arXiv:2603.09086v1 [cs.RO] 10 Mar 2026

自动驾驶的潜在世界模型：统一的分类体系、评估框架与开放挑战

Rongxiang Zeng 和 Yongqi Dong*

摘要——**新兴的生成式世界模型和视觉-语言-动作（VLA）系统正迅速重塑自动驾驶，通过支持可扩展仿真、长时域预测以及能力丰富的决策。这些方向均以潜在表示为核心计算基底：它们压缩高维多传感器观测，实现时间连贯的展开，并为规划、推理和可控生成提供接口。本文提出一个统一的潜在空间框架，综合了自动驾驶世界模型领域的最新进展。该框架依据潜在表示的目标与形式（潜在世界、潜在动作、潜在生成器；连续状态、离散词元及混合形式），以及几何、拓扑和语义的结构先验，对设计空间进行组织。在此分类体系的基础上，本文阐述五项贯穿各领域的内部机制（即结构同构、长时域时间稳定性、语义与推理对齐、价值对齐目标与后训练，以及自适应计算与审慎推理），并将这些设计选择与鲁棒性、泛化和可部署性联系起来。本文还提出了具体的评估方案，包括旨在缩小开环/闭环不匹配的闭环指标套件和资源感知的审慎推理成本。最后，本文提出了切实可行的研究方向，以推动潜在世界模型发展，使其能够支持可用于决策、可验证且资源高效的自动驾驶。**

索引词—**世界模型、潜在表示、自动驾驶、生成建模、闭环评估。**

I. 引言

AUTOMATED 驾驶将高维多传感器感知与严格安全约束下的长时域决策相结合。尽管大规模真实世界驾驶日志加速了基于学习的自主驾驶 [1], [2], ，但稀有事件和对抗性机动等安全关键交互仍然稀少，且在闭环中验证成本高昂 [3], [4], [5], ，而纯合成模拟器则会带来不可忽视的仿真到现实差距 [1], [5], [6]。此外，以人为中心且符合社会规范的行为在物理安全之外提出了更多结构性要求，因此模型必须考虑交互规范和合作意图 [7], [8], [9]。这些现实促使我们探索能够压缩原始观测、支持时间连贯的“想象”与预测，并为下游规划与控制揭示决策相关结构的表征。

（*通讯作者：*Yongqi Dong*。）Rongxinag Zeng 就职于亚琛工业大学，地址：德国亚琛 Templergraben 55, 52062（电子邮箱：rongxiang.zeng@rwth-aachen.de）。Yongqi Dong 就职于代尔夫特理工大学，地址：荷兰代尔夫特 2628 CN（电子邮箱：yongqi.dong@rwth-aachen.de）
数字对象标识符 10.1109/TITS.20XX.XXXXXXX

早期基于学习的流程展示了从感知到控制的直接映射，后来的模仿学习和混合方法则通过更丰富的监督和有针对性的场景合成提升了鲁棒性；然而，它们也暴露出在分布偏移和长尾交互下的脆弱性 [10], [11], [12], [13], [14], [15]。与此同时，强化学习（RL）将驾驶视为具有明确状态转移和奖励结构的序贯决策[16], [17], [18], [19]。学习到的动力学和潜在世界模型通过从原始感官输入中推断紧凑的潜在状态，并支持在潜在空间中基于想象的展开，拓展了这一视角 [20], [21], [22], [23]。这一转变表明，中间潜在表示的结构和时间演化往往是稳定性、可控性和泛化能力的关键。

因此，潜在空间已成为一种统一的计算基底。它们充当紧凑的信息瓶颈，将高维观测压缩为结构化表征，以实现高效的学习和推理，同时支持对几何、拓扑和动力学施加显式约束，而这些约束很难在原始传感器空间中实现。表征学习、离散潜变量建模（令牌化）以及视觉-语言预训练的发展带来了新的能力：离散潜变量支持词元化生成建模和基于序列的推理 [24], ，大规模视觉-语言预训练结合语言/语义监督，可提供可迁移的语义对齐信号，从而引导或蒸馏与推理相关的表征 [25]。面向自动驾驶的现代世界模型日益利用这些能力，超越像素级或感知层面的重建，转向决策相关的展开接口 [1], [4], [26], [27], [28]。

尽管相关领域进展迅速，现有文献仍分散于以任务和以架构为中心的不同视角，例如预测与规划、扩散模型与Transformer生成器，以及开环预测与闭环控制。这种碎片化掩盖了推动成功或导致失败的共同内部机制。许多经验层面的不足，包括长时域漂移与幻觉、跨域泛化失败，以及感知指标与闭环安全性之间的不匹配，都可以追溯到潜在表示如何编码结构、随时间演变，并与语义、目标及计算约束相对齐。以潜在表示为中心的视角为比较那些看似毫不相关的方法提供了有原则的基础，也阐明了为何视觉保真度或开环准确率的提升未必能转化为更安全的闭环行为。

为解决这些碎片化问题，本文采用一种

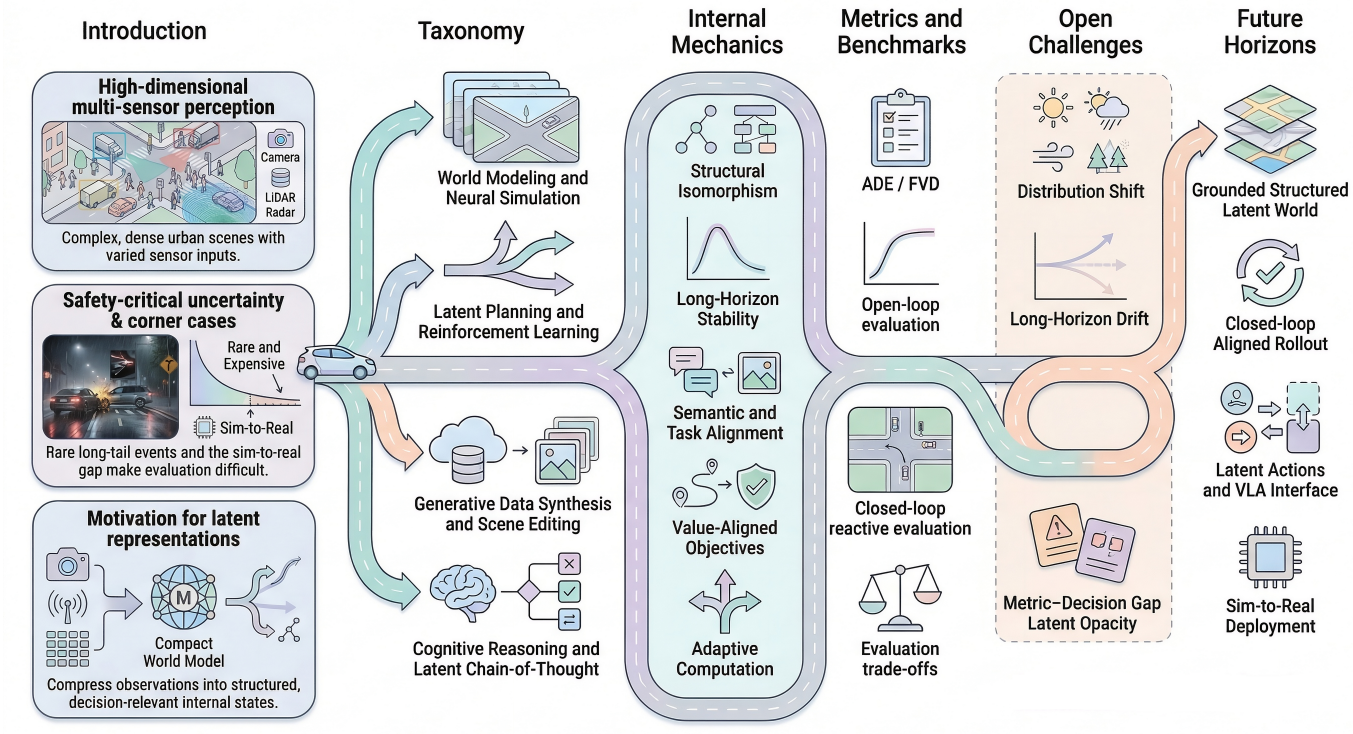

图 1. **本文的可视化路线图** .

统一的潜在空间视角与分类体系，通过表征设计和决策对齐的视角，重新审视自动驾驶世界模型的最新进展。本文并非单纯按照任务或架构对方法进行分类，而是强调潜在表示的结构与计算属性，因为这些属性会直接影响闭环部署。图1并非呈现字面意义上的流程管线，而是作为一场概念性的“虚拟旅程”，将本文展开为一份连贯的设计路线图：首先介绍原始感知与基于学习的方法在实践中面临的复杂性和挑战（如第一节 引言所示），随后构建精简的建模范式分类体系（第二节 分类体系），进而聚焦于决定推演行为与稳定性的核心内部机制（第三节 内部机制）。在此基础上，本文考察开环预测与闭环交互的评估标准有何不同（第四节 指标与基准测试），并探讨尚未解决的开放挑战（第五节 开放挑战），最后概述未来方向，重点关注结构化潜在表征与统一的评估原则（第六节 未来展望）。

简而言之，本文的主要贡献有五点：(1) 提出一个统一的分类体系，用于刻画现代自动驾驶世界模型所采用的目标、形式和结构先验，并将其置于下游决策接口的关联框架中；(2) 综合分析五类贯穿始终的内部机制（即结构同构与几何先验、长时域时间动态、语义与推理对齐、价值对齐目标与后训练，以及用于审慎推演的自适应计算），并分析这些机制如何影响闭环鲁棒性、泛化能力和可部署性；(3) 提出一套精简的评估方案，包括闭环指标套件和资源感知的审慎推理成本，旨在缩小开环感知得分与闭环安全关键结果之间的差距；(4) 推导出具有指导性的设计建议和一份优先级明确的研究议程，以推动研究从视觉上可信的推演迈向与决策相关且可验证的模型；(5) 汇总并整理具有代表性的基准和示例方法，以促进可复现的后续研究。总体而言，这些贡献使本文从目录式梳理转向一套具有指导性的设计与评估议程，旨在加速将潜在世界模型转化为可部署的自动驾驶系统。

II. 自动驾驶世界模型分类体系

图 2 展示了所提出分类体系的概念概览，说明不同建模范式如何通过共同的潜在空间视角相互关联。作为对这一概览的补充，表 I以精简的属性级比较形式，按照潜在表征形式、建模目标以及与下游决策任务的集成方式，对代表性方法和研究进行整理。图表结合起来，呈现了一个统一视角，将神经仿真、潜在空间中心规划、生成式数据合成与场景编辑，以及认知推理和潜在思维链机制纳入世界建模

