 · 巴甫洛夫、加布里埃尔 · 戈、斯科特 · 格雷、切尔西 · 沃斯、亚历克 · 拉德福德、马克 Ilya Sutskever. Zero - shot text - to - image generation. In *ICML*，2021. 57

[162] 阿迪蒂亚·拉梅什、普拉夫拉达·达里瓦尔、亚历克斯·尼科尔、凯西·朱和马克·陈. 基于CLIP隐变量的分层文本条件图像生成。 *arXiv 预印本 arXiv: 2204.06125* ，2022. 32 ， 57

[163] RAPIDS 。急流 ： 端到端 gpu 数据科学的图书馆 ， 2023 年。 URL https: / / rapids. ai . 10

魏明 任，杨卦，张哲，魏Cong，杜欣润，黄文浩，和 陈文华. Consisti2v：增强图像到视频生成中的视觉一致性。 *arXiv 预印本 arXiv: 2402.04324* ，2024. 56

[165] 扬 · 罗宾、马克 · 霍夫曼、托比亚斯 · 乌尔韦和斯特凡 · 哈默林。基于变形金刚的世界模型对 100k 互动感到满意。 *arXiv 预印本 arXiv ： 2303.07109*，2023. 55, 56

[166] Robin Rombach, Patrick Esser, and Bj ö rn Ommer. Geometry - free view synamic: Transformers a nd no 3d priors. In *ICCV* ，2021. 56

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bj ö rn Ommer. High - resolutio n
[167] 具有潜在扩散模型的图像合成。在 *CVPR* ，2022. 11 ， 19 ， 57
[168] 跑道。第 3 代 ， 2024 年。 URL https: / / runwayml. com / research / introducing - gen - 3 - alpha. 56

[169] 赛义德莫特泽·萨达特,雅各布·布赫曼,戴克·布拉德利,奥特马·希利格斯,和罗曼·M·韦伯. Lite-vae: 轻量级且高效的变分自编码器用于潜在扩散模型。 *arXiv 预印本 arXiv: 2405.14477* ，2024. 14

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemi pour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, 等人。具有深度语言理解的逼真文本到图像扩散模型。在 *NeurIPS* ，2022. 23

梅赫迪·S·萨贾迪、亨宁·迈耶、埃蒂安·波特、乌尔斯·贝尔格曼、克劳斯·格雷夫、诺哈·拉德万、苏汉尼·沃拉、马里奥·卢奇、丹尼尔·邓克沃斯、亚历克西·多索维斯基等人。场景表示变换器：通过集合隐式场景表示实现几何无关的新视角合成。在 *CVPR* ，2022. 56

[172] 保罗 · D · 桑普森。将圆锥截面拟合到 “非常分散 ” 的数据 ： 对 bookstein 算法的迭代细化。 *计算机图形学和图像处理* ，1982. 36 ， 51

[173] 约翰内斯 · 舍恩伯格、郑恩良、扬 - 迈克尔 · 弗拉姆和马克 · 波勒费斯。像素视图非结构化多视图立体的选择。在 *ECCV*，2016. 36, 41

[174] Johannes LutzSch ö nberger and Jan - Michael Frahm. Structure - from - motion revisioned. In *CVPR*，2016. 36, 41

[175] Christoph Schuhmann 。改进的美学预测器 ， 2022 。 URL https: / / github. com / christophschuhmann / 改进的美学预测 . 9

[176] 石伊春, 王鹏, 叶江龙, 麦龙, 李克杰, 肖扬. Mvdream: 三维生成的多视图扩散. *arXiv 预印本 arXiv: 2308.16512* ，2023. 56

[177] 摩哈默德·绍伊比,莫斯塔法·帕图瓦里,拉尔·普里,帕特里克·莱格雷斯利,杰拉德·卡斯per,和布莱恩·卡坦扎罗. 兆atron-lm: 使用模型并行性训练多十亿参数的语言模型。 *arXiv 预印本 arXiv ： 1909.08053* ，2019 . 29

[178] Karen Simonyan 和 Andrew Zisserman 。用于大规模图像识别的非常深度卷积网络。 *arXiv 预印本 arXiv ： 1409.1556* ，2014. 15

[179] 维尼森特·西兹曼、塞蒙·雷兹奇科夫、比尔·弗里曼、约什·泰恩布尔和弗德罗·杜兰。光场网络：具有单评估渲染的神经场景表示。 *NeurIPS* ，2021. 41

