

Vista: 高保真度与多功能可控性的通用驾驶世界模型

沈源高^{1,2}贾智杨²陈力^{2,5}Kashyap Chitta^{3,4}邱一航²

Andreas Geiger^{3,4†}张军^{1†}李宏洋^{2,5†}

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

代码和模型 : github.com/OpenDriveLab/Vista

演示页面 : opendrivelab.com/Vista

摘要

世界模型能够预见不同行动的结果，这对自动驾驶至关重要。然而，现有的驾驶世界模型在未见环境的泛化能力、关键细节的预测保真度以及灵活应用的动作可控性方面仍存在局限。本文提出了Vista，一种具有高保真度和多功能可控性的可泛化驾驶世界模型。基于对现有方法的系统诊断，我们引入了几个关键要素来解决这些局限。为了在高分辨率下准确预测现实世界的动态，我们提出了两种新颖的损失函数，以促进对移动实例和结构信息的学习。我们还设计了一种有效的潜在替换方法，将历史帧作为先验注入，以实现连贯的长时域推演。对于动作可控性，我们通过一种高效的学习策略，从高级意图（命令、目标点）到低级操作（轨迹、角度和速度）整合了一系列多功能控制。经过大规模训练后，Vista的能力能够无缝泛化到不同场景。在多个数据集上的广泛实验表明，Vista在超过70%的比较中优于最先进的通用视频生成器，并在FID和FVD上分别以55%和27%的优势超越了表现最佳的驾驶世界模型。此外，我们首次利用Vista自身的容量，在不访问真实动作的情况下，为现实世界动作评估建立了一种可泛化的奖励机制。

1 Introduction

在可扩展学习技术的推动下，自动驾驶在过去几年取得了令人鼓舞的进展[18, 58, 135]。然而，复杂和分布外的情况对于最先进的技术来说仍然难以处理[83]。一个有前景的解决方案在于世界模型[57, 76]，这些模型通过历史观测和替代行动推断世界可能的未来状态，从而评估这些行动的可行性。它们具有推理不确定性和避免灾难性错误[54, 76, 127]的潜力，从而促进自动驾驶的泛化和安全性。

尽管世界模型的主要前景是使其具备对新环境的泛化能力，但现有的驾驶世界模型仍受限于数据规模[90, 125, 127, 143, 147]和地理覆盖范围[54, 61]。如表1和图1所总结的，它们通常还局限于低帧率和低分辨率，导致关键细节的丢失。此外，大多数模型仅支持单一的控制模式，如方向盘角度和速度。这不足以表达从高级意图到低级操作的各种动作格式，并且与流行的规划算法[12, 14, 21, 56, 58, 64]的输出不兼容。此外，将动作可控性泛化到未见过的数据集的研究尚不充分。这些局限性阻碍了现有工作的应用，因此迫切需要开发一种能够克服这些局限性的世界模型。

主要联系人为高申远，邮箱：sygao@connect.ust.hk†同等指导。

表1：真实世界驾驶环境模型。基于大规模高质量驾驶数据训练，Vista在高空时分辨率下执行并支持多样的动作可控性。私人数据。

Method	Data Scale	Model Setups		Resolution	Action Control Modes		
		Frame Rate	Resolution		Angle&Speed	Trajectory	Command
DriveSim [102	7h	5 Hz	80 × 160	✓			
DriveGAN [68	160h	8 Hz	256 × 256	✓			
DriveDreamer [125	5h	12 Hz	128 × 192	✓			
Drive-WM [127	5h	2 Hz	192 × 384		✓		
WoVoGen [90	5h	2 Hz	256 × 448	✓			
ADriver-I [61	300h	2 Hz	256 × 512			✓	
GenAD [136	2000h	2 Hz	256 × 448		✓		✓
GAIA-1 [54	4700h	25 Hz	288 × 512	✓			
Vista (Ours)	1740h	10 Hz	576 × 1024	✓	✓	✓	✓

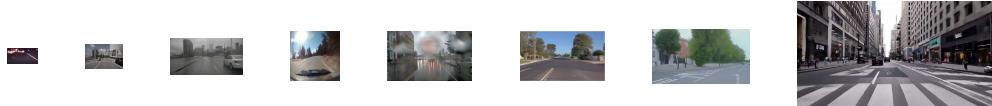


图1：分辨率比较。Vista 的预测分辨率高于以往文献。

为此，我们引入了Vista，这是一种擅长跨领域泛化、高保真预测和多模态行动控制能力的驾驶世界模型。具体而言，我们在大量全球驾驶视频[136]的基础上开发了预测模型，以提升其泛化能力。为了实现连贯的未来外推，我们在Vista中加入了三个基本动态先验条件（第3.1节）。我们不仅依赖于标准的扩散损失[5]，还引入了两种显式损失函数来增强动态效果并保留结构细节（第3.1节），从而提升Vista在高分辨率下模拟真实未来的能力。为了实现灵活的控制能力，我们整合了一系列多样化的行动格式，包括高级意图如指令和目标点，以及低级操作如轨迹、转向角度和速度。这些行动条件通过一个统一的接口注入，该接口通过高效的训练策略学习得到（第3.2节）。因此，如图2所示，Vista能够以10赫兹和576×1024像素的分辨率预见现实的未来，并在不同粒度级别上获得多样的行动控制能力。我们还展示了Vista作为可泛化的奖励函数来评估不同行动可靠性的潜力。

我们的贡献主要体现在三个方面：(1) 我们提出了Vista，一种可泛化的驾驶世界模型，能够在高时空分辨率下预测现实的未来场景。通过两种新颖的损失函数来捕捉动态并保持结构，以及详尽的动态先验知识来维持长期推演的一致性，Vista的预测准确性得到了显著提升。(2) 借助一种高效的学习策略，我们通过统一的调节接口将多样的动作控制能力整合到Vista中。Vista的动作控制能力还能以零样本的方式泛化到不同的领域。(3) 我们在多个数据集上进行了全面的实验，以验证Vista的有效性。它在最先进的通用视频生成器中表现出色，并在nuScenes数据集上创下了新的技术水平。我们的实证证据表明，Vista可以作为评估动作的奖励函数使用。

2Preliminary

我们使用预训练的Stable Video Diffusion (SVD) [5] 来初始化Vista，这是一个用于图像到视频生成的潜在扩散模型。为了提高采样灵活性，SVD采用了连续时间步公式 [66, 111]。它通过扩散过程 $p(n|x) \sim N(x, \sqrt{2}I)$ 将数据样本 x 转换为噪声 n ，并通过逐步对潜在空间进行去噪，从高斯噪声中生成新的样本，直至 $n = 0$ 。

SVD的训练可以简化为最小化 $\text{Ex}_n, nh \cdot \|D(n; \theta) - x\|^{2i}$ ，其中 D 是一个参数化的UNet去噪器， θ 是一个在此后为简洁起见省略的重新加权函数。基于此框架，SVD处理一系列噪声潜在 $n = \{n_1, n_2, \dots, n_K\} \in \mathbb{R}^{K \times C \times H \times W}$ ，并生成包含 $K = 25$ 帧的视频。生成过程由一个条件图像引导，该图像的潜在表示按通道方式连接到输入中，作为内容生成的参考。

尽管具有高审美质量，SVD 缺乏作为驾驶世界模型的几项关键属性。如第 4 节所示，SVD 预测的第一帧与条件图像不相同，这使得由于内容不一致而无法进行自回归展开。此外，SVD 在处理驾驶场景的复杂动态方面存在困难，导致出现不合理的运动。

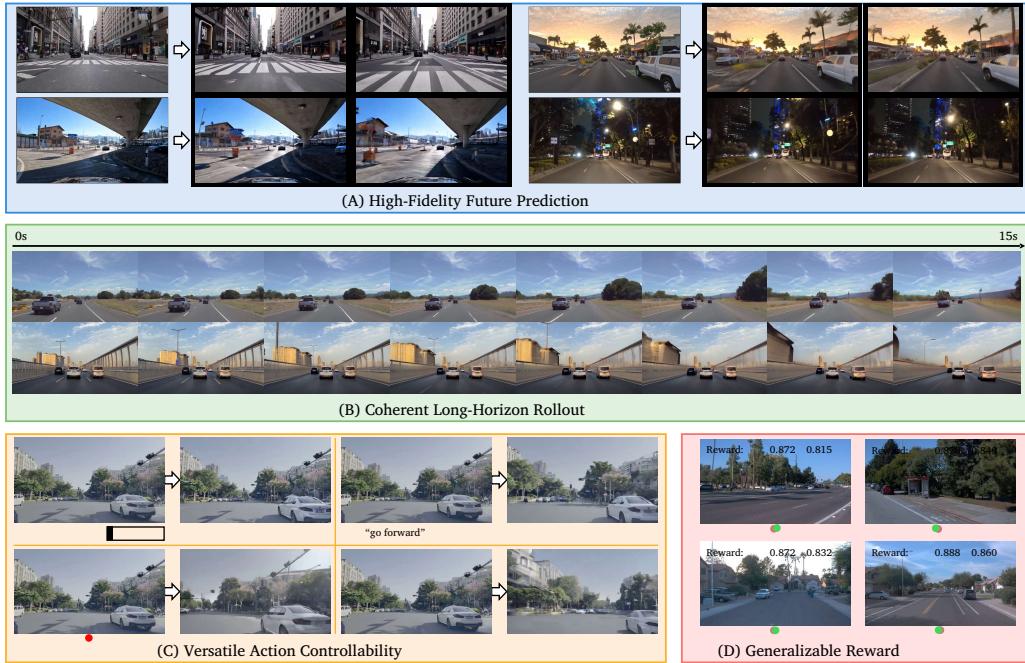


图2：Vista的功能。从任意环境出发，Vista能够预测高时空分辨率的现实且连续的未来（A-B）。它可以通过多模态动作进行控制（C），并且可以作为通用的奖励函数来评估现实世界的驾驶动作（D）。

SVD无法通过任何行动格式进行控制。相反，我们的目标是构建一个可泛化的驾驶世界模型，该模型能够预测具有现实动力学的高保真未来。它应当能够持续扩展至长时间范围，并能通过多模态行动进行灵活控制，如图2所示。

3学习一个可推广的驾驶世界模型

如图3所示，Vista采用了两阶段训练流程。首先，我们构建了一个专门的预测模型，该模型采用潜在替代方法以实现连贯的未来预测，并引入了两种新颖的损失函数来增强保真度（第3.1节）。为了确保模型对未见场景的泛化能力，我们利用了最大的公开驾驶数据集[136]进行训练。在第二阶段，我们整合了多模态动作，通过一种高效且协同的训练策略来学习动作的可控性（第3.2节）。利用Vista的能力，我们进一步提出了一种可泛化的动作评估方法（第3.3节）。

3.1 第一阶段：学习高保真未来预测

基本设置。由于世界模型旨在从当前状态预测未来，它们的预测起点应与条件图像紧密对齐。因此，我们将SVD定制为专用预测模型，将第一帧作为条件图像，并在训练过程中舍弃噪声增强[5, 49]。凭借这种预测能力，Vista能够通过迭代预测短期片段并将条件图像重置为最后片段，来执行长期推演。

动态先验注入。然而，使用上述设置进行训练通常会导致相对于历史帧的不合理动态，特别是在长期推演中。我们推测这主要源于未来运动趋势的先验信息不足所导致的模糊性，这也是现有驾驶世界模型的一个常见局限性 [54, 68, 125, 127, 136]。

估计连贯的未来场景需要至少三个基本先验条件，这些先验条件本质上支配着场景中实例的未来运动：位置、速度和加速度。由于速度和加速度分别是位置的一阶和二阶导数，因此这些先验条件可以通过使用三个连续的帧进行条件化来完全推导。具体来说，我们构建了一个帧级掩码 $\{m \in \{0, 1\}^K\}$ ，其长度为 K，用于指示条件帧的存在。掩码按时间顺序依次设置，最多有三个元素被分配为 1，以表示三个条件帧。

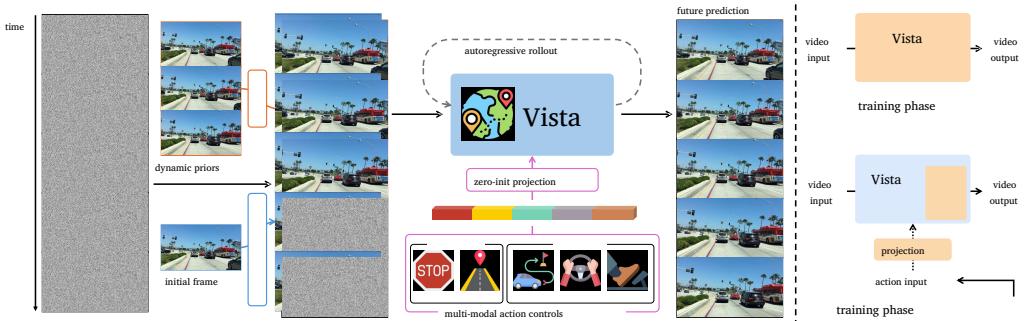


图3：[左]：Vista流水线。除了初始帧，Vista还可以通过潜在替换吸收更多关于未来动态的先验信息。其预测可以通过不同动作进行控制，并通过自回归展开扩展到长时间范围。[右]：训练过程。Vista采用两个训练阶段，第二阶段冻结预训练权重以学习动作控制。

frames. 我们通过用图像编码器编码的干净潜在变量 z_i 替换相应的噪声潜在变量 n_i 来注入新的条件帧，而不是将额外的通道连接到输入中。形式上，输入潜在变量构造为 $\hat{n} = m \cdot z + (1 - m) \cdot n$ （见图3 [左]）。为了区分干净潜在变量，我们从预训练权重中复制一个新的时间步嵌入，并根据 m 将其分配给条件帧。条件帧和预测帧的时间步嵌入分别进行训练。与通道级连接相比，我们发现替换潜在变量在吸收不同数量的条件帧时更为有效和灵活。此外，我们观察到，替换后的潜在变量在直接应用于SVD时不会降低其生成质量。因此，在训练开始时，原始性能不会受到影响。由于不需要预测观察到的条件帧，我们将它们从损失中排除，如下所示：

