

广义自动驾驶预测模型

贾志阳^{1*}高深远^{2,1*}邱一航^{1*}陈力^{3,1†}李天宇¹戴博¹
 Kashyap Chitta^{4,5}Penghao Wu¹Jia Zeng¹Ping Luo³Jun Zhang^{2‡}
 Andreas Geiger^{4,5}Yu Qiao¹Hongyang Li^{1†}

1 OpenDriveLab和上海人工智能实验室2 香港科技大学3 香港大学4 蒂宾根大学5 蒂宾根人工智能中心

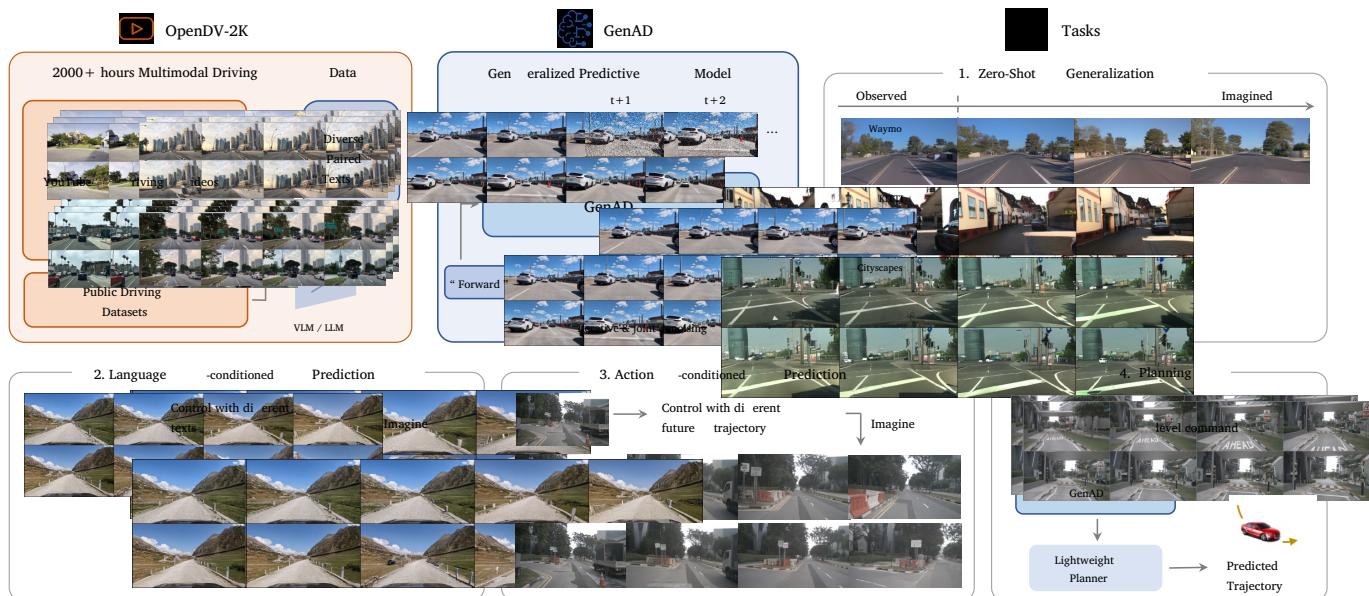


图1. GenAD范式的概述。我们的目标是通过呈现迄今为止最大的多模态驾驶视频数据集OpenDV-2K，以及一种能够根据过去视觉和文本输入预测未来的生成模型GenAD，来建立一个适用于自动驾驶的通用视频预测范式。GenAD的强大泛化能力和可控性在多种任务中得到了验证，包括零样本领域迁移、语言条件预测、动作条件预测和运动规划。

摘要

在本文中，我们介绍了自动驾驶领域首个大规模视频预测模型。为了消除高成本数据收集的限制并增强我们模型的泛化能力，我们从网络上获取了大量数据，并将其与多样化和高质量的文本描述配对。由此产生的数据集累计了超过2000小时的驾驶视频，涵盖全球各地，具有多样的天气条件和交通场景。继承了近期

我们的模型，名为GenAD，通过新颖的时间推理模块处理驾驶场景中的复杂动态。我们展示了它能够以零样本方式推广到各种未见过的驾驶数据集，超越了通用或特定于驾驶的视频预测模型。此外，GenAD可以被调整为一个动作条件预测模型或一个运动规划器，具有巨大的实际驾驶应用潜力。

1. 引言

自动驾驶代理，作为高级人工智能的潜力应用，感知周围环境

* 等贡献，按抛硬币顺序排列。‡共同指导。†项目负责人。主要联系人：
 yangjiazhi@opendrivelab.com

环境、构建内部世界模型表示、做出决策，并采取行动应对[9, 50]。然而，尽管学术界和工业界多年来投入了大量努力，这些系统的部署仍局限于某些地区或场景，无法在全球范围内无缝应用。一个关键原因是结构化自动驾驶系统中学习模型的泛化能力有限。通常，感知模型在面对地理区域变化、传感器配置、天气条件、开放集对象等多样环境时，面临泛化挑战；预测和规划模型则无法泛化到具有罕见场景和不同驾驶意图的非确定性未来[2, 16, 54]。

受人类如何感知和认知世界[27, 28, 49]的启发，我们主张采用驾驶视频作为通用接口，以适应具有动态未来的多样化环境。基于此，我们更倾向于使用驾驶视频预测模型，以全面捕捉关于驾驶场景的世界知识（图1）。通过预测未来，视频预测器本质上学习了自动驾驶的两个关键方面：世界如何运作，以及如何在野外安全操控。

最近，社区开始采用视频作为表示各种机器人任务的观察行为和动作的接口[11]。对于经典视频预测和机器人等领域，视频背景大多是静态的，机器人移动缓慢，视频分辨率较低。相比之下，对于驾驶场景，它需要应对高度动态的户外环境，涵盖更大运动范围的代理以及覆盖广泛视野的感官分辨率。这些差异给自动驾驶应用带来了重大挑战。幸运的是，在驾驶领域开发视频预测模型方面已经有一些初步尝试[4, 15, 19, 23, 25, 33, 38, 45, 47]。尽管在预测质量方面取得了令人鼓舞的进展，但这些尝试尚未达到在经典机器人任务（如操作）中所需的泛化能力，仅限于有限场景，如交通密度较低的高速公路[4]和小规模数据集[15, 23, 33, 45, 47]，或受限于生成多样化环境的困难条件[38]。如何发掘视频预测模型在驾驶中的潜力仍鲜有探索。