[180] 杨松, 詹斯查·索赫-迪克斯坦, 迪德里克·P·Kingma, 比什克·库马尔, 萨藤·埃蒙, 以及本·波尔。基于分数的生成建模通过随机微分方程。 *arXiv 预印本 arXiv: 2011.13456* ，2020. 19 ， 57

[181] Tom á s Soucek 和 Jakub Lokoc 。 Transnet v2 ： 用于快速拍摄过渡检测的有效深度网络架构。在 *ACM MM* ，2024. 7

[182] 苏 Jianlin、艾哈迈德 Murtadha、卢 Yu、潘 Shengfeng、鲍 Wen、刘 Yunfeng。Roformer：带有旋转位置嵌入的增强型变压器。 *神经计算* ，2024. 20

孙培泽, 江一, 陈守发, 张世龙, 彭冰悦, 胥萍, 袁泽欢. 自回归模型优于扩散：Llama 在可扩展图像生成中的应用. *arXiv 预印本 arXiv ： 2406.06525* ，2024. 12 ， 57

[184] quan Sun, yufeng Cui, xiaosong Zhang, fan Zhang, qiying Yu, yuze Wang, yongming Rao, jingjing Liu, tiejun Huang, and Xinlong Wang. 生成多模态模型是情境学习者。在 *CVPR* ，2024. 57

[185] 马修·坦西克, 艾than·韦伯, 奈温·恩, 罗朗·李, 布伦特·伊, 特兰斯·王, 亚历山大· kristoffersen, 杰克·奥斯汀, 卡米亚尔·萨拉希, 阿比克·阿胡ja, 等. Nerfstudio: 用于神经辐射场开发的模块化框架. 在 *ACM 图表* ，2023. 36

[186] 塔特·石涛, ◆冯立桐, 裘张辉, 陈益民, 和 张巍. 基于深度结构模型的快速视频镜头转换定位. 在 *ACCV* ，2018. 7

[187] Maxim Tatarchenko, Alexey Dosovitskiy, 和 Thomas Brox. 使用卷积网络从单张图像生成多视图 3D 模型。在 *ECCV* ，2016. 56

[188] 变色龙团队。变色龙 ： 混合模态早期融合基础模型。 *URL https: / / arxiv. org / abs / 2405.09818* ， 2024. 57

[189] Gemma 团队。 Gemma 2 ： 改进实用规模的开放语言模型 ， 2024 。 URL https: / / arxiv. org / abs / 2408.00118 . 29

[190] 1X 技术。 1xgpt ， 2024 。 URL https: / / github. com / 1x - technologies / 1xgpt . 43

[191] Zachary Teed 和 Jia Deng. Raft: 光流的递归全对场变换。在 *ECCV*，2020. 15

[192] Zachary Teed 和 Jia Deng 。 Droid - slam ： 单目 ， 立体声和 rgb - d 相机的 Deep visual slam 。在 *NeurIPS*，2021. 52

[193] 姚腾、韩时、刘贤、宁雪飞、戴国浩、王宇、李振国、刘希辉。加速自动回归文本到图像生成与无训练的推测 jacobi 解码。 *arXiv 预印本 arXiv: 2410.01699*，2024. 31

[194] Richard Tucker 和 Noah Snavely 。具有多平面图像的单视图合成。在 *CVPR*，2020. 56

谢尔盖 · 图利亚科夫、刘明玉、杨晓东和扬 · 考茨。 MoCoGAN ： 分解运动和 [195] 用于视频生成的内容。在 *CVPR*，2018. 56

托马斯·翁特拉赫、绍尔德·范·斯坦基斯特、卡罗尔·库拉赫、拉斐尔·马里纳、马辛·米哈尔斯和西尔万·杰利 。Fvd ：视频生成的新度量指标。在 *ICLR 讲习班* ，2019. 17 ， 41 ， 50

[197] Dani Valevski 、 Yaniv Leviathan 、 Moab Arar 和 Shlomi Fruchter 。扩散模型是实时游戏引擎。 *arXiv 预印本 arXiv: 2408.14837* ，2024. 55 ， 56

[198] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NeurIPS* ，2017. 14 ， 57

[199] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łuka sz Kaiser, and Illia Polosukhin.《注意：你所需的一切》.收录于 *NeurIPS* ，2017. 27 ， 56

[200] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, M ohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: 从开放领域文本描述生成变长视频。 *ICLR* ，2023. 57