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{\mathbf{z}, \sigma, \hat{\mathbf{n}}} \left[\sum_{i=1}^K (1 - m_i) \odot \|D_\theta(\hat{n}_i; \sigma) - z_i\|^2 \right],$$

其中， D 是与SVD共享相同架构的UNet去噪器。在替换的潜在变量具备足够先验信息的情况下，Vista能够全面捕捉周围实例的状态，并通过迭代展开预测出更加连贯和合理的长远未来。在实际操作中，我们在展开过程中利用预测片段的最后三帧作为下一预测步骤的动态先验。

动态增强损失。与覆盖较小空间的普通视频不同，驾驶视频捕捉到了更大范围的场景 [136]。在大多数驾驶视频中，远处的单调区域占据了大部分视野，而移动的前景实例仅占据相对较小的区域 [17]。然而，后者通常表现出更高的随机性，使其预测变得复杂。由于公式 (1) 对所有输出进行统一监督，它无法有效区分不同区域的细微差别，如图 4(b) 所示。因此，模型无法高效地学习在关键区域预测现实的动态。

由于相邻两帧之间的差异提供了丰富的运动模式 [123, 132]，我们引入了一种额外的监督机制，以促进对关键区域动态的学习。具体而言，我们首先引入了一个动态感知权重 $w = \{w_2, w_3, \dots, w_k\} \in \mathbb{R}^{K-1 \times C \times H \times W}$ ，该权重突出了预测与真实值相比运动不一致的区域：

$$w_i = \|(D_\theta(\hat{n}_i; \sigma) - D_\theta(\hat{n}_{i-1}; \sigma)) - (z_i - z_{i-1})\|^2.$$

为了确保数值稳定性，我们在每个视频片段内对 w 进行归一化处理。如图4(c)所示，权重放大了运动差异较大的存在，突出了动态区域，同时排除了单调的背景。鉴于未来预测的因果关系，即后续帧应遵循前面的帧，我们通过惩罚每个相邻帧对的后续帧来定义一个新的损失：

$$\mathcal{L}_{\text{dynamics}} = \mathbb{E}_{\mathbf{z}, \sigma, \hat{\mathbf{n}}} \left[\sum_{i=2}^K \text{sg}(w_i) \odot (1 - m_i) \odot \|D_\theta(\hat{n}_i; \sigma) - z_i\|^2 \right],$$

其中 $\text{sg}(\cdot)$ 表示停止梯度。通过自适应地重新加权标准扩散损失， $\mathcal{L}_{\text{dynamics}}$ 可以提升动态区域的学习效率，例如图 4(d) 中的移动车辆和人行道。

结构保持损失。在视频生成中，感知质量与运动强度之间的权衡已被广泛认可 [3, 32, 73, 144]，我们的情况也不例外。当

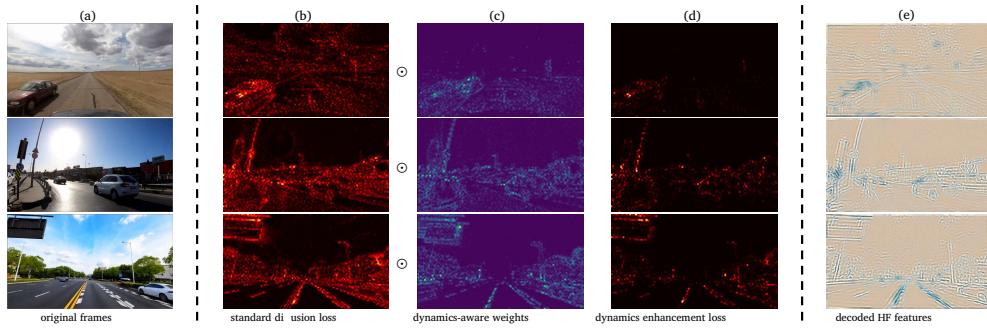


图4：损失设计的图示。与标准扩散损失（b）均匀分布不同，我们的动力学增强损失（d）能够自适应地集中于关键区域（c）（例如，移动车辆和路边）以进行动力学建模。此外，通过显式监督高频特征（e），可以增强对结构细节（例如，边缘和车道）的学习。

在高分辨率动态驾驶场景预测中，我们发现预测的结构细节会出现严重退化，表现为物体过度平滑或破碎，例如车辆的轮廓在移动过程中迅速解体（见图12）。为缓解这一问题，我们需要更加重视结构细节。鉴于结构细节（如边缘和纹理）主要存在于高频成分中，我们在频域中识别它们，具体如下：

$$z'_i = \mathcal{F}(z_i) = \text{IFFT}(\mathcal{H} \odot \text{FFT}(z_i)),$$

其中，FFT和IFFT分别表示二维离散傅里叶变换和逆离散傅里叶变换， \mathcal{H} 是一个理想的高通滤波器，用于截断低于某个阈值的低频分量。傅里叶变换分别独立地应用于 z_i 的每个通道。如图4(e)所示，通过公式(4)可以有效地强调与结构信息相关的特征。类似地，也可以从预测的潜在特征 $D_\theta(\hat{n}_i; \sigma)$ 中提取相应的特征。基于提取的高频特征，我们设计了一种新的结构保持损失，如下所示：

$$\mathcal{L}_{\text{structure}} = \mathbb{E}_{z, \sigma, \hat{n}} \left[\sum_{i=1}^K (1 - m_i) \odot \|\mathcal{F}(D_\theta(\hat{n}_i; \sigma)) - \mathcal{F}(z_i)\|^2 \right].$$

这个损失函数最小化了预测与真实值之间高频特征的差异，从而保留了更多的结构信息。我们的最终训练目标是公式(1)、公式(3)和公式(5)的加权和，其中1和2是用于平衡优化的权衡权重。

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{diffusion}} + \lambda_1 \mathcal{L}_{\text{dynamics}} + \lambda_2 \mathcal{L}_{\text{structure}}.$$

3.2阶段二：学习多功能的动作可控性

多功能动作的统一条件化。为了最大化使用灵活性，驾驶世界模型应能够利用多种具有不同特征的动作格式。例如，可以使用世界模型来评估高级策略[127]，或执行低级操作[102]。然而，现有方法仅支持有限的动作控制[54, 61, 90, 125, 127]，限制了其灵活性和适用性。因此，我们在Vista中引入了多样的动作模式集：(1) 角度和速度代表最精细的动作控制。我们将角度归一化到 $[-1, 1]$ ，并以公里/小时表示速度。不同时间戳的信号按顺序连接。(2) 轨迹是一系列以自我坐标系表示的二维位移。它广泛用作规划算法的输出[12, 21, 58, 62, 63]。我们以米为单位表示轨迹，并将其展平为序列。(3) 命令是最高级的意图。在不失一般性的前提下，我们定义了四个命令，即前进、右转、左转和停止，这些命令实现为类别索引。(4) 目标点是短期自我目标投影到初始帧上的二维坐标，作为交互接口[74]。坐标按图像尺寸进行归一化。

请注意，这些动作是异质的，不能互换使用。在将所有这些动作转换为数值序列后，我们将它们编码为傅里叶嵌入的统一连接[114, 116]（见图3）。这些嵌入可以通过学习额外的投影来扩展UNet[5]中交叉注意力层的输入维度，从而被联合摄入。新的

投影被初始化为零，以便从预训练状态逐步学习。我们通过实验发现，通过交叉注意力层引入动作条件比其他方法（如加性嵌入[128, 136]）能更快地收敛并实现更强的可控性。

高效学习。我们在第一阶段训练后学习动作可控性。由于总迭代次数对扩散训练至关重要[5, 22, 32, 99]，我们将动作控制学习分为两个阶段。在第一阶段，我们在低分辨率（ 320×576 ）下训练模型，相比原始分辨率（ 576×1024 ），训练吞吐量提高了3.5倍。这一阶段占据了大部分训练迭代。然后，我们在所需分辨率（ 576×1024 ）下进行短期微调，使得学到的可控性能够有效适应高分辨率预测。

然而，直接在较低分辨率下调整UNet[5]可能会削弱其高保真预测能力。相反，冻结所有UNet权重并单独训练新的投影会导致质量下降（见附录D），这表明有必要使UNet具有适应性。为此，我们冻结了预训练的UNet，并为每个注意力层引入了参数高效的LoRA适配器[55]。训练后，低秩矩阵可以无缝地与冻结权重结合，不会引入额外的推理延迟。因此，在低分辨率下训练时，预训练权重保持不变，避免了预训练高保真预测能力的退化。

由于在开放世界场景中无法获取相机的参数和车辆的参数，因此在推理时似乎不可能同时获得多个等效的动作条件。此外，要涵盖所有可能的动作条件组合将需要极其昂贵的训练成本。因此，与通常在训练期间激活所有条件的做法不同，我们通过在每次训练样本中仅启用其中一种动作格式来强制执行不同动作格式的独立性。剩余的动作条件将被填充为零，作为无条件的输入。如附录D所示，这种简单的约束防止了在动作组合上的训练成本浪费，并在相同的训练步骤内最大化每种单独动作模式的学习效率。

协同训练。需要注意的是，上述动作条件在OpenDV-YouTube [136]中不可用。另一方面，nuScenes [10]有足够的标注来推导这些条件。为了保持泛化性并同时学习可控性，我们引入了一种协同训练策略，利用来自两个数据集的样本，并将OpenDV-YouTube的动作条件设置为零。动作控制学习阶段采用与公式(6)相同的损失。通过从两个互补数据集中学习，Vista获得了可泛化到新数据集的多功能可控性。

3.3 可泛化的奖励函数

世界模型的一个应用是通过引入奖励模块来评估行动[40, 42, 43, 76]。Drive-WM[127]通过使用外部检测器[82, 84]来建立奖励。然而，这些检测器是在特定数据集[10]上开发的，这可能在任意场景中成为奖励估计的瓶颈。另一方面，Vista已经吸收了数百万条人类驾驶记录，展现出跨场景的强大泛化能力。基于分布外条件将导致生成多样性增加的观察结果[28, 60]，我们利用Vista自身的预测不确定性作为奖励的来源。与Drive-WM不同，我们的奖励函数无缝继承了Vista的泛化能力，而无需依赖外部模型。具体来说，我们通过条件方差来估计不确定性。为了可靠的近似，我们从相同条件帧c和动作a的随机采样噪声中进行M轮去噪。我们的奖励函数R(c, a)定义为平均负条件方差的指数：

$$\begin{aligned}\mu' &= \frac{1}{M} \sum_m D_{\theta}^{(m)}(\hat{n}; c, a), \\ R(c, a) &= \exp \left[\text{avg} \left(-\frac{1}{M-1} \sum_m (D_{\theta}^{(m)}(\hat{n}; c, a) - \mu')^2 \right) \right],\end{aligned}$$

其中 $\text{avg}()$ 对视频片段内的所有潜在值取平均。基于这种表述，不确定性较大的不利动作将导致较低的奖励。与常用的评估协议（例如 L2 误差）相比，我们的奖励函数可以在不参考地面真实动作的情况下评估动作。请注意，为了定义的简洁性，我们没有对估计的奖励进行归一化处理，但通过使用一个因子对估计的奖励进行重新缩放，可以简单地放大相对对比度。



表2：nuScenes验证集上的预测保真度比较。Vista取得了令人鼓舞的结果，显著超越了最先进的驾驶世界模型。

Metric	DriveGAN 102	DriveDreamer 125	WoVoGen 90	Drive-WM 127	GenAD 136	Vista (Ours)
FID	73.4	52.6	27.6	15.8	15.4	6.9
FVD	502.3	452.0	417.7	122.7	184.0	89.4



图5：不同模型在相同条件框架下预测的驾驶未来。我们将Vista与公开可用的视频生成模型（使用其默认配置）进行对比。尽管之前的模型产生了错位和损坏的结果，Vista并未受这些缺陷的影响。



图6：[顶部]：长时程预测。Vista能够预测15秒的高分辨率未来场景，且质量无显著下降，覆盖了较长的驾驶距离。蓝色线条的长度表示先前工作中展示的最长预测时长。[底部]：SVD的长时程扩展结果。SVD无法像Vista那样自回归地生成一致的高保真视频。

4 Experiments

在本节中，我们首先在第4.1节展示了Vista在泛化性和保真度方面的优势。然后，我们在第4.2节展示了动作控制的影响。我们还通过第4.3节验证了所提出奖励函数的有效性。最后，我们在第4.4节对关键设计进行了消融研究。有关更多实现细节和实验结果，请参阅附录C和附录D。

4.1 泛化性与保真度的比较

自动评估。由于没有公开可用的驾驶世界模型，我们通过在nuScenes上的定量结果来比较这些方法。我们从验证集中筛选出5369个有效样本进行FID [47]和FVD [115]评估。对于FID评估，我们将预测帧裁剪并调整到 256×448 的分辨率。对于FVD评估，我们使用每个视频片段中的所有25帧，并按照LVDM [46]的方法将其下采样至 224×224 。表2报告了所有方法的结果。在两种指标上，Vista都显著优于之前的驾驶世界模型。

人类评估。为了分析Vista在不同数据集上的泛化能力，我们将其与三个基于网络规模数据训练的著名通用视频生成器[5, 133, 144]进行比较（见图5）。众所周知，像FVD[115]这样的自动评估指标无法全面揭示感知的差异。