受上述讨论的启发，我们的目标是构建一个适用于自动驾驶的视频预测模型，该模型能够泛化到新的条件和环境中。为此，我们需要回答以下问题：（1）哪些数据可以以可行且可扩展的方式获取？（2）如何构建一个预测模型来捕捉动态场景的复杂演变？（3）如何将（基础）模型应用于下游任务？

缩放数据。为了实现强大的泛化能力，

大量且多样化的数据语料库是必要的。受到基础模型从互联网规模数据中学习成功的启发[1, 26, 39]，我们从网络和公开授权的数据集中构建了我们的驾驶数据集。与现有选项相比，由于其受限的收集过程，这些选项在规模和多样性上有限，而网络数据在多个方面具有极大的多样性：地理位置、地形、天气条件、安全关键场景、传感器设置、交通元素等。为了确保数据的高质量和适合大规模训练，我们全面收集了YouTube上的驾驶记录，并通过严格的人工验证移除了意外的损坏帧。此外，视频与多样化的文本级条件配对，包括由现有基础模型[30, 35]生成和精炼的描述，以及由视频分类器推断的高级指令。通过这些步骤，我们构建了OpenDV-2K，这是迄今为止最大的公开驾驶数据集，包含超过2000小时的驾驶视频，比广泛使用的nuScenes数据集大了374倍。我们的数据集公开可用，网址为<https://github.com/OpenDriveLab/DriveAGI>。

广义预测模型。学习一个广义的驾驶视频预测器面临几个关键挑战：生成质量、训练效率、因果推理以及剧烈视角变化。我们通过提出一种新颖的两阶段学习的时间生成模型来解决这些方面的问题。为了同时捕捉环境细节、增强生成质量并保持训练效率，我们基于最近在潜在扩散模型（LDMs）[37, 41]方面的成功。在第一阶段，我们通过在OpenDV-2K图像上微调LDM，将其生成分布从预训练的通用视觉领域转移到驾驶领域。在第二阶段，我们将提出的时间推理模块插入原始模型中，并学习根据过去帧和条件来预测未来。与传统的时间模块[4, 18]相比，这些模块容易受到因果混淆和大运动的影响，我们的解决方案包括因果时间注意力和解耦的空间注意力，以高效地建模高度动态驾驶场景中的剧烈时空变化。经过充分的训练后，我们的自动驾驶生成模型（GenAD）¹能够以零样本的方式推广到各种场景中。

模拟与规划扩展。在大规模视频预测预训练之后，GenAD本质上理解了世界如何演变以及如何驾驶。我们展示了如何将它学到的知识应用于实际的驾驶问题，即模拟和规划。对于模拟，我们用未来的自我轨迹作为附加条件对预训练模型进行微调，以将未来的想象与不同的自我动作联系起来。我们还赋予

¹注意，GenAD 是“生成模型”和“广义能力”的缩写。

	Dataset	Duration (hours)	Front-view Frames	Geographic Diversity Countries Cities	Sensor Setup
	KITTI [14	1.4	15k		fixed
	Cityscapes [10	0.5	25k		fixed
	Waymo Open 43	11	390k	50	fixed
	Argoverse 2 48	4.2	300k		fixed
✓	nuScenes [5.5	241k		fixed
✓	nuPlan	120	4.0M		fixed
✓	Talk2Car [12	4.7			fixed
✓	ONCE [34	144	7M		fixed
✓	Honda-HAD [24	32	1.2M		fixed
✓	Honda-HDD-Action [40	104	1.1M		fixed
✓	Honda-HDD-Cause [40	32			fixed
✓	OpenDV-YouTube (Ours)	1747	60.2M	40	244
	OpenDV-2K (Ours)	2059	65.1M	40	244
					uncalibrated
					uncalibrated

表1. OpenDV-2K与现有同类数据集在规模和多样性方面的简要比较。请注意，带有✓的数据集包含在OpenDV-2K中（最后一行）。*Waymo Open、Argoverse 2和nuPlan中的感知子集。[†]根据视频标题由GPT [36]估算。

GenAD利用轻量级规划器将潜在特征转化为自车未来轨迹，从而在具有挑战性的基准上执行规划。凭借其预训练的预测准确未来帧的能力，我们的算法在仿真一致性和规划可靠性方面均展现出良好的结果。

2. OpenDV-2K 数据集

我们推出了OpenDV-2K，这是一个用于自动驾驶的大规模多模态数据集，旨在支持训练一个通用的视频预测模型。其主要组成部分是一个庞大的高质量YouTube驾驶视频库，这些视频从世界各地收集而来，并在经过精心筛选后纳入我们的数据集中。我们使用视觉语言模型自动为这些视频创建语言注释。为了进一步提高其在传感器配置和语言表达上的多样性，我们将7个公开许可的数据集合并到我们的OpenDV-2K中，如表1所示。因此，OpenDV-2K总共包含了2059小时的视频与文本配对，其中1747小时来自YouTube，312小时来自公共数据集。我们分别使用OpenDV-YouTube和OpenDV-2K来指代YouTube部分和整体数据集。

2.1. 优先数据集的多样性

表1提供了与其他公共数据集的简要对比。除了其显著的规模外，所提出的OpenDV-2K在以下各个方面也展现了多样性。

全球地理分布。由于在线视频的全球性，OpenDV-2K涵盖了全球超过40个国家和244个城市。相较于以往的公共数据集，通常仅在少数限制区域内收集数据，这是一个巨大的改进。我们在图2中绘制了OpenDV-YouTube的具体分布情况。