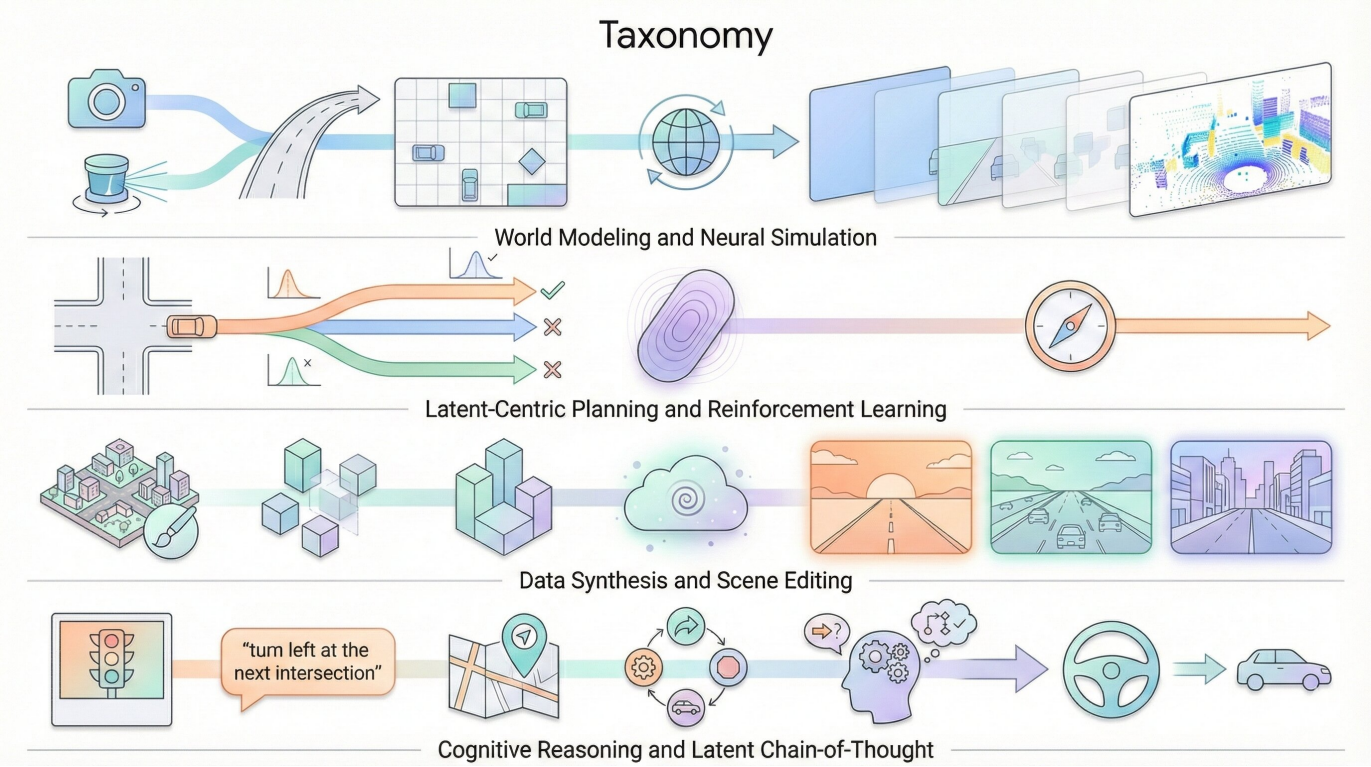

图 2. **自动驾驶世界模型的分类体系**：统一的潜在表示中心框架下，神经仿真、潜在规划、数据合成和认知推理 的概念概述。

框架之中。该分类体系并未将这些范式视为彼此孤立的方向，而是强调它们共有的结构选择及其权衡，例如表征流形（连续、离散或混合）、施加的几何或拓扑先验、时间因子分解策略以及价值对齐目标；这些因素会显著影响滚动推演保真度、推理能力以及对闭环控制的适用性。本节余下部分将详细阐述这些维度，说明不同范式如何具体体现并权衡这些基本设计原则。

A. 时空世界建模与神经仿真

此类别涵盖构建神经模拟器以近似物理世界演化的方法，重点是生成高保真、时空一致的未来观测，范围从二维（2D）视频序列到三维（3D）占用流，为下游感知与预测任务奠定基础。此处的神经仿真 是指在学习得到的潜在动力学模型中进行动作条件展开，以执行预测与决策，而非进行照片级真实感渲染。

为解决不同模态之间的几何差异，**BEVWorld** [29] 将异构输入（例如来自激光雷达和摄像头的数据）投影到统一的鸟瞰图潜在空间中，从而通过光线投射实现联合自监督重建。为进一步推进多模态统一，**OmniGen** [2] 利用共享鸟瞰图空间联合解码激光雷达和摄像头数据，通过体积渲染确保合成点云与多视角图像之间严格的几何一致性。为增强泛化能力，**DriveX** [30] 引入了一种解耦建模策略，将世界表征学习与未来状态解码分离，并利用动态感知射线采样捕捉场景的整体演变。

由于自回归误差累积，保持长时域预测的一致性仍是一项主要挑战。**Epona** [31] 通过时空因子分解应对这一问题，将时间动态与细粒度空间生成解耦。针对离散与连续之争，**Orbis** [32] 表明，连续自回归流匹配 [33]在长期稳定性方面优于离散标记化，且无需深度或地图监督。此外，**LongDWM** [34] 采用跨粒度蒸馏，以细粒度流监督粗粒度潜在动力学，实现无限视频生成。

为实现一致的环视合成，**UniMLVG** [35]将显式视点建模融入扩散 Transformer，而 **CoGen** [36] 则利用生成的三维几何条件引导视频生成过程。进一步拓展至视频之外，**DriveWorld**[26]和 **Drive-OccWorld** [37] 利用解耦的内存库（分离静态与动态特征）预测四维（4D）占用流，从而有效衔接生成式仿真与下游感知和规划任务。

表I代表性世界模型方法比较：范式与属性

方法名称缩写	潜在结构	潜在动力学模型	关键机制	输入源 → 输出
第1部分：神经仿真与世界建模				
BEVWorld [29]	多模态鸟瞰图网格	Transformer	通过光线投射渲染和扩散实现的多模态词元化	摄像头、激光雷达 →点云、视频
CoGen [36]	2D + 3D 条件	扩散	通过自适应几何条件实现 3D 一致性视频生成	摄像头 → 视频
Drive-OccWorld [37]	3D 占用	Transformer	结合语义和运动归一化的 4D 占用预测	摄像头 → 占用、流
DriveDreamer [1]	二维潜变量	扩散	结构化交通约束理解的两阶段训练	摄像头、三维边界框 →视频
DriveScape [38]	二维 + 3D 道路先验	Transformer	双向调制特征融合用于多视图合成	摄像头 → 视频
DriveWorld [26]	双重记忆	MSSM	采用动静分离的时空记忆状态空间建模	摄像头 → 4D特征
DriveX [30]	解耦	Transformer	采用动态感知射线采样的全场景建模	摄像头、激光雷达 → 3D 表征
Epona [31]	因子化	扩散	自回归潜空间中的解耦时空因子分解	摄像头 → 长视频
GAIA-2 [39]	二维潜变量	扩散	通过多智能体潜嵌入实现可控场景合成	摄像头、文本 →视频
GLAD [40]	二维潜变量	扩散	通过潜变量传播进行流式场景生成	摄像头 → 流式视频
LongDWM [34]	分层	Transformer	用于无限视频生成的跨粒度运动蒸馏	摄像头 → 无限视频
MagicDrive-V2 [41]	二维潜变量	扩散	用于自适应几何控制的时空条件编码	摄像头、文本 →视频
Orbis [32]	一维向量	流匹配	用于长时域稳定性的连续自回归流匹配	摄像头 → 视频
UniMLVG [35]	二维潜变量	扩散	用于运动状态转移一致性的显式视点建模	摄像头 → 多视角视频
第二部分：潜在空间中心规划与强化学习				
AD-R1 [42]	潜在状态	RL	用于端到端驾驶策略优化的闭环强化学习	摄像头 → 驾驶动作
BRYANT [43]	频域	Transformer	脑启发因果感知对比学习与CfC网络	摄像头 → 驾驶动作
DriveLaW [44]	视频潜变量	扩散	通过直接潜变量注入实现统一视频生成与规划	摄像头 → 轨迹，视频
EGADS [45]	一维向量	流 + 变分自编码器	利用归一化流提高样本效率的变分推断	摄像头 → 驾驶动作
GenAD [46]	实例令牌	Transformer	地图感知的实例中心场景令牌化与先验建模	摄像头 → 规划轨迹
Interp-E2E [47]	一维 + 掩码	RL	通过语义鸟瞰掩码实现的序列潜在解释	摄像头 → 驾驶动作
LatentDriver [48]	一维向量	Transformer	用于不确定性感知决策的概率混合分布	摄像头 → 规划轨迹
LAW [49]	二维潜变量	Transformer	用于端到端规划的自监督未来场景特征预测	摄像头 → 规划轨迹
MomAD [50]	一维向量	Transformer	结合拓扑轨迹匹配的动量感知规划	摄像头 → 规划轨迹
Raw2Drive [28]	一维向量	Transformer	对齐的特权到原始模型蒸馏的引导机制	摄像头 → 驾驶动作
SEM2 [51]	掩码潜变量	Transformer	用于提升鲁棒性的语义掩码与循环滤波	摄像头 → 驾驶动作
Think2Drive [4]	一维向量	RSSM	用于高效强化学习的神经世界模型模拟器	摄像头 → 转向动作
World4Drive [52]	二维特征	Transformer	通过视觉基础模型先验实现意图感知对齐	摄像头 → 规划轨迹
WorldRFT [27]	一维向量	Transformer	通过组相对策略优化进行强化微调（RFT）	摄像头 → 轨迹
第3部分：生成式数据合成与场景编辑				
LiDAR-EDIT [53]	鸟瞰图令牌	MaskGIT	用于布局编辑中射影几何一致性的球形体素化	点云 →编辑后的点云
OmniGen [2]	共享鸟瞰图	DiT + ControlNet	基于体渲染的统一多模态（激光雷达+摄像头）生成	摄像头、激光雷达 →传感器数据
Safety-Critical [3]	图潜变量	扩散	用于物理合理交通仿真的对抗引导采样	轨迹 → 安全关键轨迹
SubjectDrive [54]	二维潜变量	扩散	用于提升生成数据多样性的主体控制机制	摄像头 → 合成图像
SynDiff-AD [55]	二维潜变量	扩散	通过语义提示实现面向子群体的合成数据增强	文本 →合成图像
第4部分：认知推理与潜在思维链				
Alpamayo-R1 [56]	一维向量	Transformer	强化学习反馈下的因果链推理	摄像头 → 规划轨迹
ColaVLA [57]	元动作	Transformer	基于认知元动作嵌入的分层规划	摄像头、文本 →轨迹
CRiTIC [58]	图潜变量	Transformer	面向鲁棒预测的注意力门控因果发现	摄像头 → 智能体轨迹
Exp-Latent [59]	VAE潜变量	-	用于可解释性探测的自动潜变量扰动	摄像头 → 转向指令
FutureSight [60]	视频潜变量	扩散	用于未来状态预测的时空视觉思维链	摄像头 → 规划轨迹
FutureX [61]	潜在思维链	Transformer	用于自适应推理与反射之间动态切换的“自动思考切换”	摄像头 → 轨迹
LCDrive [62]	潜在动作	Transformer	采用动作-世界令牌交错的潜在思维链（CoT）	摄像头 → 动作令牌
MindDrive [63]	一维向量	VLM	基于视觉语言模型的多目标评估“假设”模拟	摄像头 → 轨迹
OmniDrive-R1 [64]	视觉令牌	Transformer	强化驱动的视觉定位（Clip-GRPO），用于思维链	摄像头、文本 →动作
VA-VAE [65]	视觉词元	-	用于化解优化困境的视觉基础模型对齐	摄像头 → 图像重建
VERDI [66]	一维向量	-	将视觉语言模型推理蒸馏到模块化自动驾驶技术栈中	摄像头 → 动作、推理文本
Vis-Ling [67]	一维向量	Transformer	用于视觉语言语义特征微调的多模态推理	摄像头 → 行人轨迹
VLA-R [68]	嵌入空间	检索	开放世界视觉语言动作检索范式	摄像头、文本 →驾驶动作

B. 潜在空间中心规划与强化学习

超越像素级处理后，潜在空间中心规划方法利用世界模型压缩且结构化的潜在空间，实现高效的轨迹规划与策略学习。通过利用这些模型的“推演”能力，它们能够模拟未来结果，以优化决策策略，同时避免高维感知数据带来的计算开销。

传统的端到端自动驾驶方法通常依赖遵循“感知 →预测 →规划”这一顺序的级联流程。这些系统难以明确捕捉交通参与者之间的交互式未来演变，也缺乏对轨迹分布进行结构化先验建模的方法。为弥合这一差距，**GenAD** [46]将自动驾驶重新定义为“未来生成”问题，并在共享潜在空间中统一建模智能体与自车的运动演变，通过潜在空间采样生成多模态未来轨迹分布。这一框架使多车辆运动预测与自车规划能够在单一、连贯的框架内同步进行。进一步推进统一建模的理念，**DriveLaW** [44] 消融了视频与运动规划之间的边界生成与运动规划。与级联流程不同，它将高保真视频生成器的潜在表示直接注入规划器，确保规划轨迹与预测的未来动态天然一致。

尽管自回归世界模型效果显著，但它们往往难以表示多模态决策，并且在闭环规划过程中容易出现“自我欺骗”。为了解决这些问题，**LatentDriver** [48] 将到下一状态的状态转移和自车动作建模为混合分布，并据此推导出确定性控制。此外，它将中间层生成的采样动作反馈至世界模型，以缓解自我欺骗，从而确保更稳健的长时域预测。