[201] 荷马 · 沃尔克 ， 凯文 · 布莱克 ， 亚伯拉罕 · 李 ， 莫金 ， 马克斯 · 杜 ， 郑崇义 ， 托尼 · 赵 ， 菲利普 Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levin e. Bridgedata v2: 一个用于大规模机器人学习的数据集. In *CoRL* ，2023. 16

[202] 王俊科、江义、袁泽欢、彭斌月、吴祖轩、江玉刚。 Omnotokenizer: A用于视觉生成的联合图像 - 视频标记器。 *arXiv 预印本 arXiv ： 2406.09399*，2024. 12

[203] 王鹏、刘灵杰、刘元、克里斯蒂安 · 西奥巴特、高村拓、王文平。 Neus: 学习神经隐式曲面通过体渲染进行多视图重建。在 *NeurIPS*，2021. 56

[204] 彭王, 白帅, 谭_sinan, 王世杰, 范志浩, 贝晋泽, 陈Ke Qin, 刘雪jing, 王嘉林, 何文斌, 等. Qwen2-vl: 提升视觉-语言模型在任意分辨率下对世界的感知能力。 *arXiv 预印本 arXiv ： 2409.12191* ，2024. 10

[205] 王廷春、刘明玉、朱俊艳、刘桂林、陶安德鲁、扬 · 考茨和布莱恩 · 卡坦扎罗。视频到视频合成。在 *NeurIPS*，2018. 56

[206] 王廷春、刘明玉、陶安德鲁、刘桂林、扬 · 考茨和布莱恩 · 卡坦扎罗。很少拍摄视频到视频合成。在 *NeurIPS* ，2019. 56

[207] 王廷春、阿伦 · 玛尔雅、刘明玉。视频会议的一枪自由视神经说话头合成。在 *CVPR* ，2021. 56

[208] 谢 Wang, 张世伟, 高 Changxin, 汪嘉宇, 周小强, 张 Yingya, 严 Luxin, 和 桑农. Unianimate: 控制统一视频扩散模型以实现一致的人像动画。
arXiv 预印本 arXiv ： 2406.01188 ，2024. 56

王小峰, 朱郑, 黄冠, 陈欣泽, 朱嘉纲, 陆际文. Drivedreamer: 向真实世界驱动的世界模型迈进以实现自动驾驶. *arXiv 预印本 arXiv ： 2309.09777* ，2023. 57

[210] 王新龙，张小松，罗正雄，孙全，崔宇峰，王金生，张fan，王悦泽，李震，于启莹，等. Emu3：只需下一词预测。 *arXiv 预印本 arXiv ： 2409.18869* ，2024. 57

[211] 王亚辉，陈新远，马欣，周尚辰，黄子棋，王懿，杨采圆，何燕南，于嘉硕，杨培清，等. Lavie：级联潜扩散模型生成高质量视频。 *arXiv 预印本 arXiv: 2309.15103* ，2023. 57

王毅, 李坤昌, 李新浩, 余嘉朔, 何衍南, 陈郭, 裴宝琪, 郑荣坤, 王遵, 沈衍松, 等. Internvideo2: 多模态视频理解的基础模型扩展. 在 *ECCV* ，2025. 9

[213] 王宇奇，何嘉伟，樊蕾，李鸿欣，陈云涛，张钊祥. 驱向未来：基于世界模型的多视图视觉预测与规划在自动驾驶中的应用. 在 *CVPR* ，2024. 57

王振东 ， 李兆硕 ， 阿杰 · 曼德勒卡 ， 徐振佳 ， 范娇娇 ， 亚什拉杰 · 纳朗 ， 范林西 ，
你：[214] 周玉iker, 杰吉·巴拉吉, 周明远, 等. 一步扩散策略：通过扩散精炼实现快速视知觉运动策略。 *arXiv 预印本 arXiv ： 2410.21257* ，2024. 57

周 Xia, 王 Ziyang, 王 Xintao, 李 Yaowei, 陈 Tianshui, 夏 Menghan, 洛 Ping, 和 山 Ying. Motionctrl: 视频生成统一且灵活的运动控制器. 在... *ACM 图表* ，2024. 57

[216] 翁启珍、杨凌云、余英浩、王伟、唐小川、杨国栋、张丽萍。
小心碎片 ： 调度 {GPU 共享} 具有碎片梯度下降的工作负载。
In *USENIX ATC*，2023. 11

[217] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: 端到端观点合成
来自单个图像。在 *CVPR*，2020. 56