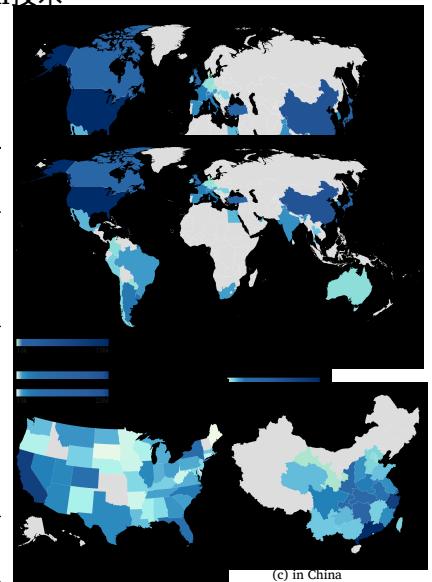


图2. OpenDV-2K的地理分布。我们的数据集涵盖了全球各地的多样化驾驶场景。

开放世界驾驶场景。我们的数据集提供了大量现实世界的驾驶体验，涵盖了如森林等罕见环境、如大雪等极端天气条件，以及在互动交通情况下的适当驾驶行为。这些数据对于多样性和泛化至关重要，但在现有的公共数据集中很少收集。

无限制的传感器配置。当前的驾驶数据集仅限于特定的传感器配置，包括内在和外在的相机参数、图像传感器类型、光学设备等，这为使用不同传感器部署学习到的模型带来了巨大的挑战[32]。相比之下，YouTube驾驶视频是在各种类型的车辆中用灵活的相机设置录制的，这有助于在使用新相机设置部署时增强训练模型的鲁棒性。

2.2. 迈向高质量多模态数据集

驾驶视频采集与筛选。从庞大的网络资源中找到干净的驾驶视频是一项繁琐且成本高昂的任务。为了简化这一过程，我们首先选择特定的视频上传者，即YouTube用户。根据平均时长和整体质量，我们收集了43名YouTube用户上传的2139段高质量前视驾驶视频。为了确保训练集和验证集之间没有重叠，我们选择3名YouTube用户的所有视频作为验证集，其余视频作为训练集。为了排除非驾驶片段，如视频介绍和订阅提醒，我们丢弃每段视频开头和结尾的特定长度片段。然后，使用VLM模型BLIP-2 [30]对每一帧进行语言上下文描述。通过手动检查这些上下文中是否存在特定关键词，进一步去除不适合训练的黑帧和过渡帧。我们给出了一个示意。

在附录C.1.1的数据集构建流程中，我们介绍了如何生成以下上下文。

YouTube视频语言标注。为了创建一个可通过自然语言控制以模拟不同未来的预测模型，为了使预测模型可控并提高样本质量[3]，将驾驶视频与有意义且多样的语言标注配对至关重要。我们为OpenDV-YouTube构建了两种类型的文本，即针对自我车辆的驾驶指令和帧描述，分别称为“指令”和“上下文”，以帮助模型分别理解自我行为和开放世界概念。对于指令，我们在Honda-HDD-Action[40]上训练了一个视频分类器，用于14种动作类型，以标注4秒序列中的自我行为。这些分类命令将进一步映射到预定义字典中的多个自由形式表达。对于上下文，我们利用已建立的视觉语言模型BLIP-2[30]来描述每帧中的主要对象和场景。有关标注的更多详情，请参阅附录C.1.2。

扩大语言谱系使用公共数据集。考虑到BLIP-2的注释是为静态帧生成的，无法理解动态驾驶场景，如交通灯转换，我们利用了多个提供驾驶场景语言描述的公共数据集[6, 7, 12, 24, 34, 40]。然而，这些数据集的元数据相对稀疏，仅包含如“阳光道路”等几个词。我们进一步使用GPT[36]提升其文本质量，形成描述性的“上下文”，并通过分类每个视频片段的记录轨迹生成“指令”。最终，我们将这些数据集与OpenDV-YouTube整合，建立了OpenDV-2K数据集，如表1最后一行所示。

3. GenAD框架

在本节中，我们介绍GenAD模型的训练和设计。如图3所示，GenAD的训练分为两个阶段，即图像域转换和视频预测预训练。第一阶段将通用文本到图像模型适配到驾驶领域（第3.1节）。第二阶段通过我们提出的时间推理块和修改后的训练方案，将文本到图像模型提升为视频预测模型（第3.2节）。在第3.3节中，我们探讨了预测模型如何扩展到动作条件预测和规划。

3.1. 图像域转换

车载摄像头捕捉到广阔的视野，包含丰富的视觉内容，如道路、背景建筑、周围车辆等，这些内容需要强大的生成能力来产生连续且逼真的驾驶场景。为了促进学习过程，我们

首先从独立图像生成开始，进入第一阶段。具体而言，我们以SDXL [37]初始化模型，这是一个用于文本到图像生成的大型潜在扩散模型（LDM），以利用其合成高质量图像并包含丰富视觉细节的能力。该模型实现为一个去噪UNet f ，包含多个堆叠的卷积和注意力块，通过去噪噪声潜在变量来学习合成图像[41]。具体来说，给定一个由前向扩散过程破坏的噪声输入潜在变量 x_t ，模型通过以下目标训练以预测 x_t 中添加的噪声 ϵ ：

$$\mathcal{L}_{\text{img}} := \mathbb{E}_{\mathbf{x}, \epsilon \sim \mathcal{N}(0, 1), \mathbf{c}, t} \left[\|\epsilon - \mathbf{f}_\theta(\mathbf{x}_t; \mathbf{c}, t)\|_2^2 \right],$$

其中， \mathbf{x} 和 \mathbf{x}_t 分别表示干净和带噪的潜在变量， t 表示不同噪声尺度的时间步， \mathbf{c} 是指导去噪过程的文本条件，它是上下文和命令的串联。为了提高训练效率，学习过程在压缩的潜在空间 [13, 37, 41] 中进行，而非像素空间。在采样过程中，模型通过迭代去噪最后一步的预测，从标准高斯噪声生成图像。然而，原始的 SDXL 是在通用领域数据上训练的，如肖像和艺术画作，这些数据与自主系统无关。为了使模型适应生成驾驶图像，我们在 OpenDV-2K 中的图文对上进行文本到图像生成的微调，目标与公式 (1) 相同。遵循 SDXL 的原始训练方式，在这一阶段，所有 UNet 的参数 都会被微调，而 CLIP 文本编码器 [39] 和自编码器 [13] 则保持冻结状态。