潜在世界模型的一个显著优势在于，它们可作为强化学习（RL）的神经模拟器，从而显著提升训练效率并降低样本复杂度。**Think2Drive** [4] 充当神经模拟器，在低维潜在空间中学习环境转移。该架构使规划器能够进行大规模并行“梦境”展开，从而以较高的训练效率实现专家级驾驶策略。

同样，**EGADS** [45] 利用强化学习智能体，借助与归一化流相结合的变分推断，将历史信息压缩到潜在空间中。这种概率方法有助于有效提取驾驶相关历史特征，同时尽量降低样本复杂度。

然而，将基于模型的强化学习（MBRL）直接应用于原始传感器数据，往往会面临训练不稳定、长期展开期间出现分布漂移以及缺乏显式状态信息等挑战。**Raw2Drive** [28] 通过提出一种双流基于模型的强化学习框架来应对这些问题。该框架首先利用特权信息训练世界模型和规划器，以建立稳定先验。随后，它采用对齐机制，使原始传感器分支（仅使用多视角图像进行训练）与特权分支同步。这一范式确保从原始输入中学习到的策略能够继承特权知识的鲁棒性，从而在闭环端到端配置中实现更强的泛化能力。为进一步使潜在表示与规划目标对齐，**WorldRFT** [27]引入了“强化微调”（RFT）。它利用结合碰撞感知奖励的组相对策略优化（GRPO） [69] ，对世界模型特征进行优化，使其更侧重于规划；与纯监督学习相比，此举显著降低了碰撞率。

C. 生成式数据合成与场景编辑

为缓解自动驾驶中的长尾分布挑战，该类别的研究利用生成模型合成罕见的安全关键场景或编辑现有传感器数据，从而为训练数据集增添多样、可控且符合物理规律的环境。

现实世界的自动驾驶数据集通常存在严重的类别不平衡，主要包含“晴朗白天”等常见条件，而“雨夜”等具有挑战性的场景则代表不足。为纠正这种导致性能下降的偏差，**SynDiff-AD** [55] 提出了一种合成数据增强流程，利用潜在扩散模型 [70]。该方法采用 ControlNet [71]保持与原始标注结构的语义一致性，并引入一种“针对子群”的提示方案，旨在生成具有特定语义密度和环境条件的照片级真实感图像。这一策略生成了天然带标签的合成数据，显著提升了分割模型和端到端驾驶模型在长尾条件下的鲁棒性。

尽管自动驾驶系统的效能依赖于海量标注数据，但真实世界数据采集的成本仍高得令人难以承受。此外，单纯扩大生成数据的规模并不能保证下游性能提升；关键决定因素在于可控多样性。**SubjectDrive** [54] 通过引入主体控制机制来应对这一需求，使生成模型能够整合多样化的外部数据源。这有助于对主体与内容维度进行精确控制，从而持续生成多样、实用且可自由标注的合成训练数据。该研究系统地验证了，提升生成数据多样性对于增强下游感知模型至关重要。

合成三维传感器数据还面临着其他独特挑战：从头生成激光雷达点云往往会造成显著的域差距，而新视角合成在生成反事实布局方面能力有限，例如重新安置车辆或改变物体配置，因此难以用于因果分析。**LiDAR-EDIT** [53] 通过先利用 VQ-VAE [72]将点云压缩为离散鸟瞰图潜在令牌，克服了这些限制。在这一结构化潜在空间中，它采用类似 MaskGIT 的 [73] 迭代预测机制，重建被操纵物体遮挡的背景区域。这一设计能够在保持几何连贯性和背景一致性的同时，对物体布局进行可控操纵，从而支持逼真且有意义的场景编辑。

高风险交互数据同样稀缺，却对鲁棒评估至关重要。**Safety-Critical** [3]通过在潜在轨迹空间中引入引导扩散 [74] 来弥补这一不足。在采样阶段施加可微的安全关键约束和对抗目标后，该模型能够合成具有方向性且符合物理规律的多车交互演变。这为生成高风险长尾交通场景提供了鲁棒的解决方案，同时在闭环仿真和评估中保持行为保真度。

D. 认知推理与潜在思维链

这一新兴范式将大型视觉语言模型（VLM）的语义推理能力融入自动驾驶技术栈。然而，该领域正迅速从基于视觉语言模型的被动解释转向主动的思维链（CoT）推理，使自主智能体能够从直觉式的“系统1”反应者（快速、由感知驱动的响应）转变为审慎式的“系统2”思考者（缓慢、基于推断的规划） [75]，通过逻辑推理进行规划。尽管自然语言具有可解释性，但对于实时控制而言，其计算效率往往较低。为解决这一瓶颈，**LCDrive** [62] 率先提出“潜在思维链”概念，以一系列动作对齐潜在令牌取代文本推理。通过将动作提议令牌与世界模型展开令牌交错处理，该方法在作出决策前，先在紧凑的潜在空间中推理未来结果。类似地，**FutureX** [61] 引入了动态“自动思考切换”，允许智能体自主决定是进入计算密集型的“思考模式”（执行潜在思维链展开，以针对复杂场景优化轨迹），还是在日常驾驶中保持“即时模式”。

为弥合高层语义与底层控制之间的差距，近期框架采用了层次化抽象和严格的落地机制。**MindDrive** [63] 建立了一个“假设情境”推理循环，由世界动作模型生成具有前瞻性的候选方案，再由面向视觉语言模型的评估器从安全性和舒适性维度进行严格评判。**ColaVLA** [57]则进一步通过认知潜变量，将场景理解压缩为“元动作嵌入”（例如“让行”“绕行”）

推理器，使后续规划器能够高效生成因果一致性轨迹。关键在于，要确保此类高层推理忠实于传感器输入，就需要主动对齐。**OmniDrive-R1** [64] 通过强化学习驱动的交错式多模态思维链来解决幻觉问题。它不依赖密集监督，而是采用“基于过程的落地奖励”（Clip-GRPO），确保视觉语言模型的文本推理与其视觉注意力焦点实时保持一致。这一方法与Alpamayo-R1 [56], 中的“因果链”框架相辅相成；后者通过强化学习优化推理轨迹，确保生成的规划在逻辑上以物理因果分析为基础。

对于资源受限部署，将推理能力蒸馏到更小的模型中仍然至关重要。**VERDI** [66]在训练期间将视觉语言模型的推理能力嵌入端到端模型的潜在空间，从而实现结构化常识推理，且不会产生推理时开销。不同于纯文本逻辑，**FutureSightDrive** [60]提出了“视觉思维链”，将生成的未来图像帧用作中间推理步骤，以可视化潜在结果。最后，为增强对无关扰动的鲁棒性，**CRiTIC** [58] 结合了一个带有注意力门控的因果发现网络，过滤掉不应影响自车决策的非因果代理。

III. 内部机制：潜在表示中的结构、对齐与动力学

潜在表示的内部机制构成了自动驾驶世界模型的骨干，决定了潜在展开在闭环交互下能否保持物理一致性、时间稳定性并与决策相关。本节不将感知、预测、规划和推理视为彼此独立的模块，而是提炼出**五种跨领域机制**，它们在不同范式中反复出现，并在很大程度上决定了鲁棒性、泛化能力和可部署性。

如图3所总结，本节将这些机制归纳为：**（A）用于实现几何与拓扑一致的潜在结构的**结构同构与几何先验；**（B）时间动态与长时域稳定性**，以减轻递归展开过程中的误差累积；**（C）语义与推理对齐**，以将潜变量建立在可迁移的抽象之上；**（D）价值对齐目标与后训练**，以将潜在展开与安全性及效用相关结果相结合；以及**（E）自适应计算与审慎推理**，以在不确定性和系统约束下分配展开深度。

A. 结构同构与几何先验 *iors*

现代世界模型的一个根本性范式转变，是从非结构化、低维的潜在瓶颈向保持几何保真度的空间同构表示的状态转移。与早期将场景压缩为抽象一维向量的方法不同，近期框架严格强制空间对应在潜在空间内保留驾驶环境的物理结构。这种结构一致性主要由**BEVWorld** [29]， 率先提出，它将潜在空间组织为显式的鸟瞰图网格，而非扁平的向量嵌入。该设计构成强归纳偏置，使解码器能够利用光线投射渲染进行自监督学习，从而确保重建观测与物理世界保持严格的几何一致性。将这种同构性扩展到体积域，**Drive-OccWorld** [37]在潜在占用记忆中纳入语义与运动条件归一化，以通过四维占用流预测占用体素的时空演化。

除单模态网格之外，结构同构近年来已成为多传感器对齐的重要桥梁。**OmniGen** [2] 通过将彼此分离的激光雷达点云和多视角相机特征投影到统一的“共享鸟瞰图空间”中，展示了这一点。通过强制采用共享几何体积，该模型借助体积渲染解码器确保生成的三维结构与二维纹理在空间上保持一致，从而有效避免非耦合生成模型中固有的几何冲突。尽管此类基于网格的布局能够提供密集的空间覆盖范围，研究人员还进一步融入拓扑约束，以更好地建模智能体与环境交互。**GenAD** [46],例如，通过地图感知实例令牌引入了以“结构化潜在空间”为形式的框架。通过将潜变量与特定的地图元素和动态智能体建立联系，该框架确保采样轨迹遵循道路网络的拓扑结构，并将物理可行性直接嵌入生成式规划过程中。

为进一步约束生成解空间并缓解结构幻觉，研究人员日益倾向于将显式几何先验注入潜在表示。**CoGen** [36]表明，高保真度三维几何条件对于实现一致的视频合成至关重要，能够引导去噪轨迹超越简单的二维布局条件。秉持相似理念，**DriveScape** [38]通过双向调制变换器融合三维道路结构先验，以确保多视角一致性。最后，在点云操作领域，**LiDAR-EDIT** [53] 利用球形体素化作为几何约束，在编辑潜在令牌时确保正确的射影几何，从而防止合成输出出现结构失真。

B. 时间动态与长时域稳定性

除空间结构外，潜在表示的时间机制在很大程度上决定了世界模型能否支持长时域想象，而不致模糊、产生运动学不一致或结构幻觉。自回归推演中一种反复出现的失效模式是误差累积和时间坍塌：潜变量不确定性与分布偏移随时间累积，逐渐损害预测未来。因此，近期研究着重关注时间因子分解、记忆结构化更新，以及能够明确缩小训练—推理差距的训练目标。

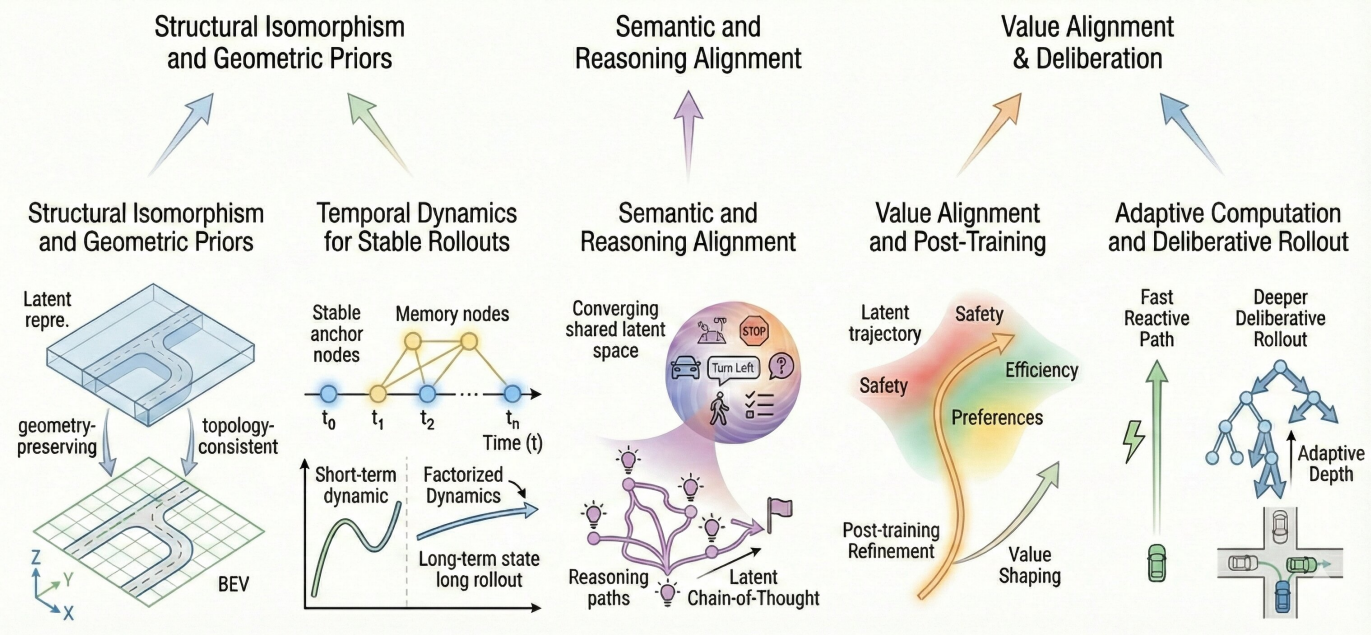

图3. **自动驾驶潜在世界模型的内部机制。**

若干方法通过构建潜在空间中的时间演化来提升长时域稳定性。**Epona** [31]主张采用解耦的时空因子分解，将随时间演化的动力学与细粒度未来生成分离，以减少自回归误差累积。**DriveWorld** [26]则通过记忆状态空间模型进一步明确时间结构，区分动态实体与静态场景内容（例如，动态记忆库与静态传播），从而施加不同的更新规则，更好地反映物理持久性。**LongDWM** [34]通过引入跨粒度蒸馏，利用细粒度运动线索约束粗粒度动力学，减轻长时间展开中的漂移，以应对短片段训练与长序列推理之间的不匹配。从在线生成的角度来看，**GLAD** [40] 采用结合潜变量传播的流式表述，将上一步的去噪潜变量作为下一帧的先验；这一设计有助于提升时间连贯性，同时减少对长窗口缓冲的需求。