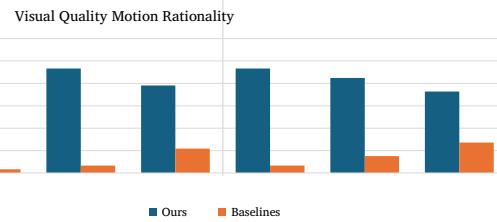


图7：人类评估结果。数值表示一个模型优于另一个模型的百分比。Vista在两项指标中均优于现有工作。

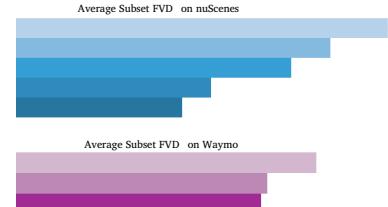


图8：行动控制的效能。应用行动控制将产生更接近真实数据的预测。



图9：多功能的动作控制能力。Vista能够在多种场景下对多模态动作条件做出响应，并预测相应的后果。更多结果见附录E。

quality [3, 6, 32, 130, 136]，更不用说真实世界的动态变化。因此，我们选择进行人工评估以进行更忠实的分析。根据最近的进展 [3, 5, 6, 15, 16, 32, 122, 126]，我们采用了Two-Alternative Forced Choice协议。具体来说，参与者会看到一对并排的视频，并被要求在两个正交的方面选择他们认为更好的视频：视觉质量和运动合理性。为了避免潜在的偏见，我们将每个视频裁剪为固定的宽高比，下采样到相同的分辨率，并在Vista生成的视频比其他视频更长时修剪多余的帧。我们只输入一个条件帧以与其他模型对齐。为了确保场景的多样性，我们从四个代表性数据集中均匀地组装了60个场景：OpenDV-YouTube-val [136]、nuScenes [10]、Waymo [112] 和 CODA [79]。这些数据集共同展示了真实世界驾驶的复杂性和多样性，例如，OpenDV-YouTube-val包括地理围栏区域，Waymo提供了与我们训练数据不同的独特领域，而CODA包含极具挑战性的边缘案例。我们从33名参与者中收集了总共2640个答案。如图7所示，Vista在两个方面均优于所有基线，展示了其对驾驶动态的深刻理解。此外，与其他仅适用于短期生成的模型不同，Vista可以容纳更多的动态先验，并生成连贯的长时程滚动预测，如图6所示。

4.2 行动可控性结果

量化结果。为了评估动作控制的影响，我们将nuScenes和未见过的Waymo数据集的验证集根据我们的命令类别分为四个子集。然后，我们使用不同模态的地面真实动作生成预测。在每个子集中测量FVD分数，然后取平均值。FVD分数越低，表示预测与地面真实视频的分布越接近，表明预测的行为与特定类型的行为更相似。图8显示，我们的动作控制能够有效地模拟相应的运动。

我们还引入了一个名为轨迹差异的新指标来评估控制一致性。遵循GenAD [136]，我们训练了一个逆动力学模型 (IDM)，该模型从视频片段中估计相应的轨迹。图13展示了IDM的示意图。然后，我们将Vista的预测结果输入到IDM中，并计算真实轨迹与估计轨迹之间的L2差异。



图10：[左]：在Waymo上不同L2误差下的平均奖励。[右]：案例研究。我们的奖励的相对对比能够适当评估L2误差无法判断的动作。



图11：动态先验的影响。注入更多的动态先验能产生与真实情况更一致的未来运动，例如白色车辆和左侧广告牌的运动。

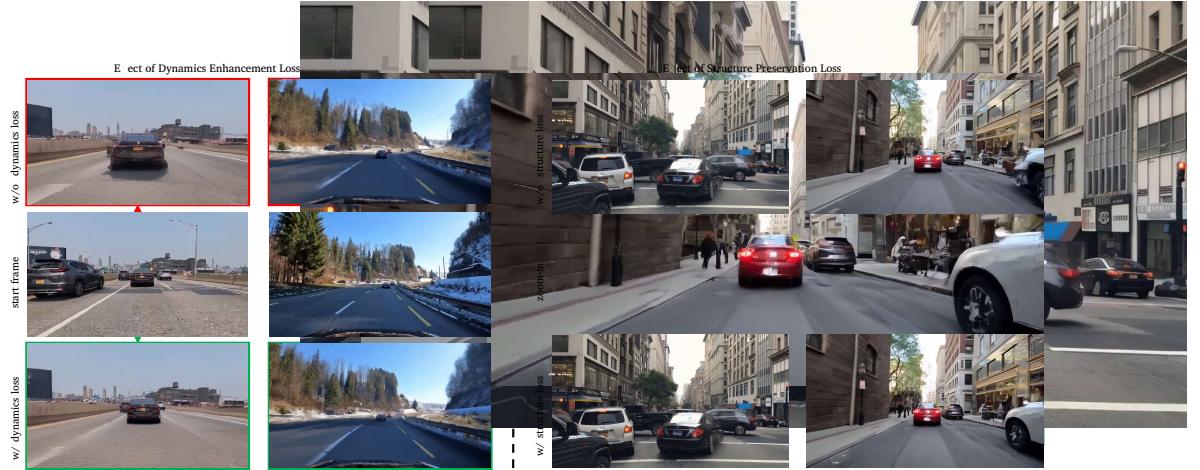


图12：[左]：动力学增强损失的效果。通过动力学增强损失监督的模型生成了更真实的动态效果。在第一个例子中，前车没有保持静止，而是正常地向前移动。在第二个例子中，当自行车向右转向时，树木自然地向左移动，符合现实世界中的几何规则。[右]：结构保持损失的效果。所提出的损失使得物体在移动时轮廓更加清晰。

差异是在2秒的时间跨度内测量的。轨迹差异越小，Vista 表现出的控制一致性越强。我们在nuScenes 和 Waymo 上进行了实验。对于每个数据集，我们收集了一个包含537个样本的子集。如表3所示，Vista 能够被不同模态的动作有效控制，从而产生与真实情况更加一致的运动。

定性结果。图9展示了我们模型的多功能动作控制能力。Vista可以通过多模态动作进行有效控制，即使在训练领域之外的未见场景中也能表现出色。在附录E中，我们还展示了Vista使用异常动作进行反事实推理的能力。

4.3 奖励建模的结果

为了验证我们奖励函数的有效性，我们将真实轨迹抖动成一系列次优轨迹。具体来说，我们计算了nuScenes训练集中每个航点的标准差作为先验分布。这些先验分布被联合缩放，以采样具有不同L2误差的扰动。然后将这些扰动作为偏移量添加到真实轨迹中。为了确保采样轨迹的合理性，我们采用了显式相关策略[35, 95]来规范偏移采样，并递归地采样新轨迹，直到它们的偏移量在趋势上保持一致。为了展示我们奖励函数的通用性，我们在Waymo[112]上进行了奖励估计，这是在训练中未见过的。通过从Waymo验证集中的每个命令类别均匀采样，总共生成了1500个案例。我们在图10中比较了具有不同L2误差的轨迹的平均奖励。当偏离真实轨迹增加时，我们的奖励减少，这突显了我们的方法作为可行奖励函数的潜力。它还具有改善当前规划评估协议中不合理性的潜力[18, 83, 141]，如图10所示的L2误差。更多关于奖励的深入分析，包括对超参数的敏感性和其他动作的奖励，在附录D中提供。

表3：不同动作条件和动态先验的影响。通过应用动作条件和动态先验，Vista能够预测与真实情况更为一致的运动。

Dataset	Condition	Average Trajectory Difference		
		with 1 prior	with 2 priors	with 3 priors
nuScenes	GT video	0.379	0.379	0.379
	action-free	3.785	2.597	1.820
	+ goal point	2.869	2.192	1.585
	+ command	3.129	2.403	1.593
	+ angle & speed	1.562	1.123	0.832
	+ trajectory	1.559	1.148	0.835
Waymo	GT video	0.893	0.893	0.893
	action-free	3.646	2.901	2.052
	+ command	3.160	2.561	1.902
	+ trajectory	1.187	1.147	1.140

4.4 消融研究

动态先验。图11展示了使用不同顺序动态先验的结果。先验的顺序对应于条件帧的数量。结果显示，动态先验在长时程推演中起着关键作用，其中与历史帧的一致性至关重要。

为进一步证明动态先验的有效性，我们在表3中进行了定量评估。具体来说，我们使用IDMin Sec. 4.2来推断预测视频的轨迹，并采用不同阶数的动态先验。轨迹差异的逐渐减小表明，引入更多先验可以有效提高预测与真实值之间的一致性。

辅助监督。为了验证第3.1节中提出的两种损失的有效性，我们设计了两个额外的变体，通过分别从包含两种损失的变体中去除每种损失。我们通过图12定性地比较它们的效果，这证实了动态增强损失可以促进对现实世界动态的学习，而结构保持损失可以强化对结构细节的预测。

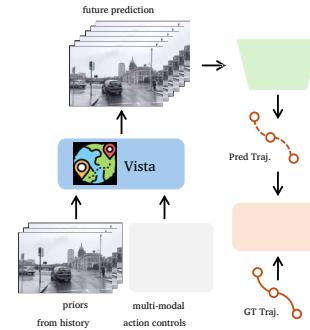


图13：表3中IDM实验的示意图。

5 Conclusion

本文介绍Vista，这是一种具有增强的保真度和可控性的可推广驾驶世界模型。基于我们的系统性研究，Vista能够以高时空分辨率预测真实且连续的未来。它还具备多样的动作可控性，这种可控性可以推广到未见过的场景。此外，它可以被形式化为奖励函数来评估动作。我们希望Vista将引领对开发可推广自主系统更广泛的兴趣。

局限性与未来工作。作为一个早期的尝试，Vista在计算效率、质量维护和训练规模方面仍存在一些限制。我们未来的工作将研究将我们的方法应用于可扩展架构 [54, 97]。更多讨论包含在附录 A 中。

致谢

本工作得到国家重点研发计划（2022ZD0160104）、国家自然科学基金（62206172）以及上海市科学技术委员会（23YF1462000）的支持。本工作还部分得到了德国联邦教育和研究部（Tübingen AI Center，项目编号：01IS18039A）、德国研究基金会（SFB 1233, TP 17, 项目编号：276693517）以及卓越集群（编号2064/1 - 项目编号：390727645）的支持。我们感谢国际马克斯·普朗克智能系统研究所（IMPRS-IS）对Kashyap Chitta的支持。我们也感谢Zetong Yang、Chonghao Sima、Linyan Huang以及OpenDriveLab的其他成员提供的宝贵反馈。我们对所有参与人类评估的匿名参与者表示衷心的感谢。

参考文献

- [1] Anurag Ajay, Seungwook Han, Yilun Du, Shuang Li, Abhi Gupta, Tommi Jaakkola, Josh Tenenbaum, Leslie Kaelbling, Akash Srivastava, 和 Pulkit Agrawal. 用于分层规划的组合基础模型。发表于 NeurIPS, 2023. 19
- [2] Bowen Baker, Ilge Akkaya, Peter Zhokov, Joost Huizinga, Jie Tang, Adrien Eco et, Brandon Houghton, Raul Sampedro, 和 Je Clune. 视频预训练 (VPT) : 通过观看未标记的在线视频学习行动。发表于 NeurIPS, 2022. 19
- [3] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, 和 Inbar Mosseri. Lumiere : 用于视频生成的时空扩散模型。arXiv 预印本 arXiv:2401.12945, 2024. 4, 8
- [4] Kevin Black, Mitsuhiro Nakamoto, Pranav Atreya, Homer Walke, Chelsea Finn, Aviral Kumar, 和 Sergey Levine. 使用预训练的图像编辑扩散模型进行零样本机器人操作。发表于 ICLR, 2024. 19
- [5] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, 和 Robin Rombach. 稳定视频扩散 : 将潜在视频扩散模型扩展到大尺度数据集。arXiv 预印本 arXiv:2311.15127, 2023. 2, 3, 5, 6, 7, 8, 21, 22, 23, 24, 26
- [6] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, 和 Karsten Kreis. 对齐您的潜在变量 : 使用潜在扩散模型进行高分辨率视频合成。发表于 CVPR, 2023. 8, 20, 21
- [7] Daniel Bogdall, Yitian Yang, 和 J Marius Zöllner. MUVO : 用于自动驾驶的多模态生成世界模型 , 采用几何表示。arXiv 预印本 arXiv:2311.11762, 2023. 21
- [8] Tim Brooks, Aleksander Holynski, 和 Alexei A Efros. InstructPix2Pix : 学习遵循图像编辑指令。发表于 CVPR, 2023. 20
- [9] Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Je Clune, Nando de Freitas, Satinder Singh, 和 Tim Rocktäschel. Genie : 生成式交互环境。arXiv 预印本 arXiv:2402.15391, 2024. 19, 20, 21
- [10] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yuxin Pan, Giancarlo Baldan, 和 Oscar Beijbom. nuScenes : 用于自动驾驶的多模态数据集。发表于 CVPR, 2020. 6, 8, 20, 21, 23, 26
- [11] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wol , Alex Lang, Luke Fletcher, Oscar Beijbom, 和 Sammy Omari. nuPlan : 基于闭环机器学习的自动驾驶规划基准。发表于 CVPR Workshops, 2021. 20
- [12] Sergio Casas, Abbas Sadat, 和 Raquel Urtasun. MP3 : 统一模型用于映射、感知、预测和规划。发表于 CVPR, 2021. 1, 5, 21
- [13] Jun Cen, Chenfei Wu, Xiao Liu, Shengming Yin, Yixuan Pei, Jinglong Yang, Qifeng Chen, Nan Duan, 和 Jianguo Zhang. 左右脑并用 : 视觉与语言规划。arXiv 预印本 arXiv:2402.10534, 2024. 19
- [14] Dian Chen, Brady Zhou, Vladlen Koltun, 和 Philipp Krähenbühl. 通过作弊学习。发表于 CoRL, 2019. 1, 21
- [15] 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. 8, 21
- [16] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, 和 Ying Shan. VideoCrafter2 : 克服数据限制的高质量视频扩散模型。发表于 CVPR, 2024. 8
- [17] Kai Chen, Enze Xie, Zhe Chen, Yibo Wang, Lanqing Hong, Zhenguo Li, 和 Dit-Yan Yeung. GeoDi usion : 文本提示的几何控制用于目标检测数据生成。发表于 ICLR, 2024. 4
- [18] 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. 1, 9, 19
- [19] Shoufa Chen, Mengmeng Xu, Jiawei Ren, Yuren Cong, Sen He, Yanping Xie, Animesh Sinha, Ping Luo, Tao Xiang, 和 Juan-Manuel Perez-Rua. GenTron : 深入探索用于图像和视频生成的扩散变换器。发表于 CVPR, 2024. 20