3.2. 视频预测预训练

在第二阶段，以一段连续视频的若干帧作为过去观察结果，GenAD被训练来推理所有视觉观察，并以合理的方式预测未来几帧。与第一阶段类似，预测过程也可以由文本条件引导。然而，由于两个根本障碍，按时间顺序预测高度动态的驾驶世界是具有挑战性的。

1. 因果推理：为了预测驾驶世界中遵循时间因果关系的未来可能性，模型需要理解所有其他代理以及自行车的意图，并理解基本的交通规则，例如，交通灯状态转变时交通将如何变化。

2. 剧烈视角变化：与主要具有静态背景且中心物体缓慢移动的传统视频生成基准不同，驾驶视角随时间发生剧烈变化。每一帧中的每个像素在下一帧中可能会移动到远处的位置。

我们提出了时间推理块来解决这些问题。如图3(c)所示，每个块由三个连续的注意力层组成，即因果

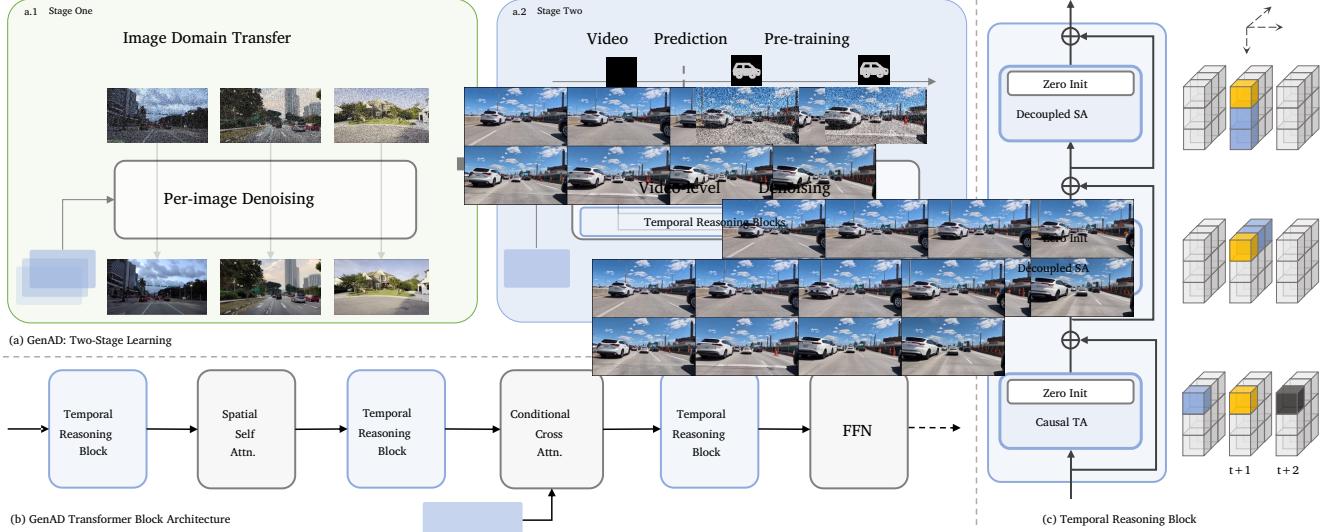


图3. GenAD框架。(a) GenAD的两阶段学习包括将图像扩散模型的图像域转移到驱动域(a.1 第一阶段)，以及用于建模视频时间依赖性的视频预测预训练(a.2 第二阶段)。(b) 在第二阶段训练中，GenAD的一个Transformer块在每个冻结层之前交错插入时间推理块，以对齐时空特征。(c) 提出的时间推理块包括一个因果时间注意力(TA)和两个解耦的空间注意力(SA)层，用于在不同轴上提取特征。查询网格不仅关注自身，还关注蓝色网格，而在因果注意力中，深灰色网格被屏蔽。每个注意力块的末尾附加了“零初始化”以稳定训练。

时间注意力层和两个解耦的空间注意力层，分别针对因果推理和驾驶场景中的大幅变化进行定制。

因果时间注意力。由于经过第一阶段训练的模型只能独立处理每一帧，我们利用时间注意力机制在不同视频帧之间交换信息。注意力发生在时间轴上，并建模每个网格特征的时间依赖性。然而，直接采用像[4, 18, 46, 51]中的双向时间注意力机制，很难获得因果推理的能力，因为预测将不可避免地依赖于后续帧而非过去条件。因此，我们通过添加因果注意力掩码来限制注意力方向，如图3(c)的最后一行所示，以鼓励模型充分利用过去的观察知识，并忠实地推理未来，就像在真实世界驾驶中一样。我们实证发现，因果约束大大规范了预测帧与过去帧的一致性。按照惯例，我们还添加了时间偏置，通过在时间轴上实现相对位置嵌入[42]来区分序列中的不同帧，以用于时间注意力。

解耦空间注意力。由于驾驶视频具有快速透视变化的特点，特定网格中的特征在不同时间步长内可能会有很大差异，且难以通过时间注意力机制进行关联和学习，因为其感受野有限。鉴于此，我们引入了空间注意力机制。

最初的关注点在于在空间轴上传播每个网格特征，以帮助收集用于时间关注的信息。我们实现了一种解耦的自注意力变体，因其具有线性计算复杂度的高效性，相比二次全自注意力。如图3(c)所示，这两个解耦的注意力层分别在水平和垂直轴上传播特征。

深度交互。直观上，第一阶段微调的空间块独立地优化每一帧的特征，使其更接近照片真实感，而第二阶段引入的时间块则对齐所有视频帧的特征，以实现连贯性和一致性。为了进一步增强时空特征的交互作用，我们将所提出的时间推理块与SDXL中的原始Transformer块交错排列，即空间注意力、交叉注意力和前馈网络，如图3(b)所示。

零初始化。与之前的做法[1, 52]类似，对于在第二阶段新引入的每个模块，我们将其最终层的所有参数初始化为零。这避免了在开始时破坏已训练良好的图像生成模型的先验知识，并稳定了训练过程。