时间鲁棒性还受到潜在流形选择（离散与连续，或混合）的影响。**Orbis** [32] 提供了受控证据，表明离散令牌动力学在长时域预测中可能较为脆弱，对令牌化细节敏感；而连续潜在动力学（例如基于流匹配的演化）往往通过避免量化伪影并强制更平滑的时间轨迹，产生更稳定的展开。为弥合这些机制，**LCDrive** [62] 提出混合交错动力学，在统一的令牌序列中对世界滚动预测和动作提议进行建模，既利用序列建模的表达能力，又保留结构化的时间依赖关系。

表示多模态性 并避免过度自信的自洽性，也会进一步影响长时域行为。**LatentDriver** [48] 对未来演变和自车动作建模为混合分布，从而支持多样化的假设，并减少长时域想象中的自我欺骗。当存在多个可行未来，且确定性滚动预测可能坍缩为不合理模式时，这种概率潜在动力学尤为重要。

干预与因果分析机制有助于诊断时间失效并揭示可操作的敏感性。**Exp-Latent** （探索潜在路径的简称） [59] 使用自动潜在扰动（ALP）来探测哪些潜在维度会影响控制输出，从而提供一种审视失效触发因素的工具。与此同时，**BRYANT** [43]强调频率感知时间解耦，将高频类噪声成分与低频趋势动力学分离，以提升时间演化的鲁棒性与可解释性。综合来看，这些观点表明，长时域能力并非仅取决于模型规模，还取决于显式设计的时间结构、流形设计、概率动力学，以及潜在空间中的诊断性干预。

C. 语义与推理对齐

为了赋予潜在空间认知推理能力和物理基础，现代框架越来越依赖超越简单重建目标的语义对齐机制。与仅根据原始传感器数据学习表征不同，**VERDI** [66] 等方法引入了一种推理蒸馏框架，在训练期间强制中间潜在特征与视觉语言模型（VLM）生成的文本嵌入对齐，从而有效地将结构化推理内化，且不会带来推理时开销。在此基础上，近期的“系统2”架构转而致力于使潜在动力学与显式推理轨迹对齐。**LCDrive** [62] 和**FutureX** [61] 进一步推进

了这一范式，引入潜在思维链（CoT），将推理表示为一系列动作对齐潜在令牌，这些令牌交织着动作提议与世界模型滚动推演，以模拟未来结果。为确保这些内部“思考”忠实于环境，**OmniDrive-R1** [64] 采用强化驱动的视觉定位机制（Clip-GRPO），强制模型的视觉注意力与文本推理保持跨模态一致性。此外，**Alpamayo-R1** [56], 则将潜在动作预测与“因果链”对齐，确保决策反映逻辑上的因果联系，而非仅仅是统计相关性。

除了语言知识，预训练视觉基础模型所蕴含的丰富视觉语义也提供了强大的归纳偏置。**VA-VAE** [65] 将视觉分词器的潜在空间直接与此类基础模型对齐，以“驯服”高维流形，从而解决重建保真度与生成质量之间的优化困境。类似地，**World4Drive** [52] 通过注入空间语义先验来丰富其表征，以实现意图感知规划。这种对齐策略如今正朝着任务级统一的方向扩展；例如，**DriveLaW**[44] 通过将生成器的潜在表示直接注入基于扩散的规划器，实现了视频生成与运动规划的统一。这确保了高保真度未来预测与可靠轨迹规划之间的内在一致性，显著缩小了预测与控制之间的差距。此外，**MindDrive** [63]和 **ColaVLA** [57] 利用元动作嵌入等语义抽象，将场景理解与多目标决策任务对齐。

在将强化学习应用于自动驾驶世界模型的背景下，使潜在表示与真实物理对齐，对于弥合现实差距至关重要。**Raw2Drive** [28] 通过特权对齐策略解决这一问题，使“原始传感器分支”与基于真实演员状态训练的“特权分支”保持同步。**WorldRFT** [27],进一步拓展了这一概念，引入利用 GRPO 的强化微调（RFT），使场景表示学习直接与安全关键规划目标对齐。通过轨迹高斯化和碰撞感知奖励，WorldRFT 优化潜在轨迹空间，对不安全行为进行惩罚，从而系统性地提升零样本安全性。最后，语义对齐促进了可控生成。**SynDiff-AD** [55]利用潜在扩散模型中固有的对齐能力，根据文本描述合成特定的极端案例；而 **OmniGen** [2] 则利用共享鸟瞰图空间，通过体渲染对齐激光雷达和摄像头生成，确保合成的多模态传感器数据在几何和语义上保持一致。

D. 价值对齐目标与后训练

潜在世界建模领域的一个显著趋势，是从以重建为中心的训练转向明确体现决策效用的目标。经典流程主要优化像素级重建或感知相似度（例如扩散去噪损失和 Fr´echet Inception Distance (FID) / Fr´echet Video Distance (FVD) 风格的代理指标 [76]），并隐含地假设，视觉上逼真的展开足以支持下游规划。在安全关键驾驶中，这一假设并不可靠：展开结果可能看起来清晰，却仍不适用于决策，例如导致碰撞、违反可行驶区域约束，或错误呈现交互式智能体行为。因此，近期系统会通过加入或替换纯感知目标，引入与价值对齐的信号，将潜在展开与规划及安全结果联系起来。

这一趋势的典型体现是强化学习式后训练方法：它们在任务级奖励下直接优化潜在表示。**WorldRFT** [27]引入强化微调，使场景表示学习与 Safety-Critical 规划目标相对齐，并利用碰撞感知奖励抑制潜在轨迹空间中的不安全行为。与此同时，以视觉-语言-动作（VLA）为核心的框架日益广泛地采用后训练，使内部推理与可靠的动作生成相对齐：**OmniDrive-R1** [64] 通过强化学习驱动的视觉定位，确保注意力与文本推理之间的跨模态一致性；而 **Alpamayo-R1** [56]则使潜在动作预测与“因果链”相对齐，以促进逻辑连贯的决策依据。价值对齐也体现在任务级统一工作中：用于预测的表征会直接提供给规划模块；例如，**DriveLaW** [44] 将生成器潜变量注入基于扩散的规划器，以减少预测–控制不匹配。

价值对齐训练会引入一些在感知指标下不易察觉的新失效模式。奖励规范可能稀疏或不完整，从而诱发奖励投机，或促使模型过度优化安全代理指标。后训练可能改变覆盖范围，使概率质量集中于保守行为，进而降低模型在复杂交互中的表现。这些观察表明，报告时应同时纳入感知质量和以决策为中心的结果，并将价值信号与可行性约束及校准不确定性相结合，以避免脆弱的过度自信。

E. 潜在展开轨迹中的自适应计算与审议

潜在世界模型日益不仅作为预测引擎，也作为规划时审议的计算基底发挥作用。驾驶需要异构推理模式：许多情况允许快速的反应式控制，而罕见或高度交互的场景则受益于更深的前瞻、反事实评估和多假设推理。潜在展开通过提供一个紧凑空间来支撑这一谱系，未来假设可在其中生成并评分，而近期的VLA/世界模型系统表明，此类审慎推理的深度与广度应当是自适应的，而非固定的。

具体实现见于近期的“推理在环”设计，这类设计将动作提议与世界模型想象交织在一起。**LCDrive** [62]和

FutureX [61]通过将审慎推理表示为一系列动作对齐潜在令牌，使潜在思维链得以落地。这些令牌在提出动作与推演其后果之间交替，从而能够在做出控制决策前进行迭代优化。与之相关，**DriveLaW** [44] 等任务级统一方法通过向规划器提供预测潜变量来支持审慎规划；而**MindDrive** [63] 和**ColaVLA** [57] 中的语义抽象机制则提供了更高层次的潜在接口，有助于进行多目标评估和候选选择。在这些设计中，自适应计算通常通过调整展开时域、分支因子或迭代优化次数来实现，其触发条件则根据不确定性、风险指标或采样未来状态之间的分歧确定。

自适应审慎推理带来了评估与系统方面的问题，而许多实验方案对此仍缺乏明确规定。只有在满足延迟、内存和功耗预算的前提下，更深入的推理才能提升安全性；反之，过于激进的提前退出启发式方法可能恰恰会在罕见的安全关键情况下失效。因此，报告应将任务指标与计算预算（例如每帧毫秒数、内存占用、展开步数和分支因子）结合起来，并描述计算资源分配对失败率的影响。实际部署很可能需要不确定性感知触发机制、安全监控器，以及在审慎推理被截断或潜在展开在分布偏移下变得不可靠时能够平稳启用的回退行为。

IV. 评估标准：指标与基准

A. 开环保真度与闭环稳定性

尽管高保真感官重建是内部模拟的基础，但世界模型的最终价值取决于其支持安全、稳健决策的能力。对这一能力的评估已从静态几何比较，发展为动态的交互感知评估，用以考察模型在不确定性和误差累积情况下的行为。

图 4 展示了开环评估与闭环评估之间的根本区别。开环测试在冻结、非反应式环境中，使用离线单步指标评估预测质量，例如平均位移误差（ADE）、Fr´echet Inception Distance（FID）、Fr´echet Video Distance（FVD）和一致性。相比之下，闭环评估通过在线多步展开，在具有反应性和交互性的世界中衡量策略质量，关注成功率、碰撞频率、规则违反和驶离道路事件等任务层面的结果。预测保真度与交互鲁棒性之间的张力，构成了潜在世界模型基准评测的核心权衡。

评估最普遍、最常见的切入点仍然是开环评估，即模型根据记录历史预测未来轨迹。在诸如**nuScenes** [77] 等数据集上，标准基准利用平均位移误差和碰撞率等指标，量化与人类专家示范的偏差。基线方法包括**GenAD** [46]

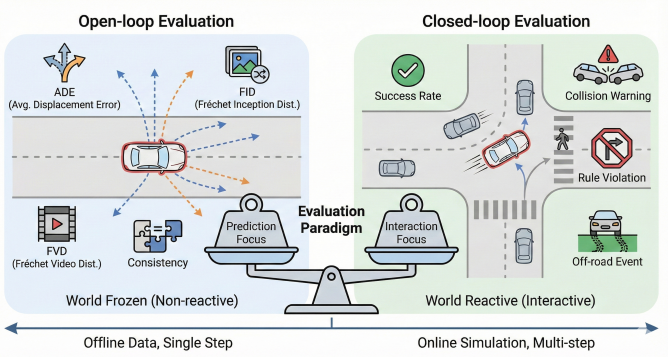

图 4. **自动驾驶中潜在世界模型的评估范式**：开环评估（左）和闭环评估（右）。

以及**DriveWorld** [26]，依靠这些确定性度量来验证所学表征的质量。近期进展，例如**LAW** [49]，表明，与从头训练的规划器相比，自监督潜在预训练显著降低了这些位移误差。尽管开环评估无处不在，但它引入了一个众所周知的关键“因果鸿沟”。由于自车状态在每个时间戳都会被重置为真实值，系统永远不会经历自身预测误差或偏差所带来的后果。这掩盖了“误差累积”效应，即早期的微小漂移会将智能体推入分布外状态，而这种失效模式在系统部署到主动环境之前一直不可见 [4]。