- [20] Kashyap Chitta, Daniel Dauner, and Andreas Geiger. SLEDGE: 使用生成模型合成驾驶代理的仿真环境。arXiv预印本arXiv:2403.17933, 2024. 21
- [21] Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Trans-Fuser: 基于Transformer的传感器融合用于自动驾驶的模仿学习。IEEE TPAMI, 2023. 1, 5, 21
- [22] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, Matthew Yu, Abhishek Kadian, Filip Radenovic, Dhruv Mahajan, Kunpeng Li, Yue Zhao, Vladan Petrovic, Mitesh Kumar Singh, Simran Motwani, Yi Wen, Yiwen Song, Roshan Sumbaly, Vignesh Ramanathan, Zijian He, Peter Vajda, and Devi Parikh. Emu: 通过海量数据中的照片级细节增强图像生成模型。arXiv预印本arXiv:2309.15807, 2023. 6
- [23] Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, et al. NAVSIM: 数据驱动的非反应性自动驾驶车辆仿真与基准测试。arXiv预印本arXiv:2406.15349, 2024. 20
- [24] Prafulla Dhariwal and Alexander Nichol. 扩散模型在图像合成上击败GAN。在NeurIPS, 2021. 21
- [25] Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. 通过文本引导的视频生成学习通用策略。在NeurIPS, 2023. 19
- [26] Yilun Du, Sherry Yang, Pete Florence, Fei Xia, Ayzaan Wahid, brian ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B. Tenenbaum, Leslie Pack Kaelbling, Andy Zeng, and Jonathan Tompson. 视频语言规划。在ICLR, 2024. 19, 21
- [27] Frederik Ebert, Chelsea Finn, Sudeep Dasari, Annie Xie, Alex Lee, and Sergey Levine. 视觉预见：基于视觉的机器人控制模型驱动的深度强化学习。arXiv预印本arXiv:1812.00568, 2018. 19, 21
- [28] Alejandro Escontrela, Ademi Adeniji, Wilson Yan, Ajay Jain, Xue Bin Peng, Ken Goldberg, Youngwoon Lee, Danijar Hafner, and Pieter Abbeel. 视频预测模型作为强化学习的奖励。在NeurIPS, 2023. 6, 19
- [29] Patrick Esser, Robin Rombach, and Bjorn Ommer. 驯服Transformer用于高分辨率图像合成。在CVPR, 2021. 21
- [30] Chelsea Finn and Sergey Levine. 深度视觉预见用于规划机器人运动。在ICRA, 2017. 19, 21
- [31] Zeyu Gao, Yao Mu, Ruoyan Shen, Chen Chen, Yangang Ren, Jianyu Chen, Shengbo Eben Li, Ping Luo, and Yanfeng Lu. 通过语义掩码世界模型增强城市自动驾驶的样本效率和鲁棒性。在NeurIPS研讨会, 2022. 21
- [32] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu Video: 通过显式图像条件化分解文本到视频生成。arXiv预印本arXiv:2311.10709, 2023. 4, 6, 8, 20, 21
- [33] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 生成对抗网络。在NeurIPS, 2014. 21
- [34] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. AnimateDi : 无需特定调整即可动画化个性化文本到图像扩散模型。在ICLR, 2024. 21
- [35] Agrim Gupta, Stephen Tian, Yunzhi Zhang, Jiajun Wu, Roberto Martín-Martín, and Li Fei-Fei. MaskViT: 用于视频预测的掩码视觉预训练。在ICLR, 2023. 9, 19, 21, 23
- [36] Tarun Gupta, Wenbo Gong, Chao Ma, Nick Pawlowski, Agrin Hilmkil, Meyer Scetbon, Ade Famoti, Ashley Juan Llorens, Jianfeng Gao, Stefan Bauer, Bernhard Krägic, Danica an Schölkopf, and Cheng Zhang. 因果关系在具身AI的基础世界模型中的关键作用。arXiv预印本arXiv:2402.06665, 2024. 25
- [37] Wes Gurnee and Max Tegmark. 语言模型表示空间和时间。在ICLR, 2024. 21
- [38] Nicholas Guttenberg and CrossLabs. 扩散与偏移噪声, 2023. 22
- [39] David Ha and Jürgen Schmidhuber. 循环世界模型促进策略进化。在NeurIPS, 2018. 21
- [40] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. 梦想控制：通过潜在想象学习行为。arXiv预印本arXiv:1912.01603, 2019. 6, 21
- [41] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. 从像素中学习潜在动态用于规划。在ICML, 2019. 21
- [42] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. 通过离散世界模型掌握Atari游戏。在ICLR, 2021. 6, 21

- [43] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, 和 Timothy Lillicrap. 通过世界模型掌握多样领域。arXiv 预印本 arXiv:2301.04104, 2023. 6, 21
- [44] Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, 和 Zhiting Hu. 使用语言模型进行推理即规划与世界模型。在 EMNLP, 2023. 21
- [45] Haoran He, Chenjia Bai, Ling Pan, Weinan Zhang, Bin Zhao, 和 Xuelong Li. 通过离散扩散的大规模无动作视频预训练用于高效策略学习。arXiv 预印本 arXiv:2402.14407, 2024. 21
- [46] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, 和 Qifeng Chen. 用于高保真长视频生成的潜在视频扩散模型。arXiv 预印本 arXiv:2211.13221, 2022. 7, 21
- [47] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, 和 Sepp Hochreiter. 通过双时间尺度更新规则训练的 GANs 收敛于局部纳什均衡。在 NeurIPS, 2017. 7
- [48] Jonathan Ho, Ajay Jain, 和 Pieter Abbeel. 去噪扩散概率模型。在 NeurIPS, 2020. 21
- [49] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, 和 Tim Salimans. 用于高保真图像生成的级联扩散模型。JMLR, 2022. 3, 20, 22
- [50] Jonathan Ho 和 Tim Salimans. 无分类器扩散指导。arXiv 预印本 arXiv:2207.12598, 2022. 22
- [51] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, 和 David J Fleet. 视频扩散模型。arXiv 预印本 arXiv:2204.03458, 2022. 21
- [52] Anthony Hu, Gianluca Corrado, Nicolas Gribble, Zachary Murez, Corina Gurau, Hudson Yeo, Alex Kendall, Roberto Cipolla, 和 Jamie Shotton. 基于模型的城市驾驶模仿学习。在 NeurIPS, 2022. 21
- [53] Anthony Hu, Zak Murez, Nikhil Mohan, Sofía Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, 和 Alex Kendall. FIERY: 从周围单目相机进行鸟瞰图的未来实例预测。在 ICCV, 2021. 19, 21
- [54] 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. 1, 2, 3, 5, 10, 20, 21
- [55] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, 和 Weizhu Chen. LoRA: 大型语言模型的低秩适应。在 ICLR, 2022. 6, 22
- [56] 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. 1, 21, 22
- [57] Yafei Hu, Quanting Xie, Vidhi Jain, Jonathan Francis, Jay Patrikar, Nikhil Keetha, Seungchan Kim, Yaqi Xie, Tianyi Zhang, Shibo Zhao, Yu Quan Chong, Chen Wang, Katia Sycara, Matthew Johnson-Roberson, Dhruv Batra, Xiaolong Wang, Sebastian Scherer, Zsolt Kira, Fei Xia, 和 Yonatan Bisk. 通过基础模型实现通用机器人：综述与元分析。arXiv 预印本 arXiv:2312.08782, 2023. 1, 20
- [58] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhui 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. 1, 5, 21, 22
- [59] Zhiting Hu 和 Tianmin Shu. 语言模型、代理模型和世界模型：机器推理与规划的 LAW。arXiv 预印本 arXiv:2312.05230, 2023. 21
- [60] Tao Huang, Guangqi Jiang, Yanjie Ze, 和 Huazhe Xu. 扩散奖励：通过条件视频扩散学习奖励。arXiv 预印本 arXiv:2312.14134, 2023. 6, 19
- [61] Fan Jia, Weixin Mao, Yingfei Liu, Yucheng Zhao, Yuqing Wen, Chi Zhang, Xiangyu Zhang, 和 Tiancai Wang. ADriver-I: 用于自动驾驶的通用世界模型。arXiv 预印本 arXiv:2311.13549, 2023. 1, 2, 5, 21
- [62] Xiaosong Jia, Yulu Gao, Li Chen, Junchi Yan, Patrick Langechuan Liu, 和 Hongyang Li. DriveAdapter: 打破端到端自动驾驶中感知与规划的耦合障碍。在 ICCV, 2023. 5
- [63] Xiaosong Jia, Penghao Wu, Li Chen, Jiangwei Xie, Conghui He, Junchi Yan, 和 Hongyang Li. 在驾驶前思考两次：通过可扩展解码器实现端到端自动驾驶。在 CVPR, 2023. 5
- [64] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, 和 Xinggang Wang. VAD: 用于高效自动驾驶的矢量化场景表示。在 ICCV, 2023. 1, 21, 22

- [65] Michael I Jordan 和 David E Rumelhart。前向模型：通过远端教师的监督学习。认知科学，1992年。19
- [66] Tero Karras, Miika Aittala, Timo Aila, 和 Samuli Laine。阐明基于扩散的生成模型的设计空间。在 NeurIPS, 2022。2, 22
- [67] Tarasha Khurana, Peiyun Hu, David Held, 和 Deva Ramanan。点云预测作为4D占用预测的代理。在 CVPR, 2023。21
- [68] 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。2, 3, 21
- [69] Seung Wook Kim, Yuhao Zhou, Jonah Philion, Antonio Torralba, 和 Sanja Fidler。学习使用 GameGAN 模拟动态环境。在 CVPR, 2020。21
- [70] Diederik P Kingma 和 Max Welling。自动编码变分贝叶斯。arXiv 预印本 arXiv:1312.6114, 2013。21
- [71] Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, 和 Joshua B Tenenbaum。通过密集对应关系从无动作视频中学习行动。在 ICLR, 2024。19
- [72] Jing Yu Koh, Honglak Lee, Yinfei Yang, Jason Baldridge, 和 Peter Anderson。Pathdreamer：用于室内导航的世界模型。在 ICCV, 2021。21
- [73] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vighnesh Birodkar, Yong Cheng, Ming-Chang Chiu, Josh Dillon, Irfan Essa, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, David Ross, Grant Schindler, Mikhail Sirotenko, Kihyuk Sohn, Krishna Somandepalli, Huisheng Wang, Jimmy Yan, Ming-Hsuan Yang, Xuan Yang, Bryan Seybold, 和 Lu Jiang。VideoPoet：用于零样本视频生成的大型语言模型。arXiv 预印本 arXiv:2312.14125, 2023。4
- [74] Hanyang Kong, Dongze Lian, Michael Bi Mi, 和 Xinchao Wang。DreamDrone。arXiv 预印本 arXiv:2312.08746, 2023。5, 21
- [75] Lei Lai, Zhongkai Shangguan, Jimuyang Zhang, 和 Eshed Ohn-Bar。XVO：通过跨模态自训练实现广义视觉里程计。在 ICCV, 2023。19
- [76] Yann LeCun。通向自主机器智能的路径。开放评论, 62, 2022。1, 6, 19, 20, 21
- [77] Hongyang Li, Yang Li, Huijie Wang, Jia Zeng, Huilin Xu, Pinlong Cai, Li Chen, Junchi Yan, Feng Xu, Lu Xiong, Jingdong Wang, Futang Zhu, Chunjing Xu, Tiancai Wang, Fei Xia, Beipeng Mu, Zhihui Peng, Dahua Lin, 和 Yu Qiao。自动驾驶中的开源数据生态系统：现状与未来。arXiv 预印本 arXiv:2312.03408, 2023。21
- [78] Hongyang Li, Chonghao Sima, Jifeng Dai, Wenhui Wang, Lewei Lu, Huijie Wang, Jia Zeng, Zhiqi Li, Jiazhui Yang, Hanming Deng, Hao Tian, Enze Xie, Jiangwei Xie, Li Chen, Tianyu Li, Yang Li, Yulu Gao, Xiaosong Jia, Si Liu, Jianping Shi, Dahua Lin, 和 Yu Qiao。深入探讨鸟瞰感知中的魔鬼：回顾、评估与方法。IEEE TPAMI, 2023。21
- [79] Kaican Li, Kai Chen, Haoyu Wang, Lanqing Hong, Chaoqiang Ye, Jianhua Han, Yukuai Chen, Wei Zhang, Chunjing Xu, Dit-Yan Yeung, Xiaodan Liang, Zhenguo Li, 和 Hang Xu。CODA：用于自动驾驶中物体检测的真实道路角落案例数据集。在 ECCV, 2022。8, 26
- [80] Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, 和 Martin Wattenberg。涌现的世界表示：探索在合成任务上训练的序列模型。在 ICLR, 2023。21
- [81] Qifeng Li, Xiaosong Jia, Shaobo Wang, 和 Junchi Yan。Think2Drive：通过在潜在世界模型中思考实现高效的强化学习，用于准现实自动驾驶（在 CARLA-v2 中）。arXiv 预印本 arXiv:2402.16720, 2024。21
- [82] Zhiqi Li, Wenhui Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, 和 Jifeng Dai。BEVFormer：通过时空变换器从多摄像头图像中学习鸟瞰表示。在 ECCV, 2022。6, 19
- [83] Zhiqi Li, Zhiding Yu, Shiyi Lan, Jiahua Li, Jan Kautz, Tong Lu, 和 Jose M Alvarez。自我状态是否足以实现开环端到端自动驾驶？在 CVPR, 2024。1, 9, 19, 21
- [84] Bencheng Liao, Shaoyu Chen, Xinggang Wang, Tianheng Cheng, Qian Zhang, Wenyu Liu, 和 Chang Huang。MapTR：用于在线矢量化高清地图构建的结构化建模与学习。在 ICLR, 2023。6, 19
- [85] Jessy Lin, Yuqing Du, Olivia Watkins, Danijar Hafner, Pieter Abbeel, Dan Klein, 和 Anca Dragan。通过语言学习建模世界。arXiv 预印本 arXiv:2308.01399, 2023。21
- [86] Hao Liu, Wilson Yan, Matei Zaharia, 和 Pieter Abbeel。使用 RingAttention 在百万长度视频和语言上的世界模型。arXiv 预印本 arXiv:2402.08268, 2024。21