训练。GenAD通过联合去噪从带有噪声的潜在表示中预测未来，这些潜在表示在过去的帧和文本条件的指导下进行。我们首先将一个视频片段的连续T帧投影到一个潜在表示的批次 $v = \{v^m, v^n\}$ 中，其中前 m 帧的潜在表示 v^m 是干净的，代表历史观察，而其他 $n = T - m$ 帧的潜在表示

v^n 表示待预测的未来。 v^n 随后通过正向扩散过程被破坏为 v^{n+1} ，其中 t 索引了一个随机采样的噪声尺度。模型被训练来预测以观察值 v^m 和文本 c 为条件的 v^{n+1} 的噪声。视频预测模型的学习目标表述如下：

$$\mathcal{L}_{\text{vid}} := \mathbb{E}_{v, \epsilon \sim \mathcal{N}(0, 1), c, t} \left[\|\epsilon - f_{\theta, \phi}(v_t^n; v^m, c, t)\|_2^2 \right],$$

其中， ϵ 表示继承的第一阶段模型，而 ϕ 代表新插入的时间推理模块。根据 [4]，我们冻结 ϵ ，仅训练时间推理模块，以避免干扰图像生成模型的生成能力，并专注于学习视频中的时间依赖关系。值得注意的是，只有来自损坏帧 v^{n+1} 的输出对训练损失有贡献，而来自条件帧 v^m 的输出则被忽略。

3.3. 扩展

基于在驾驶场景中训练有素的视频预测能力，我们进一步挖掘了预训练模型在动作控制预测和规划中的潜力，这对现实世界的驾驶系统至关重要。在此，我们探索了 nuScenes [6] 上的下游任务，该数据集提供了记录的姿态。

动作条件预测。为了使我们的预测模型能够通过精确的自行车动作进行控制，并作为模拟器 [25]，我们通过将未来轨迹作为额外条件进行微调来优化模型。具体来说，我们将原始轨迹映射到高维特征中，使用傅里叶嵌入 [44]。经过线性层的进一步投影后，将其添加到原始条件下。因此，自行车动作通过图3(b)中的条件交叉注意力层注入到网络中。

规划。通过学习预测未来，GenAD 获得了对复杂驾驶场景的强大表示能力，这些表示能力可以进一步用于规划。具体来说，我们通过冻结的 GenAD 的 UNet 编码器提取两个历史帧的时空特征，该编码器的大小接近整个模型的一半，并将这些特征输入多层感知器 (MLP) 以预测未来的路径点。通过冻结的 GenAD 编码器和一个可学习的 MLP 层，我们的规划器的训练过程可以比端到端规划模型 UniAD [22] 快 3400 倍，验证了 GenAD 学习到的时空特征的有效性。

4. 实验

4.1. 设置与协议

GenAD 在 OpenDV-2K 上分两个阶段进行学习，但具有不同的学习目标（在第3节中）和输入格式。在第一阶段，模型接收输入（图像，文本）对。

Method	Training Dataset	Pred.	nuScenes	
			FID (FVD (
DriveGAN [25]		✓	73.4	502
DriveDreamer [45]	nuScenes	✓	52.6	452
DrivingDiusion [31]			15.8	332
GenAD-nus (Ours)	nuScenes	✓	15.4	244
GenAD (Ours)	OpenDV-2K	✓	15.4	184

表2. 视频生成质量与基于nuScenes训练的最先进方法的比较。“预测”：未来预测的评估。*：需要3D布局输入。

在文本到图像生成任务上进行了训练。我们将命令注释广播到每个4秒视频序列所包含的所有帧上。该模型在32块NVIDIA Tesla A100 GPU上进行了30万次迭代训练，总批量大小为256。在第二阶段，GenAD被训练以联合去噪未来潜在变量，条件是过去的潜在变量和文本。其输入为（视频片段，文本）对，其中每个视频片段为4秒，采样率为2Hz。当前版本的GenAD在64块GPU上进行了11.25万次迭代训练，总批量大小为64。输入帧在两个阶段的训练中都被调整为 256×448 大小，并且以 $p = 0.1$ 的概率丢弃文本条件 c ，以实现无分类器指导 [17] 的采样，这在扩散模型中常用于提高样本质量。更多训练和采样细节见附录D。

4.2. 视频预测预训练结果

与近期视频生成方法的比较。我们将GenAD与OpenDV-YouTube、Waymo [43]、KITTI [14] 和 Cityscapes [10] 的未见数据集进行对比，采用零样本生成方式。图4展示了定性结果。图像到视频模型I2VGen-XL [53] 和 VideoCrafter1 [8] 无法严格遵循给定帧进行预测，导致预测帧与过去帧之间的一致性较差。在Cityscapes上训练的视频预测模型DMVFN [21] 在其预测中存在不利的形状扭曲，特别是在这三个未见数据集上。相比之下，GenAD展现了卓越的零样本泛化能力和视觉质量，尽管这些数据集均未包含在训练中。

与nuScenes专家的比较。我们还比较了GenAD与最新的、专门为nuScenes训练的驾驶视频生成模型。表2显示，GenAD在图像保真度 (FID) 和视频连贯性 (FVD) 方面均超越了以往的所有方法。具体而言，GenAD相比 DrivingDiusion [31] 显著降低了FVD达44.5%，且无需额外输入3D未来布局。为公平比较，我们仅在nuScenes 数据集上训练了一个模型变体 (GenAD-nus)。我们发现，尽管GenAD-nus表现相当，