为克服这一局限，评估已转向在 CARLA 等基于物理的交互式模拟器中采用闭环协议，在这种环境下，世界模型必须充当递归策略。在此设置中，成功率（SR）和路线完成度（RC）等生存指标成为主要指标。**Think2Drive** [4] 凸显了不同评估范式之间的显著差异：研究报告称，开环误差（如平均位移误差）相近的模型，在复杂城市场景中的成功率却可能大相径庭，范围从 20 % 到 100%。为刻画单纯生存之外的驾驶质量细微差异，**Raw2Drive** [28]采用综合驾驶评分（DS），对闯红灯和驶离道路等违规行为进行惩罚；这证明，使潜在状态与特权物理知识相一致，对于遵守交通规则而不仅仅是避开障碍物至关重要。

除了成功导航之外，规划轨迹的时间稳定性也是衡量潜在空间鲁棒性的关键指标。端到端模型往往会出现“控制抖动”：即使环境稳定，规划也会在连续帧之间剧烈振荡。针对这一问题，**MomAD** [50] 提出轨迹预测一致性（TPC）指标，以惩罚帧间不一致，从而确保潜在规划过程具备时间平滑性。作为对这一确定性视角的补充，**LatentDriver** [48]主张采用基于负对数似然（NLL）的概率评估，确保模型能够捕捉未来可能性的多模态特性，例如选择礼让

而非绕行，而不是将其简化为单一且可能不安全的平均轨迹。

先进的评估协议如今深入探究驱动这些决策的因果机制，尤其是在安全关键和长尾场景中。为区分真正的因果推理与虚假相关性，**BRYANT** [43] 在频域中分析潜在空间，证明滤除高频噪声能够揭示场景中稳定的因果驱动因素。同样，**CRiTIC** [58]利用因果发现指标，验证规划器的注意力是否聚焦于因果相关智能体，而非背景干扰项。在鲁棒性方面，**Alpamayo-R1** [56] 和 **Safety-Critical** [3] 通过对抗性扰动和非结构化地形对模型进行压力测试，并使用 越野率 量化其泛化能力，这是标准城市场景基准无法衡量的。最后，为弥合保真度—交互性权衡，像 **NAVSIM** [31], [52]这样的混合数据驱动仿真环境，如今可通过在真实世界数据上进行日志回放展开，提供大规模类闭环验证；这既减轻了完全合成引擎造成的仿真到现实差距，又保留了交互性。

B. 基准测试与仿真环境

自动驾驶的可靠性需要多样化的评估平台生态，涵盖用于表征学习的大规模静态数据集，以及用于策略验证的反应式模拟器。表II根据保真度与交互性的权衡，将主流环境（代表性数据集和仿真平台）分为三类：**(1)静态真实世界数据集**，用于开环感知和预测；**(2)交互式仿真平台**，用于闭环策略验证；以及 **(3) 数据驱动神经仿真平台**，用于混合日志仿真。

nuScenes数据集是开环评估的标准测试平台。其多样化的传感器套件（激光雷达、摄像头、雷达）和丰富的三维标注，使其成为**GenAD** [46]和**DriveWorld** [26] 等基线方法评估预测准确度（例如，L2 误差）和生成保真度（例如，FID）的首选数据集。诸如**nuScenes-Occupancy** [78] 之类的扩展，进一步支持了对**Drive-OccWorld** [37]等方法进行基于体素的评估。为验证缩放定律，**Waymo开放数据集（WOD）** [79],凭借显著更长的驾驶时长，被**GAIA-2** [39] 等基础模型采用，以确保在不同地理区域中的鲁棒性。

为评估闭环决策，学界主要依赖**CARLA**这一基于虚幻引擎的高保真开源仿真器。它是基于模型的强化学习的主要实验平台。值得注意的是，近期的**Think2Drive** [4]和**Raw2Drive** [28] 等工作，均以具有挑战性的**CARLA Leaderboard2.0** [80], 为目标；该基准引入了长距离路线和对抗性场景，以压力测试智能体的因果推理能力。尽管 CARLA 提供完美的真实物理，但与现实世界的传感器数据流相比，它存在视觉域差距。

于2025年兴起的基准（如**NAVSIM** [81]）弥合了静态日志回放与合成引擎。NAVSIM通过允许自车在真实世界日志中偏离预录轨迹，并借助几何投影合成新视角，从而实现“对数模拟”。该范式将真实数据的照片级真实感与闭环控制的交互性相结合。包括**Epona** [31], **World4Drive** [52], 和**LAW** [49], 在内的前沿框架利用NAVSIM验证：世界模型可以直接从原始传感器数据学习可泛化策略，而无需传统模拟器中那些“游戏化”的伪影。

除 CARLA 外，LGSVL、 [82], AirSim、 [83], MetaDrive、 [84], 和 Waymax [85] 等其他交互式仿真平台也可用于评估策略在不同保真度和可扩展性约束下的鲁棒性，具体总结见表 II。

C. 迈向统一的以潜在表示为中心的评估指标

现有基准通常将开环预测保真度与闭环安全评估分离开来 [11], [86],[87]。尽管 nuScenes [77] 和 WOD[88]等数据集对开环指标进行了标准化，但感知相似度与交互鲁棒性之间的差异仍被广泛观察到。为此，本文提出了三个互补的统一指标，旨在弥合潜在表示质量与基于部署行为的决策安全性之间的鸿沟。

a) 闭环安全差距（*CSG*）：为量化视觉保真度与交互安全性之间的不匹配，本研究定义了闭环安全差距：

$$\mathrm{CSG} = F_{OL} - S_{CL}, \tag{1}$$

其中， $F_{OL} \in [0,1]$ 表示归一化的开环保真度分数（例如，平均位移误差的倒数或窗口化FVD）， S_{CL}表示闭环安全分数，定义为

$$S_{CL} = 1 - \frac{\text{Collision Rate}}{\text{Route Complexity}}. \tag{2}$$

较大的正值 CSG 表明，视觉上看似合理的预测未能转化为安全的闭环执行，凸显了长期存在的预测—交互鸿沟。

b) 时间连贯性分数（*TCS*）：稳定的决策还要求轨迹在连续帧之间平滑演变。受先前关于控制抖动和概率稳定性的分析启发 [89], ，本研究提出并定义了一种归一化的时间连贯性分数：

$$\mathrm{TCS} = 1 - \frac{\mathrm{Var}(\Delta_t)}{\mathrm{Var}_{\max}}, \tag{3}$$

其中， Δ_t 表示轨迹导数在相邻帧之间的差异。较高的 TCS 值表示时间振荡减少，规划稳定性提升，且规划动态更加平滑、稳定。

c) 审慎推理成本（*DC*）：随着世界模型融入多步想象和分支展开，评估还必须考虑实时约束下的计算可行性 [12]。为系统性地基准测试认知深度与计算可行性之间的权衡，本研究将审慎推理成本（DC）定义为一种归一化的

表II评估生态系统（面向潜在世界模型）：代表性数据集与仿真平台。

基准数据集或平台	Type	主要评估重点	优点（+）与缺点（-）	代表性方法
类别1：静态真实世界数据集（开环）				
nuScenes	真实世界日志	感知、生成（FID）、预测（L2）	+ 多模态标准；丰富的标注 -因果鸿沟（无交互）	GenAD [46], DriveWorld [26], Drive-OccWorld [37]
Waymo开放数据集（WOD）	真实世界日志	大规模扩展、长时域预测	+ 基础模型所需的大量数据 - 计算成本高	GAIA-2 [39], LatentDriver [48]
类别 2：交互式仿真平台（闭环）				
CARLA（LB 2.0）	物理引擎	规划策略、安全性、规则遵循	+ 完全交互式；动力学可复现 -视觉仿真到现实差距	Think2Drive [4], Raw2Drive [28], MomAD [50],Alpamayo-R1 [56]
LGSVL（SVL模拟器）	物理引擎	完整自动驾驶技术栈集成	+高保真度传感器模型；ROS兼容性 -高度依赖系统	–
AirSim	物理引擎	感知 + 控制仿真	{v1>照片级真实感渲染；多模态传感器-有限的真实交通交互	–
MetaDrive	轻量级仿真器	RL式策略评估	+训练速度快且可扩展 - 简化的视觉真实感	–
Waymax	基于场景的仿真	日志条件闭环推演	+ 大规模场景回放 - 物理动力学建模有限	LatentDriver [48],AWMs [90]
类别 3：数据驱动神经仿真平台（混合）				
NAVSIM / OpenDV	对数模拟	可泛化策略，端到端规划	+真实视觉效果 +闭环自由度 -视点受传感器边界限制	World4Drive [52], Epona [31],LAW [49]

效率指标，用于衡量获得边际安全性提升所需的资源占用：

$$\mathrm{DC} = \frac{w_T \widehat{T} + w_B \widehat{B} + w_E \widehat{E} + w_M \widehat{M}}{\widehat{\Delta G}}. \tag{4}$$

分子概括了审慎推理的计算特征。这里， T 表示推理延迟，使用尾部统计量（例如，毫秒/帧）进行测量，以反映实时运行的最坏情况界限； B 是平均分支因子，用于表示并行假设或展开的广度； E 表示每次决策消耗的能耗（焦耳/次决策）； M 表示工作内存占用（MB）。分母 ΔG表示相较于纯反应式（零审慎推理）基线，审慎推理所带来的预期安全收益。

为确保尺度不变性，带帽变量表示已按硬件和平台特定的参考 $\widehat{\ }$ $\widehat{\ }$预算进行归一化，例如 $T = T/T_{\mathrm{ref}}$和 $\Delta G = \Delta G/\Delta G_{\mathrm{ref}}$。非负权重 $w_{\{T,B,E,M\}}$ 用于编码部署优先级；例如，在高速场景中，可以通过为 w_T赋予更大的值来提高对延迟的重视程度。较低的 DC 值表示更高效的审慎推理，从而推动评估同时考虑安全性收益和可在边缘设备部署的资源约束。

V. 前沿与瓶颈：潜在世界建模面临的持续挑战

尽管用于自动驾驶的潜在世界模型发展迅速，但要将强劲的基准测试结果转化为稳健的现实世界行为，仍然面临挑战。第四节 中回顾的评估标准明确揭示了这一差距：开环保真度和短时域准确性往往无法预测长时域展开、领域偏移和交互反馈条件下的闭环安全性。因此，许多失败并不仅仅是工程产物，而是源于潜在表示如何随时间演化、想象过程中如何分配计算，以及如何在不确定性下对决策进行 grounded 和验证。

图 5 总结了构成本节框架的五个反复出现的瓶颈。A面板突出显示了由累积展开推演误差导致的长时域幻觉和漂移；B面板体现了实时部署约束，这些约束在车载延迟、内存和功耗预算的限制下，制约了审慎推理深度。C面板强调了持续存在的仿真到现实差距和跨域泛化差距，而D面板则指出有限的可解释性和薄弱的因果推理使诊断与

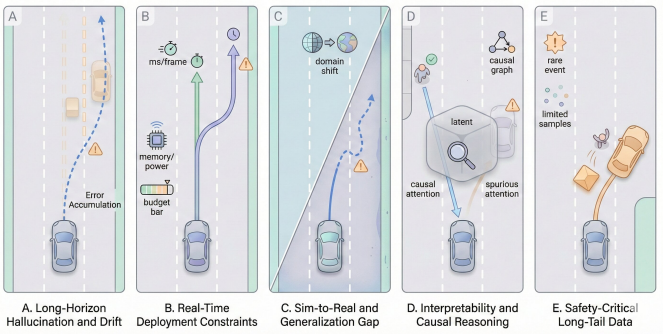

图5。 **自动驾驶中潜在世界模型的开放挑战。**

验证更加复杂。E面板进一步强调了安全关键长尾交互的稀缺性，促使人们在分布内基准测试之外开展压力测试与评估。以下小节将逐一讨论各个瓶颈，并总结具有代表性的缓解方向。