- [87] Ilya Loshchilov 和 Frank Hutter。解耦权重衰减正则化。arXiv 预印本 arXiv:1711.05101 , 2017。22
- [88] William Lotter, Gabriel Kreiman, 和 David Cox。深度预测编码网络用于视频预测和无监督学习。在 ICLR , 2017。21
- [89] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, 和 Jun Zhu。DPM-Solver : 一种用于扩散概率模型采样的快速 ODE 求解器, 仅需约 10 步。在 NeurIPS , 2022。20
- [90] Jiachen Lu, Ze Huang, Jiahui Zhang, Zeyu Yang, 和 Li Zhang。WoVoGen : 用于可控多摄像机驾驶场景生成的世界体积感知扩散。arXiv 预印本 arXiv:2312.02934 , 2023。1, 2, 5, 7, 21
- [91] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, 和 Hang Zhao。潜在一致性模型 : 用少量步骤推理生成高分辨率图像。arXiv 预印本 arXiv:2310.04378 , 2023。20
- [92] Russell Mendonca, Shikhar Bahl, 和 Deepak Pathak。从人类视频中构建结构化世界模型。在 RSS , 2023。21
- [93] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, 和 Tim Salimans。关于引导扩散模型的蒸馏。在 CVPR , 2023。20
- [94] Vincent Micheli, Eloi Alonso, 和 François Fleuret。Transformer 是样本高效的世界模型。在 ICLR , 2023。21
- [95] Anusha Nagabandi, Kurt Konolige, Sergey Levine, 和 Vikash Kumar。用于学习灵巧操作的深度动力学模型。在 CoRL , 2020。9, 23
- [96] Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Richard L Lewis, 和 Satinder Singh。使用 Atari 游戏中的深度网络进行动作条件视频预测。在 NeurIPS , 2015。21
- [97] William Peebles 和 Saining Xie。可扩展的扩散模型与 Transformer。在 ICCV , 2023。10, 20
- [98] AJ Piergiovanni, Alan Wu, 和 Michael S Ryoo。通过梦境学习现实世界的机器人策略。在 IROS , 2019。21
- [99] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, 和 Robin Rombach。SDXL : 改进潜在扩散模型用于高分辨率图像合成。在 ICLR , 2024。6, 20
- [100] 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。21
- [101] Tim Salimans 和 Jonathan Ho。扩散模型的快速采样渐进蒸馏。在 ICLR , 2023。20
- [102] Eder Santana 和 George Hotz。学习驾驶模拟器。arXiv 预印本 arXiv:1608.01230 , 2016。2, 5, 7, 21
- [103] Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, 和 Robin Rombach。使用潜在对抗扩散蒸馏进行快速高分辨率图像合成。arXiv 预印本 arXiv:2403.12015 , 2024。20
- [104] Ingmar Schubert, Jingwei Zhang, Jake Bruce, Sarah Bechtel, Emilio Parisotto, Martin Riedmiller, Jost Tobias Springenberg, Arunkumar Byravan, Leonard Hasenclever, 和 Nicolas Heess。用于控制的通用动力学模型。arXiv 预印本 arXiv:2305.10912 , 2023。21
- [105] Max Schwarzer, Ankesh Anand, Rishab Goel, R Devon Hjelm, Aaron Courville, 和 Philip Bachman。使用自预测表示的数据高效强化学习。在 ICLR , 2021。21
- [106] Dhruv Shah, Ajay Sridhar, Nitish Dashora, Kyle Stachowicz, Kevin Black, Noriaki Hirose, 和 Sergey Levine。ViNT : 视觉导航的基础模型。在 CoRL , 2023。19
- [107] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, 和 Hongyang Li。DriveLM : 通过图视觉问答进行驾驶。arXiv 预印本 arXiv:2312.14150 , 2023。20, 21
- [108] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, 和 Yaniv Taigman。Make-A-Video : 无需文本-视频数据的文本到视频生成。在 ICLR , 2023。21
- [109] Jiaming Song, Chenlin Meng, 和 Stefano Ermon。去噪扩散隐式模型。在 ICLR , 2021。22
- [110] Yang Song, Prafulla Dhariwal, Mark Chen, 和 Ilya Sutskever。一致性模型。在 ICML , 2023。20
- [111] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, 和 Ben Poole。通过随机微分方程进行基于分数的生成建模。在 ICLR , 2021。2

- [112] 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. 自动驾驶感知中的可扩展性：Waymo开放数据集。在CVPR，2020年。8, 9, 20, 25, 26
- [113] Richard S Sutton. 智能决策者的通用模型探索。arXiv预印本arXiv:2202.13252，2022年。20, 21
- [114] 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域中学习高频函数。在NeurIPS，2020年。5, 21
- [115] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, 和 Sylvain Gelly. 迈向准确的视频生成模型：一个新的度量与挑战。arXiv预印本arXiv:1812.01717，2018年。7
- [116] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, 和 Illia Polosukhin. 注意力就是你所需要的。在NeurIPS，2017年。5, 21
- [117] Vikram Voleti, Alexia Jolicoeur-Martineau, 和 Chris Pal. MCVD: 用于预测、生成和插值的掩码条件视频扩散。在NeurIPS，2022年。21
- [118] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforet, Robin Rombach, 和 Varun Jampani. SV3D: 利用潜在视频扩散从单张图像进行新颖的多视图合成和3D生成。arXiv预印本arXiv:2403.12008，2024年。22
- [119] Carl Vondrick, Hamed Pirsiavash, 和 Antonio Torralba. 利用场景动态生成视频。在NeurIPS，2016年。21
- [120] Hanqing Wang, Wei Liang, Luc Van Gool, 和 Wenguan Wang. DREAMWALKER: 连续视觉语言导航的精神规划。在ICCV，2023年。21
- [121] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, 和 Shiwei Zhang. ModelScope文本到视频技术报告。arXiv预印本arXiv:2308.06571，2023年。21
- [122] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, 和 Jiaying Liu. VideoFactory: 在时空扩散中交换注意力以进行文本到视频生成。arXiv预印本arXiv:2305.10874，2023年。8
- [123] Xiang Wang, Shiwei Zhang, Hangjie Yuan, Zhiwu Qing, Biao Gong, Yingya Zhang, Yujun Shen, Changxin Gao, 和 Nong Sang. 通过无文本视频扩展文本到视频生成的方法。在CVPR，2024年。4
- [124] Xiang Wang, Shiwei Zhang, Han Zhang, Yu Liu, Yingya Zhang, Changxin Gao, 和 Nong Sang. VideoLCM: 视频潜在一致性模型。arXiv预印本arXiv:2312.09109，2023年。20
- [125] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, 和 Jiwen Lu. DriveDreamer: 迈向真实世界驱动的自动驾驶世界模型。arXiv预印本arXiv:2309.09777，2023年。1, 2, 3, 5, 7, 21
- [126] 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: 利用级联潜在扩散模型生成高质量视频。arXiv预印本arXiv:2309.15103，2023年。8, 20
- [127] Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, 和 Zhaoxiang Zhang. 驶入未来：利用世界模型进行多视角视觉预测和规划的自动驾驶。在CVPR，2024年。1, 2, 3, 5, 6, 7, 19, 20, 21
- [128] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Tianshui Chen, Menghan Xia, Ping Luo, 和 Ying Shan. MotionCtrl: 视频生成的统一且灵活的运动控制器。arXiv预印本arXiv:2312.03641，2023年。6, 20, 21
- [129] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, 等。Argoverse 2: 下一代自动驾驶感知和预测数据集。在NeurIPS数据集和基准测试，2023年。20
- [130] Jay Zhangjie Wu, Guiyan Fang, Haoning Wu, Xintao Wang, Yixiao Ge, Xiaodong Cun, David Junhao Zhang, Jia-Wei Liu, Yuchao Gu, Rui Zhao, Weisi Lin, Wynne Hsu, Ying Shan, 和 Mike Zheng Shou. 迈向更好的文本到视频生成度量。arXiv预印本arXiv:2401.07781，2024年。8
- [131] Jialong Wu, Haoyu Ma, Chaoyi Deng, 和 Mingsheng Long. 利用野外视频预训练上下文世界模型以进行强化学习。在NeurIPS，2023年。21

- [132] 吴鹏昊, 陈力, 李红阳, 贾小松, 严俊驰, 乔宇. 通过自监督几何建模进行自动驾驶的策略预训练. 在 ICLR, 2023. 4
- [133] 邢金波, 夏梦涵, 张勇, 陈浩鑫, 王新涛, 黄天进, 单颖. DynamiCrafter: 利用视频扩散先验生成开放领域动画图像. arXiv预印本arXiv:2310.12190, 2023. 7, 22
- [134] 严伟森, 张云子, Pieter Abbeel, Aravind Srinivas. VideoGPT: 使用VQ-VAE和Transformer生成视频. arXiv预印本arXiv:2104.10157, 2021. 21
- [135] 闫旭, 张海明, 蔡英杰, 郭景明, 邱伟超, 高斌, 周凯强, 赵越, 金焕, 高建涛, 李振, 江丽辉, 张伟, 张鸿波, 戴登新, 刘冰冰. 为自动驾驶打造视觉基础模型: 挑战、方法与机遇. arXiv预印本arXiv:2401.08045, 2024. 1
- [136] 杨家志, 高申元, 邱义航, 陈力, 李天宇, 戴博, Kashyap Chitta, 吴鹏昊, 曾佳, 罗平, 张军, Andreas Geiger, 乔宇, 李红阳. 自动驾驶的广义预测模型. 在CVPR, 2024. 2, 3, 4, 6, 7, 8, 20, 21, 23, 25, 26
- [137] 杨梦娇, 杜一伦, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, Pieter Abbeel. 学习交互式现实世界模拟器. 在ICLR, 2024. 19, 21
- [138] 杨雪儿, Jacob Walker, Jack Parker-Holder, 杜一伦, Jake Bruce, Andre Barreto, Pieter Abbeel, Dale Schuurmans. 视频作为现实世界决策的新语言. arXiv预印本arXiv:2402.17139, 2024. 20
- [139] 杨世元, 侯亮, 黄海滨, 马崇阳, 万鹏飞, 张迪, 陈晓东, 廖静. Direct-a-Video: 用户指导的摄像头运动和物体运动定制化视频生成. arXiv预印本arXiv:2402.03162, 2024. 21
- [140] 杨泽彤, 陈力, 孙燕娜, 李红阳. 视觉点云预测实现可扩展自动驾驶. 在CVPR, 2024. 21
- [141] 翟江天, 冯泽, 杜金豪, 毛永强, 刘江江, 谭子昌, 张一夫, 叶小青, 王井东. 重新思考nuScenes中端到端自动驾驶的开环评估. arXiv预印本arXiv:2305.10430, 2023. 9, 19
- [142] 张亚历克斯, Khanh Nguyen, Jens Tuyls, Albert Lin, Karthik Narasimhan. 语言引导的世界模型: AI控制的基于模型的方法. arXiv预印本arXiv:2402.01695, 2024. 21
- [143] 张伦俊, 熊宇文, 杨泽, Sergio Casas, 胡锐, Raquel Urtasun. 通过离散扩散学习自动驾驶的无监督世界模型. 在ICLR, 2024. 1, 21
- [144] 张世伟, 王家宇, 张英雅, 赵康, 袁航杰, 秦志武, 王翔, 赵德利, 周景仁. I2VGen-XL: 通过级联扩散模型实现高质量图像到视频合成. arXiv预印本arXiv:2311.04145, 2023. 4, 7, 20, 21, 22
- [145] 赵国胜, 王晓峰, 朱政, 陈鑫泽, 黄冠, 鲍晓毅, 王兴刚. DriveDreamer-2: 增强LLM的世界模型用于多样化的驾驶视频生成. arXiv预印本arXiv:2403.06845, 2024. 20, 21
- [146] 赵文亮, 白璐佳, 饶永明, 周杰, 鲁文. UniPC: 扩散模型快速采样的统一预测-校正框架. 在NeurIPS, 2023. 20
- [147] 郑文昭, 陈伟良, 黄元辉, 张博睿, 段月琪, 鲁文. OccWorld: 学习自动驾驶的三维占用世界模型. arXiv预印本arXiv:2311.16038, 2023. 1, 21
- [148] 郑文昭, 宋瑞琪, 郭祥达, 陈龙. GenAD: 生成式端到端自动驾驶. arXiv预印本arXiv:2402.11502, 2024. 19
- [149] 周云松, 黄林燕, 卜庆文, 曾佳, 李天宇, 邱航, 朱红子, 郭敏毅, 乔宇, 李红阳. 驾驶场景的具身理解. arXiv预印本arXiv:2403.04593, 2024. 20