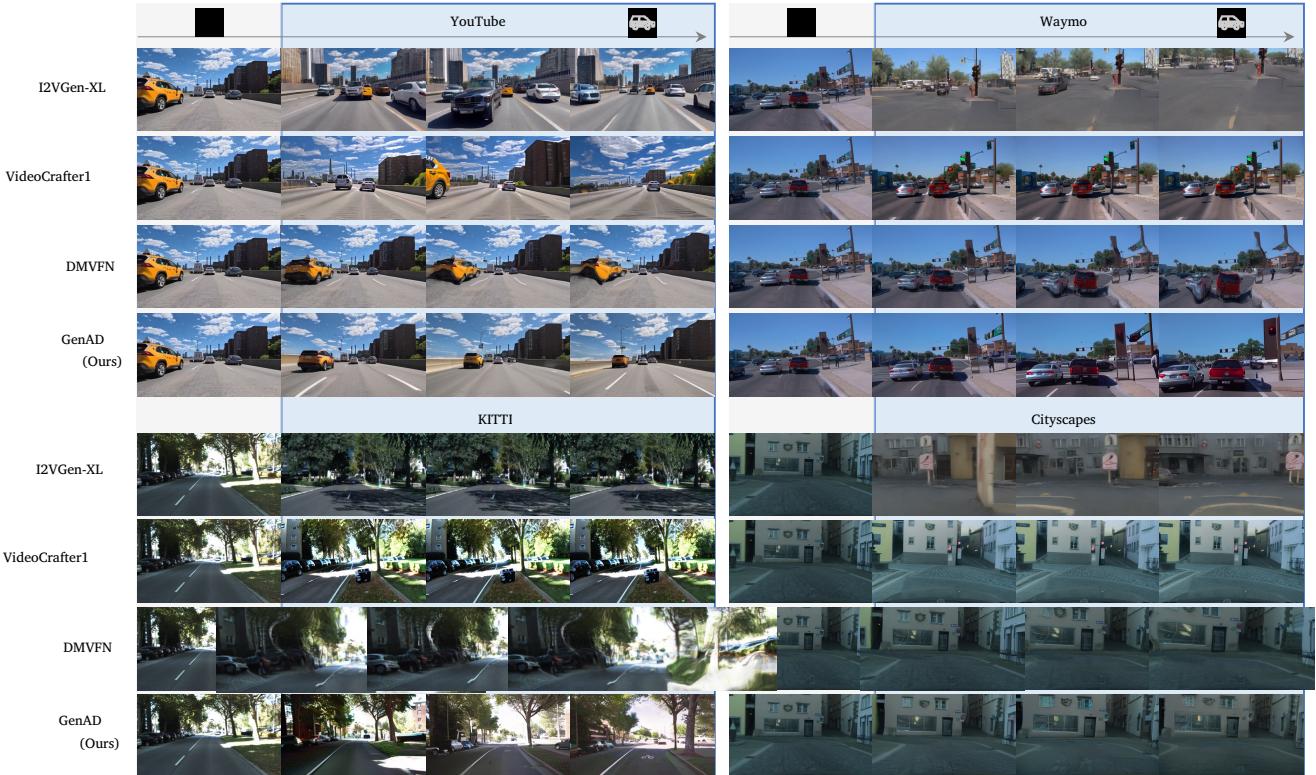


图4. 针对未见场景的零样本视频预测任务。我们展示了在相同初始帧下不同模型的生成结果（蓝色框内）。GenAD在未见数据集（场景）上做出了更为稳健、真实且合理的未来预测。
更多比较和可视化内容展示在附录中。



图5. 语言条件预测任务。给定交叉路口雨天场景的两帧图像和三个高级文本条件，GenAD相应地模拟出合理的未来情景。

在nuScenes上使用GenAD时，它难以推广到像Waymo这样的未见过的数据集，生成的结果退化为nuScenes的视觉模式。相比之下，在OpenDV-2K上训练的GenAD展现了强大的跨数据集泛化能力，如图4所示。

我们在图5中提供了nuScenes上的语言条件预测样本，其中GenAD模拟了各种未来情况。

从相同的起点出发，遵循不同的文本指令。其令人印象深刻的生成质量体现在环境的复杂细节和自我运动的自然过渡中。

消融研究。我们通过对OpenDV-2K的一个子集进行75K步的训练来执行消融实验。从仅使用普通时间注意力的基线模型[4, 18]开始，我们逐步引入我们提出的组件。值得注意的是，通过将时间块与空间块交错排列，FVD显著提升（-17%），这是由于更充分的时空交互。时间因果性和解耦的空间注意力均有助于提升CLIP-SIM，增强了未来预测与条件帧之间的时间一致性。需要澄清的是，FID和FVD在表3的第四行和第三行中略有增加，这并不准确反映生成质量的下降，如[4, 5, 37]中所讨论的。每个设计的效果如图6所示。

4.3. 扩展结果

行动条件预测。我们在图7和表4中进一步展示了在nuScenes上微调的行动条件模型GenAD-act的性能。给定两个起始帧和一个由6个未来路径点组成的轨迹 w ，



图6. 模型设计案例研究。所有组件均有助于缓解伪影并提高未来预测的一致性。

Method	YouTube		
	FID (FVD (CLIPSIM (
Baseline	18.32	244.44	0.8405
+ Deep Interaction	17.96	201.69	0.8409
+ Temporal Causality	16.54	207.45	0.8550
+ Decoupled Spatial Attn.	17.67	189.54	0.8652

表3. 在GenAD中模型设计的消融研究。所有提出的设计都对最终性能有所贡献。

Method	Condition	nuScenes	
		Action	Prediction Error (
Ground truth		0.90	
GenAD	text	2.54	
GenAD-act	text + traj.	2.02	

表4. 基于动作条件的预测任务。与仅使用文本条件的GenAD相比，GenAD-act能够实现更精确的、符合动作条件的未来预测。

points, GenAD-act 设想了沿着轨迹序列的 6 个未来帧。为了评估输入轨迹 w 与预测帧之间的一致性，我们在 nuScenes 上建立了一个逆动力学模型 (IDM) 作为评估器，该模型将视频序列投影到相应的自我轨迹。我们利用 IDM 将预测帧转换为轨迹 \hat{w} ，并计算 w 和 \hat{w} 之间的 L2 距离作为动作预测误差。具体而言，与带有文本条件的 GenAD 相比，GenAD-act 显著降低了 20.4% 的动作预测误差，从而实现了更准确的未来模拟。