A. 长时域一致性与幻觉困境

尽管近期生成式驾驶世界模型在短时域视觉保真度方面取得了长足进展，但长时域展开仍面临稳定性退化和幻觉带来的根本性挑战。随着展开长度增加，潜在状态误差的累积以及分布偏移会逐步放大，导致渲染模糊、运动学漂移，更严重的是，还会违反物理和几何约束（例如物体自发消失、多视角不一致，或自车/智能体状态在动力学上不合理）。既有研究已从互补视角出发，开始阐明这些失败的成因及潜在的缓解策略。**Orbis** [32] 提供了受控证据，表明离散标记化在长时域预测中可能较为脆弱，对分词器设计表现出显著敏感性；相比之下，连续潜在动力学（例如流匹配）往往能够产生更加稳定的展开。**LongDWM** [34] 将长时域退化归因于短片段监督与长序列自回归部署之间的训练与推理不匹配，并通过跨粒度动态蒸馏缓解漂移。**Epona** [31] 通过解耦的时空因子分解和前向链式训练机制应对自回归误差累积，并进一步证实了长时域

世界模型在规划基准上的决策效用。总的来说，这些结果凸显了一个更广泛的开放问题：当前领域仍缺乏针对长时域稳定性的统一、可解释且与闭环相关的评估协议。一个关键方向是，在长时域感知指标（例如窗口化FVD）、几何一致性、物理可行性与闭环失败率（碰撞/驶出道路/险些碰撞）之间建立有原则的对应关系，从而推动世界模型从仅具备视觉上的合理性，发展为能够服务于决策的模型。

B. 实时效率与部署

尽管基于扩散的生成器和大型Transformer世界模型进展迅速，实时部署仍是自动驾驶的主要瓶颈：最先进的生成式滚动推演往往计算密集、内存需求高，并且难以在严格的车载延迟和功耗预算内执行（例如，汽车级片上系统（SoC））。近期工作探索了互补方向，但重要局限依然存在。

GLAD [40]通过逐帧流式处理以及潜变量传播与缓存，将扩散式驾驶视频生成推向在线使用，从而减少缓冲完整序列的需要，并实现无限长度合成。然而，其流式表述仍以迭代去噪为基础，若没有激进加速（蒸馏、少步采样、量化）和审慎的系统集成，对于硬实时车载推理可能仍然难以承受；此外，仅靠流式生成并不能保证交互反馈下的闭环可控性或安全性。

EGADS [45]主要从训练角度解决效率问题，借助潜在空间信念建模和示范增强强化学习，针对城市驾驶中的样本复杂度与泛化。这可以降低数据和训练成本，但并不能直接解决部署时的推理时约束（延迟/吞吐量），而且其对结构化奖励和示范信号的依赖，可能会限制在目标、传感器套件或驾驶风格不同的领域之间的可扩展性。

Think2Drive [4] 通过将紧凑的潜在世界模型用作神经模拟器来执行并行滚动推演，从而提升闭环学习效率，并加速复杂基准中的策略优化。该方法仍需承担维护高容量世界模型和规划器的开销，部署可行性取决于组合技术栈能否满足实时约束，并在分布偏移下保持保真度；以模拟器为中心的训练还可能留下残余的仿真到现实差距，并需要稳健的安全监控器以供上路使用。

在系统级方法方面，仍缺乏能够联合优化以下因素的统一方案：(i) 展开/模型保真度，(ii) 推理延迟和内存占用，以及 (iii) 交互反馈下的闭环安全性。值得探索的方向包括低步数生成式滚动推演（蒸馏和一致性类目标）、硬件感知压缩（量化/剪枝）、模块化调度（采用安全关键型回退机制的异步世界模型更新），以及在现实的汽车领域约束下，同时报告任务性能和实时预算（毫秒/帧、内存、功耗）的基准。

C. 仿真到现实差距与泛化

生成式世界模型和端到端驾驶策略往往在分布内表现强劲，但在分布偏移下会急剧下降，从而形成持续存在的仿真到现实与跨域泛化差距。在特定日志分布（例如特定城市、传感器套件或天气特征）上训练的模型，在部署到道路拓扑、交通规范 [91],感知特性或长尾交互不同的区域时可能会失效。这一困难源于多种相互纠缠的偏移来源：仿真中的渲染与动力学不匹配、传感器噪声与校准差异、其他智能体的行为异质性，以及数据集特定偏差，因此单一基准测试上的改进很少能转化为稳健的域外安全性。

近期研究探索了多种互补的解决方案，但每种方案都有明显局限。**Raw2Drive** [28] 强调特权对齐，将训练时的特权信号（可在仿真中获取，或通过更丰富的标注获得）提炼到基于原始观测学习的表征中，以此弥合模态差距和域差距。这一策略能够提高学习过程对外观变化的稳定性，但其对特权监督的依赖引发了可移植性方面的担忧：特权线索可能无法获取、前后不一致，或与现实世界结构匹配不佳；而对齐目标也可能对设计选择高度敏感，从而降低跨域可复现性。**DriveX** [30]强调可泛化的世界知识，主张通过扩大规模并设计能够促进可迁移结构、而非数据集特定记忆的训练目标来实现这一点。然而，仅凭这种规模扩张趋势，并不能证明所获取的知识具有因果性或与安全相关。泛化方面的主张往往仅得到代理指标和有限跨域协议的支持，而由此产生的模型通常更庞大，也更难在实时约束下部署。**World4Drive** [52] 引入基础模型先验以提升域外鲁棒性，利用更强的预训练表征来降低对狭窄训练分布的依赖。这种优势通常主要体现在感知和表征层面，而在行为和规则变化下，交互层面的决策和闭环安全性仍可能失效。此外，融入大规模预训练先验会使归因和验证更加复杂，也让系统级安全论证更具挑战性。

这一领域的进步需要采用能够明确考验域外条件（新城市、传感器配置、罕见交互）的评估协议，并在报告开环准确率的同时报告闭环失败统计。从方法上看，值得探索的方向包括在潜在空间中将外观与几何和动力学分离，为交互建模学习域不变因果因子，

以及设计在分布偏移下可证明安全的适应或校准机制（例如不确定性感知回退策略和安全监控器）。如果没有此类协议和系统方面的进展，生成保真度或分布内规划精度的提升不太可能转化为可靠的真实世界部署。

D. 可解释性与因果推理

基于学习的驾驶策略和世界模型仍然难以审计：优异的性能可能与虚假线索、脆弱注意力模式以及不忠实于实际决策机制的事后叙述并存。可解释性方法开始揭示内部敏感性，但要建立具有因果性且与动作相关的解释，并使其能够预测分布偏移下的失效模式，仍然是一个开放性挑战。

Exp-Latent [59]通过 ALP 推进了机制可解释性，探究哪些潜在维度驱动控制输出，并支持受控的假设分析。然而，由此产生的解释往往是模型内部的：潜在轴可能无法与物理上可干预的因素（例如智能体位置或道路几何）对齐，可能在不同训练运行之间不稳定，并且通常只能提供局部敏感性，而非闭环保证。**CRiTIC** [58] 引入了一种由因果发现驱动的注意力门控机制，用于多智能体预测，促使注意力跟踪具有因果相关性的实体，并提高对非因果干扰的鲁棒性。其有效性依赖于来自观测数据的强可识别性假设，仍容易受到隐藏混杂因素和部分可观测性的影响，也无法直接转化为规划与控制的因果保证。
Alpamayo-R1 [56] 通过“因果链”将自然语言因果推理与轨迹输出相结合，朝着面向决策的解释迈进，为实现人类可审计的驾驶依据提供了一条路径。然而，该方法仍面临忠实性风险（看似合理的合理化解释与动作的真正驱动因素之间可能存在差异）、依赖学习型评判器时的评估偏差，以及对长尾因果结构覆盖有限等问题。

要取得进展，需要采用能够评估解释忠实性（即解释是否能在干预下预测动作变化）的评估协议，将解释与闭环安全结果关联起来，并支持在具有物理意义的变量层面开展反事实分析。方法论方向包括解耦且有依据的潜在因素、通过显式干预进行因果表征学习，以及利用因果结构预测故障而非仅在事后描述故障的安全监控器。

E. Safety-Critical 与长尾数据

安全关键事件（碰撞、险些碰撞、极端加塞、罕见的规则违反）在真实世界驾驶日志中极为稀少，导致训练和评估都面临严重的数据不平衡。因此，基于常规驾驶数据拟合的模型可能无法应对罕见交互，而标准基准测试往往对主要风险模式的采样不足。

生成式增强是一种自然的应对方式，但也引发了有关覆盖范围、真实性和安全相关性的棘手问题。

Safety-Critical [3] 通过引导将潜在扩散模型引向危险工况。其关键局限在于有效性：引导可能会优化“危险”的代理概念，并生成在物理或行为上不合理的场景，因此在合成分布上的收益未必能降低现实世界的风险。覆盖范围仍不确定，因为引导可能会坍缩到极少数极端案例。

SynDiff-AD [55]合成针对性极端案例，以丰富长尾数据。其有效性取决于严格的质量控制、在避免不可行性的同时保持多样性、标签的正确性，以及与传感器物理规律和交通规则的一致性。如果缺乏过滤和闭环检查，合成数据可能引入伪影、损害校准或使规划器产生偏差；目前仍缺少用于衡量(i)长尾覆盖率、(ii)真实性/因果合理性，以及(iii)闭环安全影响的标准化协议。

更有力的基准测试应报告罕见事件风险（碰撞／驶出道路／险些碰撞），并支持场景级验证（规则遵循、物理可行性、多视角一致性）。安全约束生成、校准不确定性以及仿真器在环验证，很可能是将合成长尾数据转化为可衡量的闭环安全增益的先决条件。

VI. 未来展望

要推动潜在世界模型走向自动驾驶的可靠真实世界部署，需要在模型稳定性、评估对齐、表征结构和系统集成方面协同取得进展。尤其是，稳定长时域展开对于缓解分布偏移下的误差累积和幻觉至关重要。同时，协调开环预测指标与闭环交互性能，是缩小长期存在的指标—决策鸿沟的关键。最后，以可解释性和任务落地为目标构建潜在空间，有助于在规划流水线中实现可控且语义明确的推理。

图 6 总结了推动潜在世界模型发展、实现可用于决策且可在真实世界部署的自动驾驶的五个互补研究方向：(A) 有落地且结构化的潜在世界；(B) 具备闭环对齐的长时程推演稳定性；(C) 潜在动作和 VLA接口；(D) 仿真到现实泛化与适应；(E) 面向资源约束下实时部署的系统级优化。这些方向共同致力于构建世界模型，使其不仅具有丰富的感知表达能力，还能在交互式驾驶环境中做到因果落地、计算高效且安全性可验证。

A. 有落地且结构化的潜在世界

当前的潜在世界模型常常将外观、几何与动力学纠缠在一起，从而限制了分布偏移下的可控性与可靠性。一个有前景的方向是学习带有显式几何与运动学锚点（例如，鸟瞰图/占用基元）的因子化潜在状态

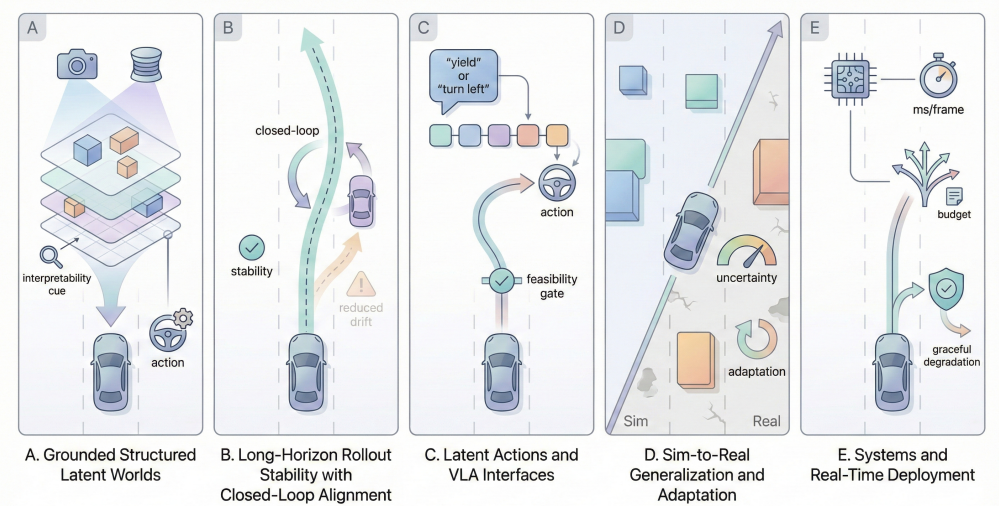

图 6. **自动驾驶中潜在世界模型的未来发展方向与研究方向示意图。**

同时保持高保真视图合成。结构化潜在表征能够实现具有物理意义的干预与规划时约束。关键挑战依然存在，包括因子的可辨识性、稳定的跨视图对齐，以及将潜在结构与闭环失效率联系起来的评估协议。