附录

A 讨论19

相关工作20

B.1世界模型	20
B.2视频生成	21

C 实现细节21

C.1模型	21
C.2数据集.	21
C.3训练	22
C.4采样.	22
C.5人类评估.	22
C.6奖励估计	23
C.7消融研究	23

D 附加实验23

D.1 奖励估计的参数敏感性	23
D.2 命令的奖励估计	23
D.3 动作独立性约束	23
D.4 三角形引导方案	24
D.5 LoRA 适应	24
D.6 动作控制一致性	24
D.7 使用 GenAD 进行人类评估	25

E 附加的可视化内容25

E.1 泛化能力.	25
E.2 长期预测.	25
E.3 动作可控性.	25
E.4 反事实推理能力.	25
E.5 人类评估案例.	26

资产许可26

A Discussions

为了帮助全面理解这项工作，我们讨论了可能会提出的直观问题。

Q1. 为什么至少需要位置、速度和加速度才能预测连贯的未来？

位置确保预测的未来与当前状态连续衔接。速度表现物体的移动状态，例如，它们是向左转还是向右转。加速度代表速度随时间的变化，例如，周围环境是移动得更快还是更慢。如果不利用加速度作为线索，一辆超车的车辆可能会在下一个自回归预测步骤中突然被超越。这三个先验提供了关键线索，使得未来的延伸与历史观察保持一致。

Q2. 提出的奖励函数的具体形式是如何定义的？

与VIPER [28]和Division Reward [60]这两种都进行离散预测的模型不同，我们的模型预测的是连续的潜在变量。因此，我们的奖励是根据条件方差来估计的，而不是对数似然或熵。此外，使用对数似然来衡量不确定性需要将预测与真实值进行比较。由于我们在任何场景中部署奖励，VIPER的方法对于我们的目标来说是不可行的。需要注意的是，我们的奖励计算经过精心设计，以满足Kolmogorov公理，即它是非负的，并且整个样本空间的度量范围是[0, 1]。

Q3. 奖励估计效率与基于检测器的方法[127]的比较。

尽管我们的奖励估算涉及多轮去噪，但它不会比Drive-WM [127]中基于检测器的奖励函数消耗更多的计算资源。具体来说，Drive-WM从感知结果中获取奖励。由于检测器[82, 84]以图像序列为输入，Drive-WM必须在感知之前完成所有去噪步骤。不同的是，我们的奖励函数通过世界模型自身的不确定性来估算奖励，而不依赖于其他感知模型。因此，不确定性估算不需要完成整个生成过程，只需对每个样本进行几步去噪即可实现。实际上，如附录C中所述，每次情况下的奖励估算所需的总计算量（10步，5轮）并不大于生成整个视频（我们的模型为50步）的计算量，正如Drive-WM所做的那样。需要注意的是，我们的奖励估算的计算成本可以灵活降低以进一步提高其效率。如图14所示，使用5步去噪（默认计算量的50%）也能得到令人满意的奖励估算结果。

第四季度。提议的奖励函数的应用。

(1) 如第4.3节所述，所提出的奖励函数有可能作为缓解现有开环评估中担忧的驾驶行为替代度量[18, 83, 141]。(2) 如图10所示，更好的动作通常会通过我们的奖励函数获得更高的奖励。利用这一特性，我们的奖励函数有很大的潜力被用作批评模块[76]，通过执行最大化估计奖励的最佳动作来实现模型预测控制[27, 30, 35]。这一过程可以与基于分布的规划器[53, 148]结合进行，这些规划器能够提出动作以减少搜索空间。

Q5. Vista 还有哪些潜在应用？

(1) 作为一种可泛化的预测模型，Vista 可以被用作前向动力学模型 [13, 26]，以模拟短期动力学并辅助长期规划任务，如视觉导航 [106]。(2) 利用 Vista 作为隐式驾驶策略也颇具吸引力，这种策略是通过未来预测自发获得的 [1, 25]。在合成视频计划后，我们可以通过非因果逆动力学模型 [4, 65, 71] 将生成的图像轨迹转换为可执行的动作，这种模型可以从比模仿学习流程 [2, 9] 少得多的数据中高效学习。在自动驾驶中，逆动力学模型可以通过视觉里程计来实现 [75]。(3) 在与奖励函数协作的情况下，也值得研究 Vista 是否能通过提高现实世界场景中的采样效率来促进基于模型的强化学习 [137]。

Q6. 与GenAD的差异[136]。

这两项工作在控制灵活性和预测准确性方面存在根本差异。首先，Vista是一个可泛化的世界模型，能够通过多模态动作条件进行控制。尽管GenAD也训练了一个基于轨迹条件的扩展模型，但其权重完全在nuScenes上进行了微调，其动作控制的泛化性从未得到验证。相比之下，Vista集成了多样的动作可控性，能够以零样本方式泛化到新场景。与GenAD需要对OpenDV-YouTube进行命令和文本标注不同，我们的协作训练策略巧妙地避免了这种可能产生累积噪声和冲突的劳动[136]。此外，Vista (10 Hz, 576 × 1024) 在帧率和分辨率上远超GenAD (2 Hz, 256 × 448) 在时间和空间轴上的能力。与GenAD不同，我们还提出了几种专为高保真预测设计的方法。我们发现，尽管模型复杂度较低，Vista在FID和FVD评分上远优于GenAD (见表2)。

Q7. 局限性、失败案例及可能的解决方案。

作为一项开创性工作，Vista仍存在一些局限性，需要未来的研究来解决。(1) 由于Vista以极高的时空分辨率预测未来，因此在计算上不可避免地会非常昂贵，尤其是在下游应用中。潜在的解决方案可能包括更快的采样技术[89, 146]和基于训练的蒸馏方法[91, 93, 101, 103, 110, 124]。(2) 在长时间滚动预测或剧烈视角变化的情况下，预测可能会出现明显的退化。对预测结果进行额外的改进[6, 49, 54, 99, 126, 144]可能会有所帮助。推测性地，将我们的方法应用于更具扩展性的架构[54, 97]也有望解决这一限制。(3) 与其他可控视频生成方法[128]类似，我们的动作控制仍有失败的可能性，特别是在意图模糊的情况下，如图8所示的命令和目标点。结合更多带有动作标注的数据集[11, 112, 129]进行协同训练可能会有所裨益。使用组合的无分类器引导[8, 19, 32]来放大动作条件的个体影响也可能有所帮助（但会增加推理计算的成本）。(4) 尽管我们的训练数据基于最大的公开驾驶数据集[136]，但这远未涵盖互联网上的所有驾驶数据，因此Vista仍有巨大的潜力可以进一步扩展其能力。

Q8. 为什么不将Vista框架扩展到环绕视图生成？

确实，支持环视生成将进一步有助于驾驶。现有的工作[127, 145]已经在nuScenes[10]上探索了环视设置。然而，本文我们专注于前视设置，主要基于以下三个原因：(1) 前视设置允许利用多样化的数据源[54, 136]。相反，来自不同数据集的多视图视频中的差异，如不同数量的摄像头，阻碍了统一的建模和数据扩展。(2) 专注于前视的模型可以无缝应用于不同的数据集而无需适应[107]，从而扩大了其在不同数据集中的适用性。(3) 尽管不完整，前视通常包含驾驶所需的主要关键信息。如NAVSIM[23]所示，仅使用前视摄像头与使用五个环视摄像头相比，碰撞率仅下降1.1%。

Q9. 更广泛的影响。

尽管取得了令人鼓舞的进展，但在处理高度复杂情况的实际应用中，我们的工作远非完美。由于Vista基于扩散框架，引入了随机结果和不可忽略的延迟，直接将其部署到自动驾驶车辆中可能会带来安全风险。虽然它还不是万能的解决方案，但我们期望Vista能够激发社区进一步挖掘驾驶世界模型的能力和应用。作为可泛化的驾驶世界模型的原型，我们希望Vista能够促进对开发自动驾驶和机器智能通用系统的研究。

相关工作

B.1 世界模型

智能代理即使在未见过的情况下也应能做出有效决策[9, 57, 76, 113, 138, 149]。这需要对世界有根本性的理解，能够泛化到罕见的情况。

这种知识的内在表现形式，世界模型预测了在给定潜在行动下世界的合理未来 [9, 40, 69, 76, 96, 113, 137]。原则上，它不仅预测环境随时间的发展，还推导出潜在的物理动力学和主体行为。这些特性对于表示学习 [35, 45, 88, 105, 131]、基于模型的强化学习 [39, 40, 42, 43, 94, 96, 98] 以及模型预测控制 [27, 30, 35, 41, 92, 104, 142] 都很有用。最近的方法 [37, 44, 80, 85, 86] 还从大型语言模型中诱导出基于语言的世界模型，但受限于文本空间，难以与物理基础相结合 [26, 59]。

尽管世界模型在模拟游戏[40, 42, 43]和室内实体[72, 92, 120]中得到了广泛应用并取得了重大突破，但在自动驾驶领域的研究仍相对滞后[127, 143]。与其它任务不同，自动驾驶的世界建模面临独特的挑战，主要源于广阔的视野和高度动态的运动。一些实践设想在鸟瞰图（BEV）空间中构建世界模型[20, 31, 52, 53, 78, 81]。近期实践将世界状态建模为原始传感器观测数据，如点云[7, 67, 140, 143, 147]和图像[54, 61, 68, 90, 102, 125, 127, 131, 145]。后者，即视觉世界模型，由于传感器的灵活性和数据的易获取性，更具有扩展潜力。然而，现有方法局限于特定数据集[61, 77, 90, 125, 127, 143, 145, 147]或模拟器[7, 131]，限制了它们在新领域的泛化能力。同时，这些努力缺乏针对驾驶领域的系统设计，仅在相对较低的帧率和分辨率下建模世界，忽略了细粒度细节，损害了其表达真实世界行为的能力。此外，大多数方法局限于特定的控制模式[54, 61, 90, 125]，这阻碍了它们适应主流规划算法[12, 14, 21, 56, 58, 64]并扩展到更多应用，如决策[127]或用户交互[74]。此外，现有方法很少探索跨不同数据集的零样本动作可控性。其泛化性、保真度和可控性的不足共同阻碍了现有自动驾驶世界模型在推动自动驾驶发展中的广泛应用。

B.2视频生成

视频生成是建模世界的一种有效方式，并且近年来取得了显著的进展。开创性的工作[119, 134]研究了各种类型的生成模型[29, 33, 70]。受到扩散模型[24, 48, 100]成功的启发，基于扩散的视频生成方法大量涌现[6, 34, 46, 51, 108, 117, 121]。近期的工作[5, 15, 32, 144]将重点转向图像到视频的生成，因为这种生成方式在内容描述上更为精细，并且在训练数据的可扩展性上表现更好。然而，大多数方法并不是严格的预测模型，无法从条件图像开始生成视频。此外，现有方法在处理从自我视角出发的驾驶场景中的复杂动态时表现不佳[136]，这限制了它们作为驾驶世界模型的可行性。

尽管大多数现有方法生成的视频不具备显式的可控性，但最近的两项工作[128, 139]引入了摄像机运动控制到视频生成中。然而，摄像机运动在概念上与车辆动作不同，且这两项工作都是基于文本生成视频的方法，不具备任何预测能力。相反，我们开发的模型是一个预测性的世界模型，能够生成真实的动态效果，并为自动驾驶提供多样的动作控制。

实现细节

C.1 Model

我们采用SVD [5]框架作为Vista的架构，该架构总共包含2.5B参数，其中包括1.6B UNet参数。对于动作调节，我们将每个动作序列的值编码为具有128通道的傅里叶嵌入[114, 116]。

C.2 Dataset

我们使用经过严格筛选的OpenDV-YouTube [136]进行训练，并在动作控制学习阶段加入了nuScenes训练集 [10]。具体而言，我们手动剔除了OpenDV-YouTube中15小时的无关内容，最终获得了约1735小时的无标签驾驶视频。由于nuScenes存在严重偏差 [83, 107, 127]，我们根据指令类别平衡其样本，以促进对罕见动作的学习。视频片段以每秒10帧的速度采样，共25帧。尽管nuScenes [10] 记录频率为12 Hz，但我们发现将其处理为10 Hz并无负面影响。

以10 Hz的速度录制视频。模型输入通过裁剪和调整这些剪辑的大小以达到目标分辨率。我们将一个动作定义为由25帧组成的序列。为了将动作分类为命令，我们遵循规划中的既定惯例[56, 58, 64]，并定义当自车的最终位移在相对于其初始航向的正交方向上超过2米时，命令为“向右转”或“向左转”。为了允许更精确的分类，当向前行驶距离小于2米时，我们还引入了一个“停止”命令。

C.3Training

在第一训练阶段，我们在128个A100 GPU上以 576×1024 的分辨率训练所有UNet参数，进行20K次迭代，总共耗时约8天。我们累积了2步的梯度，得到有效批量大小为256。遵循SVD，我们的模型使用EDM框架[66]进行训练。我们使用AdamW优化器[87]，学习率为 1×10^{-5} 。空间层的学习率通过折扣因子0.1进行调整。方程(6)中的系数1和2分别设置为1.0和0.1。偏移噪声[38]的强度为0.02，有助于提高时间平滑度。我们随机抽取不同顺序的动态先验，概率递增，即分别为0、1、2、3条件帧的 $1/15$ 、 $2/15$ 、 $4/15$ 、 $8/15$ 。噪声增强[49]被禁用，以保留更多来自条件帧的细节。