规划结果。表5展示了在nuScenes上的规划结果，其中自车车辆的地面真值姿态可用。通过冻结GenAD编码器并仅进行优化



图7. 动作条件预测任务（模拟）。给定相同的起始帧和不同的未来轨迹

（在第一列中以黄色点显示），GenAD-act 可以模拟不同的未来，遵循不同的自我意图。更可视化内容见附录。

Method	# Trainable Params.	nuScenes	
		ADE (FDE (
ST-P3	20	10.9M	2.65
UniAD	22	58.8M	1.03
GenAD (Ours)	0.8M	1.23	2.31

表5. 开环规划任务。使用冻结的GenAD的轻量级MLP在仅使用前视图像的情况下，以73倍更少的可训练参数获得了具有竞争力的规划结果。*：多视角输入。评估协议与UniAD [22]一致。

在其之上增加一个额外的MLP，模型能够有效学习规划。值得注意的是，通过预先提取GenAD的UNet编码器的图像特征，整个规划适应的学习过程在单个NVIDIA Tesla V100设备上仅需10分钟，这比UniAD规划器的训练效率高3400倍[22]。

5. 限制与讨论

我们研究了GenAD的系统级开发，GenAD是一个大规模的广义视频预测模型，用于自动驾驶。我们还验证了GenAD学习到的表示对驾驶任务的适应性，即学习“世界模型”和运动规划。尽管我们在开放领域的泛化方面取得了改进，但增加的模型容量在训练效率和实时部署方面带来了挑战。我们设想，统一的视频预测任务将作为未来研究表示学习和策略学习的可扩展目标。另一个有趣的方向涉及将编码的知识提炼出来，以应用于更广泛的下游任务[29]。

参考文献

- [1] Jean-Baptiste Alayrac, Je Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, 和 Karen Simonyan. Flamingo: 一种用于少样本学习的视觉语言模型。在 NeurIPS, 2022. 2, 5
- [2] Mohammadhossein Bahari, Saeed Saadatnejad, Ahmad Rahimi, Mohammad Shaverdikondori, Amir Hossein Shahidzadeh, Seyed-Mohsen Moosavi-Dezfooli, 和 Alexandre Alahi. 车辆轨迹预测有效，但并非处处适用。在 CVPR, 2022. 2
- [3] Fan Bao, Chongxuan Li, Jiacheng Sun, 和 Jun Zhu. 为什么条件生成模型比无条件生成模型更好？arXiv 预印本 arXiv:2212.00362, 2022. 4
- [4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, 和 Karsten Kreis. 对齐你的潜在变量：使用潜在扩散模型生成高分辨率视频。在 CVPR, 2023. 2, 5, 6, 7
- [5] Tim Brooks, Janne Hellsten, Miika Aittala, Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei Efros, 和 Tero Karras. 生成动态场景的长视频。在 NeurIPS, 2022. 7
- [6] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yuxin Pan, Giancarlo Baldan, 和 Oscar Beijbom. nuScenes: 一种用于自动驾驶的多模态数据集。在 CVPR, 2020. 3, 4, 6
- [7] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric M. Wol , Alex Lang, Luke Fletcher, Oscar Beijbom, 和 Sammy Omari. nuPlan: 一种用于自动驾驶车辆的闭环基于机器学习的规划基准。在 CVPR Workshops, 2021. 3, 4
- [8] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, 和 Ying Shan. VideoCrafter1: 用于高质量视频生成的开源扩散模型。arXiv 预印本 arXiv:2310.19512, 2023. 6
- [9] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, 和 Hongyang Li. 端到端自动驾驶：挑战与前沿。arXiv 预印本 arXiv:2306.16927, 2023. 2
- [10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, 和 Bernt Schiele. Cityscapes 数据集：用于语义城市场景理解的基准数据集。在 CVPR, 2016. 3, 6
- [11] Yilun Dai, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, 和 Pieter Abbeel. 通过文本引导的视频生成学习通用策略。在 NeurIPS, 2023. 2
- [12] Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Luc Van Gool, 和 Marie-Francine Moens. Talk2Car: 控制你的自动驾驶汽车。在 EMNLP 2019 上。3, 4
- [13] Patrick Esser, Robin Rombach, 和 Bjorn Ommer. 驯服变换器以进行高分辨率图像合成。在 CVPR 2021 上。4
- [14] Andreas Geiger, Philip Lenz, Christoph Stiller, 和 Raquel Urtasun. 视觉与机器人学：KITTI 数据集。IJRR, 2013. 3, 6
- [15] Agrim Gupta, Stephen Tian, Yunzhi Zhang, Jiajun Wu, Roberto Mart'ın-Mart'ın, 和 Li Fei-Fei. MaskViT: 用于视频预测的掩码视觉预训练。在 ICLR 2023 上。2
- [16] Niklas Hanselmann, Katrin Renz, Kashyap Chitta, Apratim Bhattacharyya, 和 Andreas Geiger. KING: 通过运动梯度生成关键驾驶场景以实现鲁棒模仿。在 ECCV 2022 上。2
- [17] Jonathan Ho 和 Tim Salimans. 无分类器扩散指导。在 NeurIPS 研讨会 2021 上。6
- [18] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, 和 David J Fleet. 视频扩散模型。在 NeurIPS 2022 上。2, 5, 7
- [19] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, 和 Gianluca Corrado. GAIA-1: 用于自动驾驶的生成世界模型。arXiv 预印本 arXiv:2309.17080, 2023. 2
- [20] Shengchao Hu, Li Chen, Penghao Wu, Hongyang Li, Junchi Yan, 和 Dacheng Tao. ST-P3: 通过时空特征学习实现端到端视觉自动驾驶。在 ECCV 2022 上。8
- [21] Xiaotao Hu, Zhewei Huang, Ailin Huang, Jun Xu, 和 Shuchang Zhou. 一种动态多尺度体素流网络用于视频预测。在 CVPR 2023 上。6
- [22] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, 和 Hongyang Li. 面向规划的自动驾驶。在 CVPR 2023 上。6, 8
- [23] Fan Jia, Weixin Mao, Yingfei Liu, Yucheng Zhao, Yuqing Wen, Chi Zhang, Xiangyu Zhang, 和 Tiancai Wang. Adriver-i: 一种通用的自动驾驶世界模型。arXiv 预印本 arXiv:2311.13549, 2023. 2
- [24] Jinkyu Kim, Teruhisa Misu, Yi-Ting Chen, Ashish Tawari, 和 John Canny. 将人类对车辆的建议应用于自动驾驶车辆。在 CVPR 2019 上。3, 4
- [25] Seung Wook Kim, Jonah Philion, Antonio Torralba, 和 Sanja Fidler. DriveGAN: 实现可控的高质量神经模拟。在 CVPR 2021 上。2, 6
- [26] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, 和 Ross Girshick. 分割一切。在 ICCV 2023 上。2
- [27] Yann LeCun. 通往自主机器智能版本 0.9 的路径。2, 2022-06-27. Open Review, 62, 2022. 2
- [28] Yann LeCun, Yoshua Bengio, 和 Geoffrey Hinton. 深度学习。Nature, 2015. 2