[218] 米切尔·沃兹曼, 彼得·J·刘, 莱乔·肖, 凯蒂·厄文, 亚历克斯·阿莱米, 本·阿德兰, 约翰·D·科-雷耶斯, 伊泽丁·古尔, 阿比舍克·库马尔, 罗曼·诺瓦克, 等. 小规模代理用于大规模变压器训练不稳定性。 *arXiv 预印本 arXiv ： 2309.14322* ，2023. 21 ， 27 ， 28

[219] 吴晨飞 ， 梁健 ， 季雷 ， 范洋 ， 方岳建 ， 江大新 ， 南段。女娃 ： 视觉
神经视觉世界创造的综合预训练。在 *ECCV*，2022. 57

[220] 吴浩宁、张二立、廖亮、陈朝峰、侯静文、王安南、孙文秀、琼
严 ， 林维思。从美学和
技术观点。在 *ICCV*，2023. 9

菲利普·吴,亚历杭德罗·埃斯孔特雷拉,达尼贾尔·哈弗纳,皮埃尔·阿贝尔和肯·戈尔德伯格。《白日梦者：物理机器人学习的世界模型》一文。 *CoRL* ，2023. 56

[222] 姚晨, 张卓阳, 陈jun宇, 唐浩天, 李大成, 方云浩, 朱令恒, 谢enze, 于鸿旭, 伊力, 等. Vila-u: 统一的基础模型整合视觉理解与生成。 *arXiv 预印本 arXiv ： 2409.04429* ，2024. 57

[223] 吴玉新、何开明. 集团正常化. 在 *ECCV*，2018. 14

[224] 徐德佳、聂伟丽、刘超、刘思飞、扬 · 考茨、王张扬和阿拉什 · 瓦达。卡姆科：相机可控的 3d 一致性图像到视频生成。 *arXiv 预印本 arXiv ： 2406.02509*，2024. 41, 42, 43, 57

[225] 徐晶晶, 徐孙, 张志远, 赵广祥, 林俊阳. 认识和提高层归一化。在 *NeurIPS*，2019. 21

[226] 芙照雪，余康陈，达成李，青蒿胡，立耕朱，秀宇李，云皓方， Hao Tian 汤，尚阳，刘治jan 等. Longvila：面向长视频的长上下文视觉语言模型的扩展。 *arXiv 预印本 arXiv ： 2408.10188* ，2024. 10

[227] 严威尔逊 ， 张云志 ， 彼得 · 阿比尔和阿拉文 · 斯里尼瓦斯。 Videogpt ： 使用 vq - vae 和变压器的视频生成。 *arXiv 预印本 arXiv ： 2104.10157* ，2021. 12 ， 57

[228] 安阳, 杨宝松, 张北辰, 惠彬元, 郑波, 于 Bowen, 李成远, 刘 Dayiheng, 黄飞, 魏浩然, 等. Qwen2.5 技术报告。 *arXiv 预印本 arXiv ： 2412.15115* ，2024. 29

[229] 杨 Cheng-Yen, 黄 Hsiang-Wei, ﷺ Wenhao, 姜 Zhongyu, 和 黄 Jenq-Neng. Samurai: 使分割一切模型适应具有运动意识的记忆的零样本视觉跟踪。 *arXiv 预印本 arXiv ： 2411.11922* ，2024. 39

杨嘉智, 高申源, 秦义航, 陈李, 李天宇, 戴博, 谢卡扬, 吴鹏浩, 曾嘉, 洛平. 自动驾驶的一般预测模型. 在 *CVPR* ，2024. 57

[231] 梅job 阳，杜一伦，卡米亚尔·加希米波尔，乔纳森·汤普森，戴尔·舒尔曼，以及皮特·阿贝尔。学习交互式的现实世界模拟器。 *arXiv 预印本 arXiv ： 2310.06114* ，2023. 55 ， 56

[232] 杨灼毅, 蓝嘉宴, 郑文迪, 丁明, 黄世宇, 徐家正, 杨远明, 洪文逸, 张晓涵, 馮冠宇, 等. CogVideoX: 基于专家变换器的文本到视频扩散模型。 *arXiv 预印本 arXiv: 2408.06072* ，2024. 12 ， 56

[233] 天维 Yin, 张强, 张瑞, 威廉姆·T·弗里曼, 费多·杜兰, 西蒙·谢克特曼, 和 黄欣. 从缓慢的双向生成到快速的因果视频生成器。 *arXiv 预印本 arXiv ： 2412.07772* ，2024. 58

[234] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance field from one or few images. In *CVPR* ，2021. 56