关于动作控制学习阶段，我们冻结了预训练的权重，并在UNet的所有注意力块中添加了LoRA[55]和投影层。LoRA的秩设置为16。然后，我们在 320×576 分辨率下以批量大小8和学习率 5×10^{-5} 训练新权重，共进行120K次迭代。在可清晰观察到控制能力后，我们继续在 576×1024 分辨率下对未冻结的权重进行另外10K次迭代微调。我们以15%的比例随机丢弃每个激活的动作模式，以允许无分类器引导[50]。在此训练阶段，OpenDV-YouTube和nuScenes的采样比例为1:1。整个动作控制训练过程在8个A100 GPU上大约需要10天，其中低分辨率阶段约8天，高分辨率阶段约2天。

C.4Sampling

我们使用DDIM采样器[109]进行50步采样，采样从max为700.0开始。由于我们的模型以自回归方式预测长期未来，标准无分类器引导导致的过饱和问题会迅速累积。因此，与SVD线性增加引导尺度不同，我们采用了一种三角形无分类器引导方案[118]，以允许真正的长期展开。具体来说，对于每个K帧中要预测的第i帧，我们按以下方式分配其引导尺度s(i)：

$$s(i) = \begin{cases} s_{\min} + \frac{2i}{K}(s_{\max} - s_{\min}) & \text{if } i < \frac{K}{2}, \\ s_{\max} - \frac{2(K-i)}{K}(s_{\max} - s_{\min}) & \text{if } i \geq \frac{K}{2}, \end{cases} \quad (9)$$

其中，smin 和 smax 分别表示时间轴上的最小和最大引导尺度。在我们的实验中，我们将smin 定义为 1.0，smax 定义为 2.5。这种三角形方案为将在下一预测轮中用作条件的帧分配了适中的引导尺度。由于充分的时间交互，中间帧的质量也能传播到引导尺度较低的帧。如图 15 所示，该技术巧妙地缓解了饱和漂移问题，同时增强了细节。为了提高感知连续性，我们在将生成的潜在表示发送至视频感知解码器[5]之前，将其分割成重叠 3 帧的片段。解码后，重叠的帧按像素进行平均。

C.5 人类评估

回顾一下，我们要求参与者从视觉质量和运动合理性两个方面对并排的视频对进行评判。为了保证反馈的可信度，我们为人类评估的各个方面都提供了详细的注释。在视觉质量方面，我们让参与者关注生成内容的连贯性和和谐性。在运动合理性方面，我们鼓励参与者更多地关注自车和其他代理移动方式的可信度，例如，它们是否遵守交通规则并表现出安全行为。对于我们比较的所有公开模型，我们都使用官方的检查点和配置进行推理，没有进行微调。对于需要文本输入的模型[133, 144]，我们将提示设置为“真实的驾驶视角”。

表4：Waymo上的命令奖励。地面真实命令通常比随机命令输入获得更高的奖励，这表明所提出的奖励函数可以用作命令选择的可靠指标。

Condition	Average Reward
GT Com.	0.892
random Com.	0.878 (-0.014)

C.6 奖励估算

表5：动作独立性的影响。不失一般性，我们选择轨迹作为评估的代表性动作。所提出的约束加速了动作的学习。

Strategy	Action	Subset FVD			
		forth	right	left	stop
w/o A.I.	w/o Traj.	163.0	273.9	428.3	497.1
	w,Traj.	138.8	232.9	368.2	132.3
w,A.I.	w/o Traj.	156.2	263.7	402.9	463.7
	w,Traj.	130.7	230.8	345.7	118.9

对于每个条件框架和动作对，我们累积了一个大小为 $M = 5$ 的集成模型，以获得可靠的不确定性估计。每个集成样本在进行 10 步去噪推理时，我们发现对于不确定性估计来说，生成高质量结果并非必要。相关策略 [35, 95] 中的系数设为 0.5。

C.7 消融研究

对于损失函数的消融实验，我们在OpenDV-YouTube [136]上以 576×1024 的空间分辨率对每个变体进行10K步的训练。所有消融实验，包括附录D中的额外消融实验，均通过加载SVD [5]的预训练检查点进行初始化，并使用8块A100 GPU进行实验。

附加实验

D.1 奖励估计的参数敏感性

为了研究去噪步数和集成规模如何影响所提出的奖励函数的表现，我们使用不同的超参数设置重复了第4.3节中的奖励估计过程。我们首先使用5个去噪步数和每个样本5个集成规模进行实验。然后，我们测试了两种变体，分别通过将去噪步数增加到10（我们在附录C中的默认设置）和将集成规模增加到10，从而将计算成本翻倍。根据第4.3节，我们在图14中绘制了这三种变体的估计奖励与L2误差的相关性。结果显示，增加去噪步数可以大大增强奖励的相对对比度，这表明在相同的计算预算下，去噪步数是比集成规模更重要的奖励估计因素。

D.2 命令奖励估计

为了证明所提出的奖励函数同样适用于其他行动，我们估算了Waymo的真实指令的奖励，并与随机指令的奖励进行了比较。表4的结果表明，我们的奖励函数在指令选择方面同样有效。

D.3 行动独立性约束

为了证明我们提出动作控制学习策略的有效性，我们通过移除第3.2节中提出的动作独立性约束来进行对比实验。我们在nuScenes数据集上以 320×576 像素的分辨率训练了两个变体，训练步数为62K步。对比结果如表5所示。

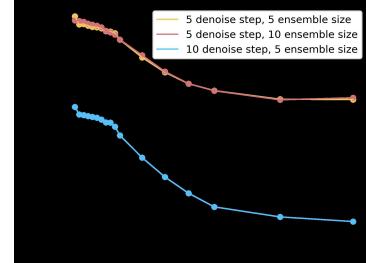


图14：奖励估计对超参数的敏感性。增加去噪步骤的数量可以产生更具区分性的奖励，而增加集成规模可以略微稳定估计。



图15：指导尺度效果。我们使用不同的CFG方案预测15秒的长期视频。我们的方法在细节生成和饱和度维持之间达到了最佳平衡。



图16：LoRA适应的必要性。单独训练新增加的投影而不使用LoRA会导致视觉损坏。比较的变体在nuScenes上训练并在Waymo上推断。

D.4三角形引导方案

我们进一步将引入的无分类器指导方案与普通方案和线性方案[5]进行比较，以验证其必要性。图15显示，我们的三角缩放方法在视觉质量和饱和度保留之间达到了最佳平衡。

D.5LoRA适配

为了展示在第3.2节中应用LoRA的必要性，我们在 320×576 像素的分辨率下训练了两个变体，进行了30K次迭代。在保持预训练的UNet权重不变的情况下，我们让其中一个变体训练LoRA和注意力块中的动作投影层，而另一个变体仅调整新的投影层。如图16所示，添加LoRA对于动作控制学习是至关重要的。

D.6行动控制一致性

在表6中，我们报告了图8的完整FVD分数，这进一步验证了各种动作控制的有效性。需要注意的是，由于我们的“停止”子集包含的样本最终位移在2米以内，目标点通常不会出现在这些视频中。因此，对于在“停止”子集上使用目标点作为动作条件的实验，大多数样本的生成方式与无动作模式相同。

表6：不同动作类别的完整FVD评分。我们根据命令类别将数据集分为四个子集，并计算了FVD评分。所有类型的动作控制在各个类别中均表现有效。

Dataset	Condition	Subset FVD				
		forth	right	left	stop	average
nuScenes	action-free	135.6	405.6	513.8	414.1	367.2
	+ goal point	122.4	315.6	439.6	413.5	322.7
	+ command	122.2	299.7	485.6	261.6	292.2
	+ angle & speed	122.8	285.6	397.8	114.1	230.0
	+ trajectory	125.2	229.2	357.7	118.5	207.6
Waymo	action-free	145.9	407.6	529.9	164.1	311.8
	+ command	122.5	331.5	496.9	143.9	273.7
	+ trajectory	126.3	285.5	527.6	136.5	268.9

D.7与GenAD的人类评估

据我们所知，目前尚未有公开可用的特定于驾驶的世界模型，这使得进行定性的人类评估变得困难。因此，我们主要通过表2中官方报告的FID和FVD分数，将Vista与现有的方法进行比较。

为了展示在视觉质量和运动合理性上的显著提升，我们进行了额外的人类评估，采用了最先进的GenAD模型[136]。由于GenAD每次处理4秒的视频，我们通过自回归预测将Vista的输出扩展到5秒，然后裁剪掉最后一秒以与GenAD的时长对齐。为了避免分辨率和频率带来的任何偏差，我们将Vista的输出（576×1024分辨率，10 Hz）下采样至256×448分辨率，2 Hz。评估过程遵循第4.1节中规定的相同步骤。

我们从未见过的OpenDV-YouTube-val数据集中收集了25多个样本，并邀请了20名志愿者进行评估。我们请志愿者选择他们认为更好的视频。结果显示，Vista在视觉质量和运动合理性方面分别以94.4%和94.8%的比例被优先选择。这表明，尽管Vista由于降采样而经历了较大的感知损失，但在生成质量上仍显著优于GenAD。我们还在图17中比较了GenAD和Vista的预测结果，显示了Vista在分辨率和保真度上的优势。

附加视觉化展示

E.1 泛化能力

我们进一步展示了Vista在不同野外场景中的强大泛化能力。图18和图19的结果表明，Vista能够在非常多样化的场景中做出高保真的预测。

E.2长期预测

除了图6，我们在图20中提供了更多关于长时间预测的定性可视化结果。Vista能够连续地预测长期未来的内容和动作，保持一致性。

E.3动作可控性

我们在图21中提供了更多不同动作输入的预测结果。在OpenDV-YouTube-val [136] 和 Waymo [112] 上的结果显示，Vista的多功能可控性可以零样本方式轻松迁移到不同领域。

E.4反事实推理能力

反事实推理能力是世界模型涌现能力之一[36]。如图22所示，Vista能够有效预测异常行为所导致的反事实后果。

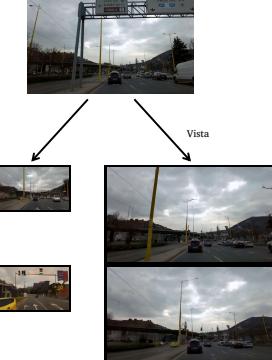


图17：GenAD与Vista之间的感知差异。

E.5 人类评估案例

为了展示用于人工评估（第4节）的场景的多样性，我们在图23中展示了从OpenDV-YouTube-val [136]、nuScenes [10]、Waymo [112]和CODA [79]中收集的所有案例。

资产许可证

我们的训练和评估使用了来自四个公开授权数据集的数据[10, 79, 112, 136]。我们的实现基于SVD[5]的代码库，该代码库采用MIT许可证。SVD的预训练检查点则根据stable video di usion非商业社区许可证进行分发。

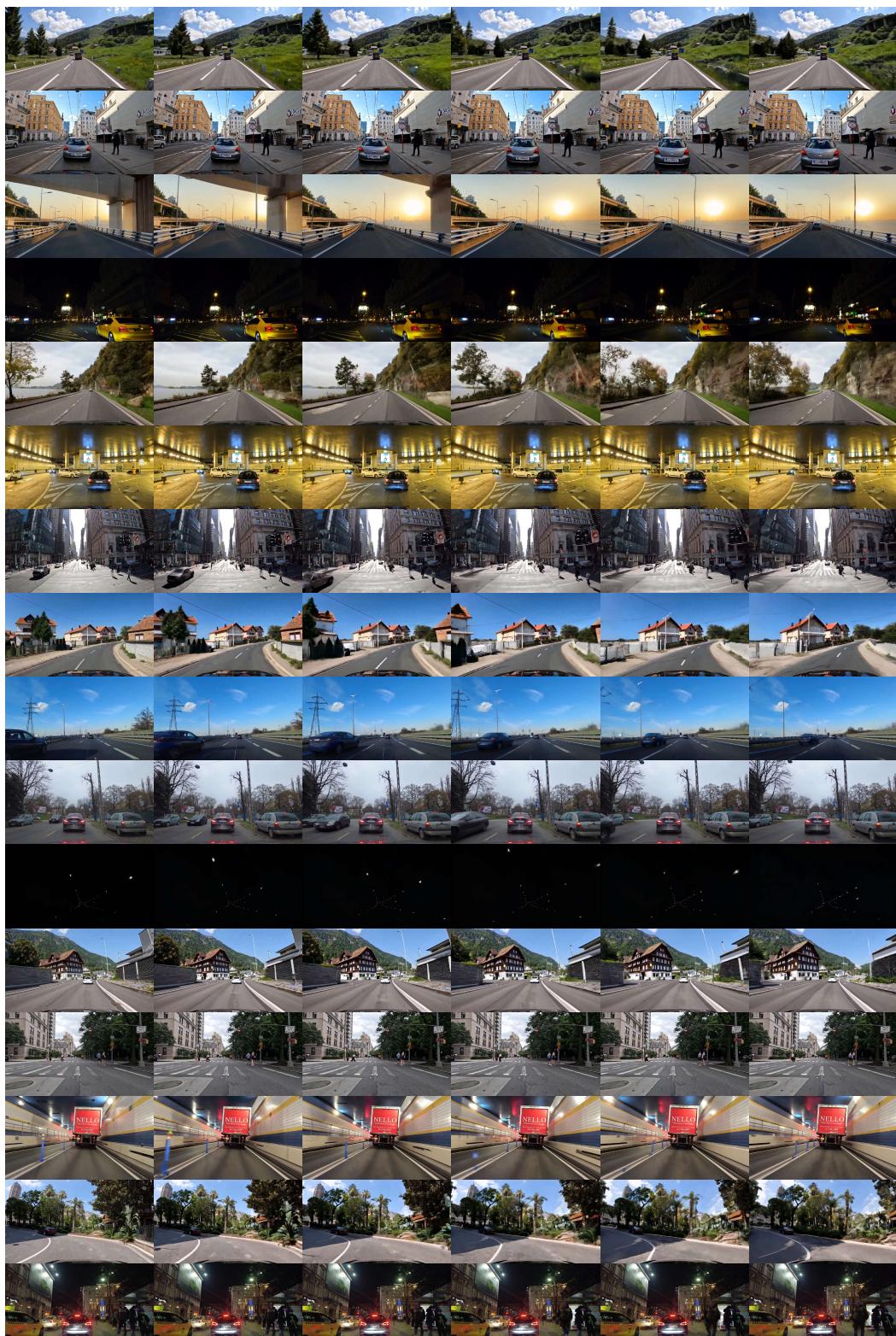


图18：Vista的泛化能力。我们在多种场景（如乡村和隧道）中应用Vista，这些场景包含未见过的相机姿态（例如双层巴士的视角）。我们的模型能够预测具有车辆和行人生动行为的高分辨率未来画面，展现出强大的泛化能力和对世界知识的深刻理解。最佳查看方式是放大。