B. 通过闭环对齐实现长时域展开稳定性

长时域展开仍容易受到累积预测误差和生成式幻觉的影响；感知指标的改善并不一定意味着闭环行为更加安全。未来研究应通过多步监督、不确定性感知推演和仿真器在环验证，使训练目标与推演时约束相一致，并直接惩罚碰撞、驶离道路事件或险些碰撞等安全关键故障。亟待解决的问题包括稀有事件的可扩展评估、可靠的不确定性校准，以及防止生成视觉上看似合理、但物理上无效的内容。

C. 潜在动作与 *VLA*接口

要将潜在世界模型与 VLA 风格的决策连接起来，动作表征既要能够通过语言实现语义对齐，也要兼容控制器。未来工作可以探索混合动作空间，将用于语义对齐的离散词元与用于控制的连续潜变量相结合。在此类系统中，推理模块（例如基于大语言模型（LLM）/视觉语言模型（VLM）的规划器）可以生成候选动作，并通过潜在世界展开中的可行性检查进行验证。主要挑战包括在不同平台间统一动作词元表示，以及确保语言条件策略在对抗提示或分布外情境下仍保持稳健。

D. 仿真到现实泛化与适应

即使是表现强劲的分布内模型，在从仿真迁移到现实环境，或在运行于不同城市、不同传感器配置和不同交通规范下时，性能也会下降。要解决仿真到现实的鸿沟和域适应问题，需要学习域不变因果因子，制定有原则且附带明确有效性检查的特权到原始蒸馏策略，并引入由校准不确定性和安全监控器把关的测试时自适应机制。至关重要的是，这类自适应机制必须能够被验证为安全且可审计，而不能只提升基准测试的平均表现。

E. 系统与实时部署

大型扩散和基于 Transformer 的世界模型，在满足汽车领域的延迟、内存和功耗预算方面面临重大挑战。因此，实际部署需要模型层面和系统层面的多方面改进：(i) 高效的生成式滚动推演，例如采用少步数/减步数扩散采样器（通过蒸馏或一致性模型实现），以及结合周期性重新规划的短时域/截断式展开策略；(ii) 硬件感知的压缩与加速，例如量化、剪枝、混合精度和内核级优化；(iii) 模块化调度架构，在保障安全关键回退行为的同时，支持异步世界模型更新。此外，评估协议应在报告任务得分的同时，提供与部署相关的资源指标，包括延迟、内存占用、能耗/功耗（平均值和峰值）以及所用硬件平台。这些资源感知基准使研究人员不仅能够评估算法性能，还能评估其在现实世界中的可行性。

VII. 结论

本文综合近期进展，从三个关键视角建立了一个适用于自动驾驶世界模型的统一潜在空间框架：以潜在空间为中心的分类体系、对内部模型机制的分析，以及一套精简的评估方案，其中包括所提出的 CSG/TCS/DC 指标，以便更好地使开环预测

保真度与闭环安全性能相一致。分析揭示了感知真实感与闭环稳定性之间的根本张力：尽管当前模型能够生成视觉上可信的传感器观测，但在交互式控制环路中部署时，往往无法保证行为在物理上一致且与决策相关。本研究的五项核心贡献，即以潜在空间视角构建的分类体系、对内部机制的综合梳理、精简的评估方案、具有指导性的设计建议，以及经过筛选的基准和示例方法，共同为将潜在世界模型转化为实用的自动驾驶系统提供了结构化且面向部署的研究议程。

展望未来，结构化几何先验、语言锚定推理、大规模预训练结合高效微调/蒸馏、资源自适应计算以及严格的闭环评估相结合，将加速面向自动驾驶的通用型具身基础模型的发展。具体而言，这些模型和方法应在统一且紧凑的潜在空间中整合感知、预测与决策，并提供可验证的接口（经过校准的不确定性和反事实查询），同时支持兼具设计即安全特性的资源自适应审议，从而打造可用于决策的系统，有效缩小仿真到现实差距。归根结底，这些架构有望成为稳健的高阶自动驾驶系统的认知核心，助力在复杂的现实环境中实现更安全、更具适应性的自主运行。

参考文献

[1] X. Wang, Z. Zhu, G. Huang, X. Chen, J. Zhu, 和 J. Lu, 《DriveDreamer：面向自动驾驶的真实世界驾驶世界模型》，载于《计算机视觉——ECCV 2024》，A. Leonardis、E. Ricci、S. Roth、O. Russakovsky、T. Sattler 和 G. Varol 编。瑞士查姆：施普林格·自然瑞士，2025，第 55–72 页。[2] T. Tang、E. Ma、x. zhou、L. Wang、T. Yan、X. Zhang、K. Zhan、P. Jia、X. Lang、J.-W. Bian、K. Yu 和 X. Liang, 《OmniGen：面向自动驾驶的统一多模态传感器生成》，2025 年 12 月，arXiv:2512.14225 [cs]。 [在线]。可访问：http://arxiv.org/abs/2512. 14225[3]M. Peng、R. Yao、X. Guo、Y. Xie、X. Chen 和 J. Ma, 《采用引导潜扩散模型的安全关键交通仿真》，2025 年 5 月，arXiv:2505.00515 [cs]。 [在线]。可访问：http://arxiv.org/abs/2505 00515 。Q. Li、X. Jia、S. Wang 和 J. Yan, 《[4] Think2Drive：通过在潜在世界模型中思考实现高效强化学习，适用于准真实自动驾驶（CARLA-v2）》，2024 年 7 月，arXiv:2402.16720 [cs]。 [在线]。可访问：http://arxiv.org/abs/2402. 16720[5] J. Yang、K. Chitta、S. Gao、L. Chen、Y. Shao、X. Jia、H. Li、A. Geiger、X. Yue 和 L. Chen, 《Resim：面向自动驾驶的可靠世界仿真》，载于《神经信息处理系统进展》（NeurIPS），2025。[6] X. Hu、S. Li、T. Huang、B. Tang、R. Huai 和 L. Chen, 《仿真如何助力自动驾驶：仿真到现实、数字孪生和平行智能综述》，《IEEE智能车辆汇刊》，第 9 卷，第 1 期，第 593–612 页，2024。[7] L. Zhang、Y. Dong、H. Farah 和 B. van Arem, 《基于驾驶风险场和模型预测轮廓控制的自动驾驶车辆社会感知规划与控制：以穿越环岛为例》，载于《2023 IEEE系统、人与控制论国际会议（SMC）》，2023，第 3297–3304 页。[8] W. Schwarting、A. Pierson、J. Alonso-Mora、S. Karaman 和 D. Rus, 《自动驾驶车辆的社会行为》，《美国国家科学院院刊》，第 116, 50, 24 972 24 978, 卷。期。页 – 2019 [ ] 10 1073 。在线。可访问：https://www.pnas.org/doi/abs/ 。/pnas. 1820676116

[9] Y. Dong, B. van Arem, 和 H. Farah, “自动驾驶车辆社会感知规划与控制的进展、专家见解与概念框架”, Communications in Transportation Research，第5卷，第100207页，2025。 [在线]。可获取：https: //www.sciencedirect.com/science/article/pii/S2772424725000472

[10] F. Codevilla、M. Muller、A. Lopez、V. Koltun和A. Dosovitskiy，“通过条件模仿学习实现端到端驾驶”，arXiv预印本 arXiv:1710.02410，2017。[11] M. Bansal、A. Krizhevsky和A. Ogale，“Chauffeurnet：通过模仿最优并合成最差来学习驾驶”，arXiv预印本 arXiv:1812.03079，2018。[12] L. Chen、P. Wu、K. Chitta、B. Jaeger、A. Geiger和H. Li，“端到端自动驾驶：挑战与前沿”，IEEE模式分析与机器智能汇刊，2024。[13] F. Arasteh、M. Elmahgiubi、B. Khamidehi、H. Mirkhani、W. Zhang、C. Tongtong和K. Rezaee，“基于失效的有效性学习：缓解自动驾驶车辆规划中的分布偏移”，载于2025年IEEE国际机器人与自动化会议（ICRA）。IEEE，2025，第15 680–15 686页。[14] M. Hallgarten、J. Zapata、M. Stoll、K. Renz和A. Zell，“车辆运动规划能否泛化到真实的长尾场景？”，载于2024年IEEE/RSJ国际智能机器人与系统会议（IROS）。IEEE，2024，第5388–5395页。[15] J. Ge、Z. Liu、L. Fan、Y. Jiang、J. Su、Y. Li、Z. Zhang和S. Chen，“揭示合成数据对端到端自动驾驶的影响”，arXiv预印本 arXiv:2503.18108，2025。[16] T. M. Moerland、J. Broekens、A. Plaat和C. M. Jonker，“基于模型的强化学习：综述”，Found. Trends Mach. Learn.，第16卷，第1期，第1–118页，2023年1月。 [在线]。可获取：https://doi.org/10 1561 2200000086 。/ H. Yuan、P. Li、[17] B. Van Arem、L. Kang、H. Farah和Y. Dong，“基于深度强化学习的安全、高效、舒适性与节能的环岛自动驾驶”，载于2023年第26届IEEE智能交通系统国际会议（ITSC），2023，第6074–6079页。[18] Y. Dong、T. Datema、V. Wassenaar、J. Van de Weg、C. T. Kopar和H. Suleman，“面向多种模拟驾驶操作的自动驾驶深度强化学习综合训练与评估”，载于2023年第26届IEEE智能交通系统国际会议（ITSC），2023，第6165–6170页。[19] T. Yan、T. Tang、X. Gui、Y. Li、J. Zhesng、W. Huang、L. Kong、W. Han、X. Zhou、X. Zhang、Y. Zhan、K. Zhan、C. zhong Xu和J. Shen，“Ad-r1：采用无偏世界模型的端到端自动驾驶闭环强化学习”，ArXiv，第abs/2511 20325, 卷。. 2025 [ ] 在线。可获取：https://api.semanticscholar.org/CorpusID: 283250930

[20] D. Ha和J. Schmidhuber，“世界模型”，arXiv预印本 arXiv:1803.10122，2018。[21] D. Hafner、J. Pasukonis、J. Ba和T. Lillicrap，“借助世界模型掌握多样领域”，2024。 [在线]。可获取：https://arxiv.org/abs/2301 04104 。C. Gumbsch、[22] N. Sajid、G. Martius和M. V. Butz，“从离散潜在动力学中学习具有自适应时间抽象的分层世界模型”，载于第十二届国际学习表征会议，2023。[23] T. Saanum、P. Dayan和E. Schulz，“利用软状态不变世界模型简化潜在动力学”，神经信息处理系统进展，第37卷，第38 355–38 382页，2024。[24] A. van den Oord、O. Vinyals和K. Kavukcuoglu，“神经离散表示学习”，arXiv预印本 arXiv:1711.00937，2017。[25] A. Radford、J. W. Kim、C. Hallacy、A. Ramesh、G. Goh、S. Agarwal、G. Sastry、A. Askell、P. Mishkin、J. Clark、G. Krueger和I. Sutskever，“通过自然语言监督学习可迁移视觉模型”，arXiv预印本 arXiv:2103.00020，2021。[26] C. Min、D. Zhao、L. Xiao、J. Zhao、X. Xu、Z. Zhu、L. Jin、J. Li、Y. Guo、J. Xing、L. Jing、Y. Nie和B. Dai，“DriveWorld：面向自动驾驶的基于世界模型的4D预训练场景理解”，载于2024年IEEE/CVF计算机视觉与模式识别会议（CVPR）。美国华盛顿州西雅图：IEEE，2024年6月，第15 522–15 533页。[在线]。可获取：https://ieeexplore.ieee.org/document/10655336/ [27] P. Yang、B. Lu、Z. Xia、C. Han、Y. Gao、T. Zhang、K. Zhan、X. Lang、Y. Zheng和Q. Zhang，“WorldRFT：面向自动驾驶的基于强化微调的潜在世界模型规划”，2025年12月，arXiv:2512.19133 [cs]。 [在线]。可获取：http://arxiv.org/abs/2512 19133 。Z. Yang、X. Jia、[28] Q. Li、X. Yang、M. Yao和J. Yan，“Raw2Drive：面向端到端的对齐世界模型强化学习