[235] 花玉、陈浩峰、王欣、鲜文启、陈颖颖、刘芳晨、瓦希什·马达万、特雷弗·达尔利。BDD100K：用于异质多任务学习的多样化驾驶数据集。 *CVPR* ，2020. 16

[236] 于嘉惠，徐远中，余 Jing Yu，卢强，Ganjan Baid，王梓瑞，Vijay Vasudevan，Alexander Ku，杨音飞，Ayan Burcu Karagol等. 扩大规模自回归模型以生成内容丰富的文本到图像转换。 *TMLR* ，2022. 57

[237] 玉立jun, 成永, 孙希虎, 约瑟夫·莱扎马, 张寒, 常会文, 霍尔根·G·豪普特曼, 杨明轩, 郝远, 伊夫恩·埃萨, 和 江路. MAGVIT: 遮罩生成视频变换器. 在 *CVPR* , 2023. 57 , 58

李军、约瑟夫·莱萨马、尼特什·巴哈德瓦杰·甘达瓦普、卢卡·韦斯拉里、基希亚克·索恩、大卫·米尼恩、永成、阿格姆·古普塔、西岳·顾、亚历山大·G·豪普特曼、博庆·贡、明-勋宣杨、伊尔凡·埃萨、大卫·A·罗斯和陆江。语言模型超越扩散——分词器是视觉生成的关键。在
ICLR , 2024. 57

[239] 岑希航, 马克·韦伯, 董雪qing, 沈晓辉, 丹尼尔·克雷默斯, 和 陈亮池. 重建和生成一幅图像需要32个标记。 *arXiv 预印本 arXiv ： 2406.07550* , 2024. 57

[240] Yu Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilative diffusion models in pr ojected tlating space. In *CVPR* , 2023. 57

杨增, 魏国强, 郑嘉仪, 邹佳欣, 肖洋, 张越尘, 李航. 让像素起舞：高动态视频生成. 在 *CVPR* , 2024. 57

[242] 翟晓华、巴兹尔 · 穆斯塔法、亚历山大 · 科列斯尼科夫和卢卡斯 · 拜尔。语言图像预训练的 Sigmoid 损失。在 *ICCV* , 2023. 55

[243] 张彪和 Rico Sennrich. 均方根层归一化。在 *NeurIPS*, 2019. 21

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The reasonable
[244] 深度特征作为感知度量的有效性。在 *CVPR* , 2018. 36

魏甫张, 王刚, 孙健, 余添远, 和 黄高. Storm: 高效基于随机变换器的世界模型在强化学习中的应用. 在 *Neur IPS* , 2024. 56

[246] 赵国生，王小锋，朱正，陈鑫泽，黄冠，鲍晓仪，王星gang. Drivedreamer-2：增强型LLM驱动的世界模型在多样化驾驶视频生成中的应用。 *arXiv 预印本 arXiv: 2403.06845* , 2024. 56

[247] Yue Zhao, Yuuan Jun Xiang, and Philipp Kr ä henb ü hl. Image and video tokenization with binary s pherical quanization. *arXiv 预印本 arXiv ： 2406.07548* , 2024. 57

[248] 周晨亭，于丽丽，阿鲁班·巴布，库沙尔·提鲁马拉，三村幸生，列昂尼德·夏米斯，雅各布·汗，马雪哲，卢克·泽特莱莫耶和奥默·雷维。Transfusion：使用一个跨模态模型预测下一个标记并扩散图像。 *arXiv 预印本 arXiv: 2408.11039* , 2024. 58

[249] 周思远，杜一伦，陈佳本，李岩东，杨迪- Yan，甘疮。Robodreamer：学习组合的世界模型以实现机器人想象。在 *ICML* , 2024. 57

[250] 周庭辉，理查德·塔克，约翰·弗莱恩，格雷汉姆·菲夫，以及诺亚·斯内维。立体放大：使用多平面图像学习视图合成。 *ACM 图形事务处理 (TOG)* , 2018. 36 , 41 , 56

[251] 竿齐 竺, 吴洪涛, 郭松, 刘宇潇, 陈立兰, 孔涛. Irasim: 学习交互式的实物机器人动作模拟器。 *arXiv 预印本 arXiv ： 2406.14540* , 2024. 46 , 48

[252] 翟 Wentao，黄玉芳，谢秀峰，刘文贤，邓锦灿，张德兵，王章阳，刘骥. 自动截帧：短视频数据集及最先进的镜头边界检测. *CVPR 研讨会* , 2023. 7