- [29] Daiqing Li, Huan Ling, Amlan Kar, David Acuna, Seung Wook Kim, Karsten Kreis, Antonio Torralba, 和 Sanja Fidler. DreamTeacher: 使用深度生成模型预训练图像骨干网络。在 ICCV, 2023. 8
- [30] Junnan Li, Dongxu Li, Silvio Savarese, 和 Steven Hoi. BLIP-2: 通过冻结图像编码器和大型语言模型引导语言-图像预训练。在 ICML, 2023. 2, 3, 4
- [31] Xiaofan Li, Yifu Zhang, 和 Xiaoqing Ye. DrivingDiffusion: 利用布局引导的多视角驾驶场景视频生成与潜在扩散模型。arXiv 预印本 arXiv:2310.07771, 2023. 6
- [32] Zhenyu Li, Zehui Chen, Ang Li, Liangji Fang, Qinhong Jiang, Xianming Liu, 和 Junjun Jiang. 通过自训练实现单目3D物体检测的无监督域适应。在 ECCV, 2022. 3
- [33] Jiachen Lu, Ze Huang, Jiahui Zhang, Zeyu Yang, 和 Li Zhang. Wovogen: 世界体积感知扩散用于可控的多摄像头驾驶场景生成。arXiv 预印本 arXiv:2312.02934, 2023. 2
- [34] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Hanxue Liang, Xiaodan Liang, Yamin Li, Chao Ye, Wei Zhang, Zhenguo Li, Jie Yu, Hang Xu, 和 Chunjing Xu. 自动驾驶的一百万个场景: ONCE 数据集。在 NeurIPS 数据集与基准测试, 2021. 3, 4
- [35] OpenAI. Gpt-4 技术报告。arXiv 预印本 arXiv:2303.08774, 2023. 2
- [36] Long Ouyang, Je rey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, 和 Ryan Lowe. 通过人类反馈训练语言模型以遵循指令。在 NeurIPS, 2022. 3, 4
- [37] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, 和 Robin Rombach. SDXL: 改进潜在扩散模型以进行高分辨率图像合成。arXiv 预印本 arXiv:2307.01952, 2023. 2, 4, 7
- [38] Xiaojuan Qi, Zhengzhe Liu, Qifeng Chen, 和 Jiaya Jia. 3D 运动分解用于 RGBD 未来动态场景合成。在 CVPR, 2019. 2
- [39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, 和 Ilya Sutskever. 从自然语言监督中学习可迁移的视觉模型。在 ICML, 2021. 2, 4
- [40] Vasilii Ramanishka, Yi-Ting Chen, Teruhisa Misu, 和 Kate Saenko. 迈向驾驶场景理解: 一个用于学习驾驶员行为和因果推理的数据集。在 CVPR, 2018. 3, 4
- [41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, 和 Björn Ommer. 使用潜在扩散模型进行高分辨率图像合成。在 CVPR, 2022. 2, 4
- [42] Peter Shaw, Jakob Uszkoreit, 和 Ashish Vaswani. 带有相对位置表示的自注意力。arXiv 预印本 arXiv:1803.02155
- [43] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott M. Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, 和 Dragomir Anguelov. 自动驾驶中的感知可扩展性: Waymo 开放数据集。在 CVPR, 2020. 3, 6
- [44] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, 和 Ren Ng. 傅里叶特征让网络在低维度领域学习高频函数。在 NeurIPS, 2020. 6
- [45] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, 和 Jiwen Lu. DriveDreamer: 面向真实世界驱动的自动驾驶世界模型。arXiv 预印本 arXiv:2309.09777, 2023. 2, 6
- [46] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, Yuwei Guo, Tianxing Wu, Chenyang Si, Yuming Jiang, Cunjian Chen, Chen Change Loy, Bo Dai, Dahua Lin, Yu Qiao, 和 Ziwei Liu. LAVIE: 使用级联潜在扩散模型生成高质量视频。arXiv 预印本 arXiv:2309.15103, 2023. 5
- [47] Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, 和 Zhaoxiang Zhang. 驶向未来: 使用世界模型进行多视角视觉预测和规划的自动驾驶。arXiv 预印本 arXiv:2311.17918, 2023. 2
- [48] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, 和 James Hays. Argoverse 2: 下一代用于自动驾驶感知和预测的数据集。在 NeurIPS 数据集和基准测试, 2021. 3
- [49] Daniel M Wolpert, Zoubin Ghahramani, 和 Michael I Jordan. 用于感觉运动整合的内部模型。Science, 1995. 2
- [50] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, 和 Tao Gui. 大型语言模型基础代理的崛起与潜力: 一项调查。arXiv 预印本 arXiv:2309.07864, 2023. 2
- [51] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, 和 Mike Zheng Shou. Show-1: 结合像素与潜在扩散模型进行文本到视频生成。arXiv 预印本 arXiv:2309.15818, 2023. 5
- [52] Lvmin Zhang, Anyi Rao, 和 Maneesh Agrawala. 为文本到图像扩散模型添加条件控制。在 ICCV, 2023. 5
- [53] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, 和

, 2018. 5

周京仁。I2VGen-XL：通过级联扩散模型实现高质量的图像到视频合成。arXiv预印本arXiv:2311.04145，2023。6

[54] 朱瑞泽，黄鹏，Eshed Ohn-Bar，和Venkatesh Saligrama。学习随处驾驶。在CoRL，2023。2