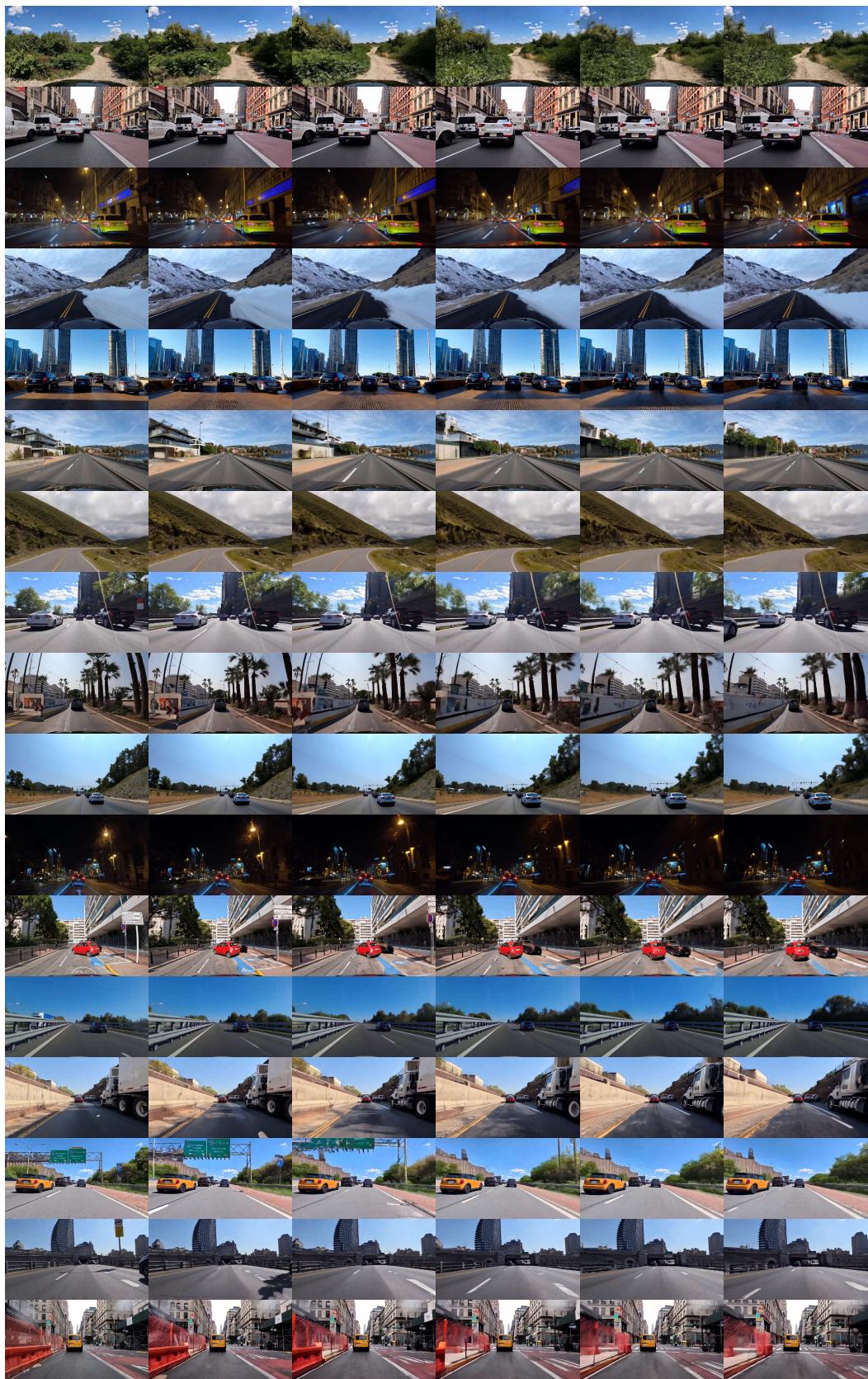


图19：Vista在更多场景中的泛化能力。最佳查看方式为放大查看。

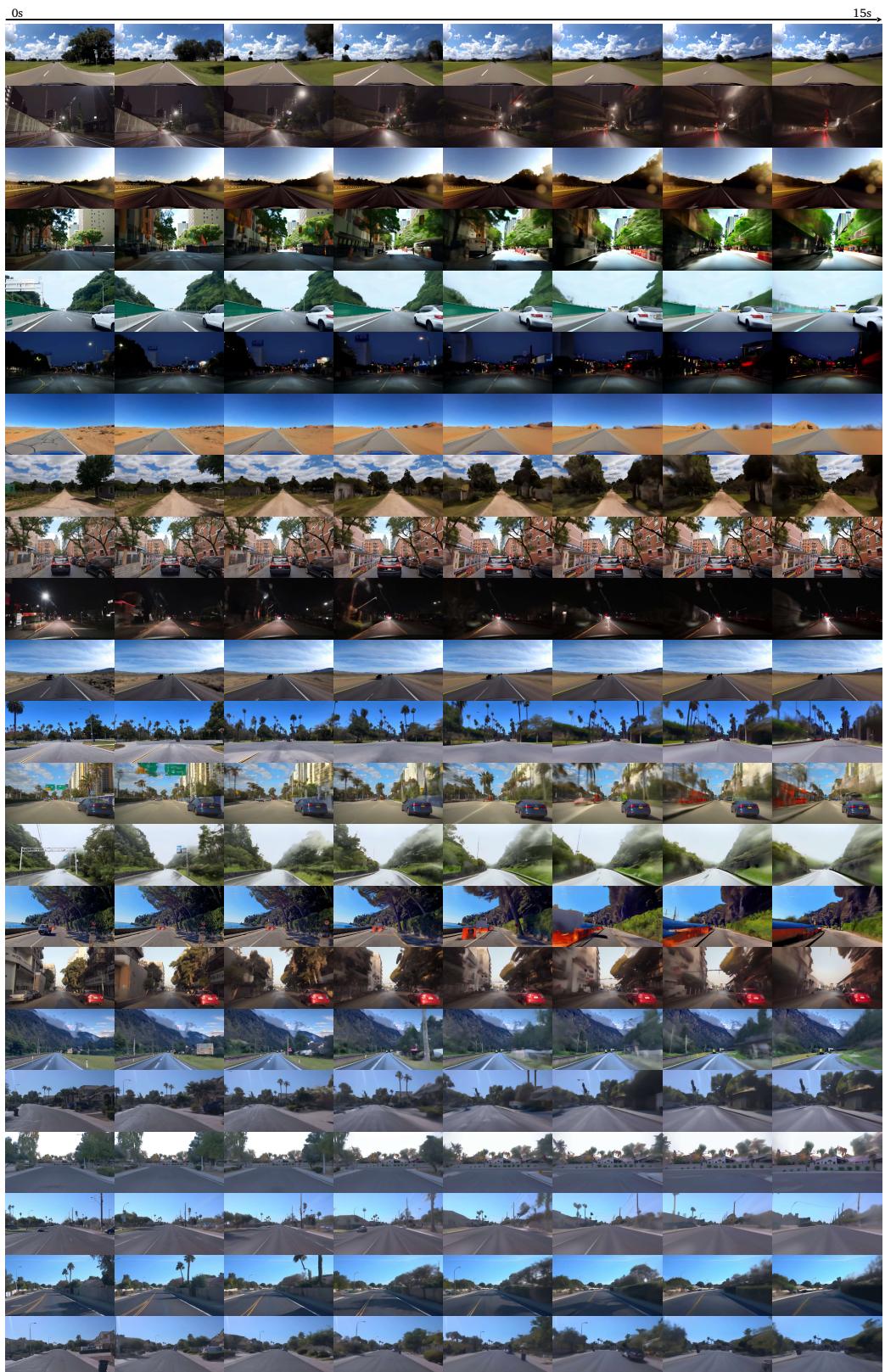


图20：长时间预测的额外结果。我们的模型可以自回归地模拟长时间驾驶体验，且质量下降极小。所有视频均以10赫兹的频率持续播放15秒。

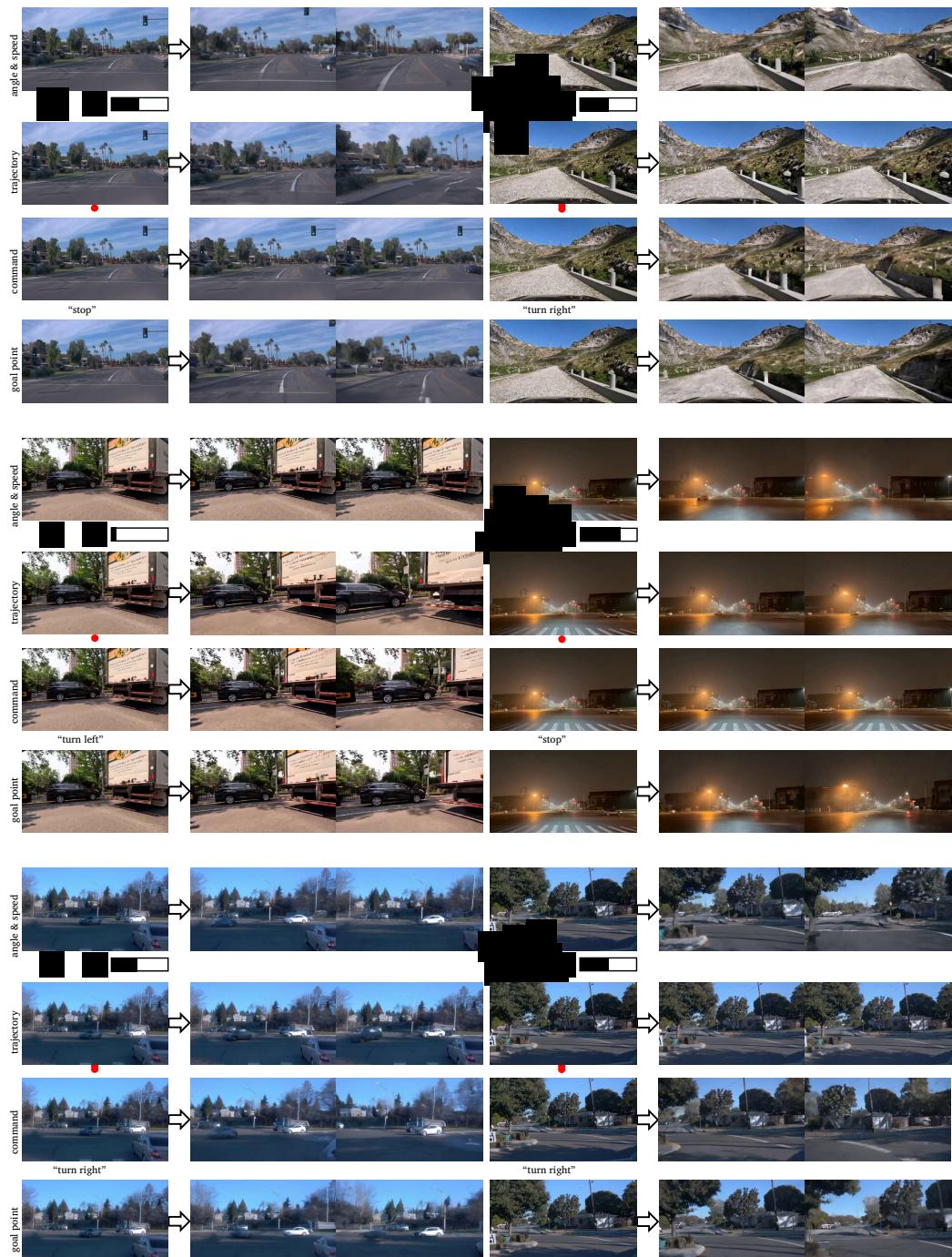


图21：动作可控性的额外结果。我们在OpenDV-YouTube-val和Waymo的多个场景中试验了不同的动作条件。通过各种干预措施，可以持续控制自行车的行为。



图22：反事实推理论。通过施加违反交通规则的动作，我们发现Vista也能够预测异常干预的后果。在第一个例子中，自行车按照我们的指示越过道路边界，冲入灌木丛。在第二个例子中，当我们强制自行车在十字路口继续行驶时，经过的车辆停止并等待，以避免碰撞。这展示了Vista在促进闭环模拟方面的潜力。



图23：为人类评估收集的多样化场景。我们从OpenDV-YouTube-val、nuScenes、Waymo和CODA中精心挑选了60个场景。每个数据集的独特属性共同代表了真实世界环境的多样性，从而允许进行全面的人类评估。

如果有大量文档或图书需要翻译，请联系我们，为您提供更专业的服务。

邮箱: yilong2001@126.com

公众号: CloudAI技术

请发邮件或公众号留言