自动驾驶（CARLA v 2 2025, 2505 16394 ）”，10月。[ ] [ ] 2505 arXiv：。cs。在线。可获取：http://arxiv.org/abs/16394
[29] Y. Zhang、S. Gong、K. Xiong、X. Ye、X. Li、X. Tan、F. Wang、J. Huang、H. Wu和H. Wang，“BEVWorld：基于场景级鸟瞰图潜变量的自动驾驶多模态世界模拟器”，2025年5月，arXiv:2407.05679 [cs]。 [在线]。可获取：http://arxiv.org/abs/2407 05679 。C. Shi、S. Shi、K. Sheng、B. Zhang和L. Ji
[30] ang，“DriveX：面向自动驾驶中可泛化世界知识学习的全场景建模”，2025年5月，arXiv:2505.19239 [cs]。 [在线]。可获取：http://arxiv.org/abs/2505 19239 。K. Zhang、Z. Tang、X. Hu、X. Pan、
[31] X. Guo、Y. Liu、J. Huang、L. Yuan、Q. Zhang、X.-X. Long、X. Cao和W. Yin，“Epona：面向自动驾驶的自回归扩散世界模型”，2025年6月，arXiv:2506.24113 [cs]。[在线]。可获取：http://arxiv.org/abs/ 2506.24113[32] A. Mousakhan、S. Mittal、S. Galesso、K. Farid和T. Brox，“Orbis：克服驾驶世界模型中长时域预测的挑战”，2025年7月，arXiv:2507.13162 [cs]。 [在线]。可获取：http://arxiv.org/abs/2507 13162 。Y. Lipman、R. T. Chen、H. Ben-Hamu、M.
[33] Nickel和M. Le，“面向生成建模的流匹配”，arXiv预印本 arXiv:2210 02747，2022。[34] X. Wang、Z. Wu和P. Peng，“LongDWM：通过跨粒度蒸馏构建长期驾驶世界模型”，2025年6月，arXiv:2506.01546 [cs]。 [在线]。可获取：http://arxiv.org/abs/ 2506.01546[35] R. Chen、Z. Wu、Y. Liu、Y. Guo、J. Ni、H. Xia和S. Xia，“UniMLVG：面向自动驾驶、具备全面控制能力的多视角长视频生成统一框架”，2025年3月，arXiv:2412.04842 [cs]。 [在线]。可获取：http://arxiv.org/abs/ 2412.04842[36] Y. Ji、Z. Zhu、Z. Zhu、K. Xiong、M. Lu、Z. Li、L. Zhou、H. Sun、B. Wang和T. Lu，“CoGen：面向自动驾驶的基于自适应条件化的3D一致性视频生成”，4月。[ ] [ ] 2503 2025, 2503 22231 arXiv：。cs。在线。可获取：http://arxiv.org/abs/ 。22231
[37] Y. Yang、J. Mei、Y. Ma、S. Du、W. Chen、Y. Qian、Y. Feng和Y. Liu，“在占据世界中驾驶：面向自动驾驶的基于世界模型的以视觉为中心的四维占据预测与规划”，2025年1月，arXiv:2408.14197 [cs]。 [在线]。可获取：http://arxiv.org/abs/2408 14197 。W. Wu、X. Guo、W. Tang、T.
[38] Huang、C. Wang和C. Ding，“DriveScape：基于多视角特征融合的高分辨率驾驶视频生成。”[39] L. Russell、A. Hu、L. Bertoni、G. Fedoseev、J. Shotton、E. Arani和G. Corrado，“GAIA-2：面向自动驾驶的可控多视角生成式世界模型”，2025年3月，arXiv:2503.20523 [cs]。[ ] 2503 20523 在线。可获取：
[40] http://arxiv.org/abs/ 。B. Xie、Y. Liu、T. Wang、J. Cao和X. Zhang，“GLAD：一种流式场景生成器，用于”，2025。[41] R. Gao、K. Chen、B. Xiao、L. Hong、Z. Li和Q. Xu，“MagicDrive-V2：面向自动驾驶的自适应控制高分辨率长视频生成”，2025年7月，arXiv:2411.13807 [cs]。 [在线]。可获取：http://arxiv.org/abs/2411 13807 。T. Yan、T. Tang、X. Gui、Y. Li、J.
[42] Zhesng、W. Huang、L. Kong、W. Han、X. Zhou、X. Zhang、Y. Zhan、K. Zhan、C. zhong Xu和J. Shen，“Ad-r1：采用无偏世界模型的端到端自动驾驶闭环强化学习”，2025。 [在线]。可获取：https://arxiv.org/abs/2511 20325 。C. Wang、L. Wang、Z. Lu、C.
[43] Wu、G. Shou和X. Wen，“基于脑启发因果感知对比学习与时频预测的自动驾驶”，IEEE物联网期刊，第12卷，第14期，第26 371–26 386页，2025年7月。 [在线]。可获取：https://ieeexplore.ieee.org/abstract/document/10960432
[44] T. Xia、Y. Li、L. Zhou、J. Yao、K. Xiong、H. Sun、B. Wang、K. Ma、G. Chen、H. Ye、W. Liu和X. Wang，“DriveLaW：在潜在驾驶世界中统一规划与视频生成”，2025年12月，arXiv:2512.23421 [cs]。 [在线]。可获取：http://arxiv.org/abs/2512. 23421[45] Z. Tang、X. Chen、Y. Li和J. Chen，“基于潜在深度强化学习与示范的高效泛化端到端自动驾驶系统”，载于《数据库中的机器学习与知识发现》。研究轨道，R. P. Ribeiro、B. Pfahringer、N. Japkowicz、P. Larra˜naga、A. M. Jorge、C. Soares、P. H. Abreu和J. Gama编。Cham：施普林格·自然瑞士，2026，第179–197页。

[46] W. Zheng、R. Song、X. Guo、C. Zhang 和 L. Chen，《GenAD：生成式端到端自动驾驶》，载于《计算机视觉——ECCV 2024》，A. Leonardis、E. Ricci、S. Roth、O. Russakovsky、T. Sattler 和 G. Varol 编。瑞士沙姆：施普林格·自然瑞士，2025年，第87–104页。[47] J. Chen、S. E. Li 和 M. Tomizuka，《利用潜在深度强化学习实现可解释的端到端城市自动驾驶》，《IEEE智能交通系统汇刊》，第23卷，第6期，第5068–5078页，2022年6月。 [在线]。获取地址：https: // ieeexplore.ieee.org/document/9346000/ [48] L. Xiao、J.-J. Liu、S. Yang、X. Li、X. Ye、W. Yang 和 J. Wang，《在自动驾驶中从潜在世界模型学习多个概率决策》，载于2025年IEEE国际机器人与自动化会议（ICRA），2025年5月，第1279–1285页。 [在线]。获取地址：https://ieeexplore.ieee.org/abstract/document/11127996
[49] Y. Li、L. Fan、J. He、Y. Wang、Y. Chen、Z. Zhang 和 T. Tan，《利用潜在世界模型增强端到端自动驾驶》，arXiv预印本 arXiv:2406.08481，2024年。[50] Z. Song、C. Jia、L. Liu、H. Pan、Y. Zhang、J. Wang、X. Zhang、S. Xu、L. Yang 和 Y. Luo，《别猛打方向盘：端到端自动驾驶中的动量感知规划》。[51] Z. Gao、Y. Mu、C. Chen、J. Duan、P. Luo、Y. Lu 和 S. Eben Li，《通过语义掩码世界模型提升端到端城市自动驾驶的样本效率和鲁棒性》，《IEEE智能交通系统汇刊》，第25卷，第10期，第13 067–13 079页，2024年10月。 [在线]。获取地址：https://ieeexplore.ieee.org/abstract/document/10838211
[52] Y. Zheng、P. Yang、Z. Xing、Q. Zhang、Y. Zheng、Y. Gao、P. Li、T. Zhang、Z. Xia、P. Jia 和 D. Zhao，《World4Drive：基于意图感知物理潜在世界模型的端到端自动驾驶》，2025年7月，arXiv:2507.00603 [cs]。 [在线]。获取地址：http://arxiv.org/abs/2507 00603 。S.-H. Ho、B. Thach 和 M. Zhu，《
[53] LiDAR-EDIT：通过编辑真实场景中的物体布局生成激光雷达数据》，载于2025年IEEE国际机器人与自动化会议（ICRA），2025年5月，第214–220页。[在线]。获取地址：https://ieeexplore. ieee.org/document/11127587[54] B. Huang、Y. Wen、Y. Zhao、Y. Hu、Y. Liu、F. Jia、W. Mao、T. Wang、C. Zhang、C. W. Chen、Z. Chen 和 X. Zhang，《SubjectDrive：通过主体控制扩展自动驾驶中的生成式数据》，《AAAI人工智能会议论文集》，第39卷，第4期，第3617–3625页，2025年4月。 [在线]。获取地址：https://ojs.aaai.org/index.php/AAAI/article/view/32376
[55] H. Goel 和 S. S. Narasimhan，《研讨会：利用潜在扩散模型生成的合成数据改进端到端自动驾驶》。[56] 英伟达，：，Y. Wang、W. Luo、J. Bai、Y. Cao、T. Che、K. Chen、Y. Chen、J. Diamond、Y. Ding、W. Ding、L. Feng、G. Heinrich、J. Huang、P. Karkus、B. Li、P. Li、T.-Y. Lin、D. Liu、M.-Y. Liu、L. Liu、Z. Liu、J. Lu、Y. Mao、P. Molchanov、L. Pavao、Z. Peng、M. Ranzinger、E. Schmerling、S. Shen、Y. Shi、S. Tariq、R. Tian、T. Wekel、X. Weng、T. Xiao、E. Yang、X. Yang、Y. You、X. Zeng、W. Zhang、B. Ivanovic 和 M. Pavone，《Alpamayo-R1：通过长尾场景中的推理与动作预测实现可泛化的自动驾驶》，2025年，版本号：1 [ ] 2511 00088 。在线。获取地址：https://arxiv.org/abs/。Q. Peng、X. Chen、C. Yang、S. Shi 和 H. Li，《ColaVLA：利用认知潜在推理实现自动驾驶中的分层并行轨迹规划》，2025年12月，arXiv:2512.22939 [cs]。 [ ] 2512 22939 在线。[58] 获取地址：http://arxiv.org/abs/。E. Ahmadi、R. Mercurius、S. Alizadeh、K. Rezaee 和 A. Rasouli，《控制你的注意力：用于自动驾驶鲁棒轨迹预测的因果注意力门控》，载于2025年IEEE国际机器人与自动化会议（ICRA），2025年5月，第779–786页。 [在线]。获取地址：https://ieeexplore.ieee.org/document/11128367
[59] A. Bairouk、M. Maras、S. Herlin、A. Amini、M. Blanchon、R. Hasani、P. Chareyre 和 D. Rus，《探索潜在路径：利用变分自编码器增强自动驾驶的可解释性》，载于2024年IEEE/RSJ国际智能机器人与系统会议（IROS），2024年10月，第9968–9975页，ISSN：2153-0866。 [在线]。获取地址：https://ieeexplore.ieee.org/abstract/document/10801915
[60] S. Zeng、X. Chang、M. Xie、X. Liu、Y. Bai、Z. Pan、M. Xu 和 X. Wei，《FutureSightDrive：利用时空思维链进行视觉思考以实现自动驾驶》，2025年10月，arXiv:2505.17685 [cs]。 [在线]。获取地址：http://arxiv.org/abs/2505 17685 。H. Lin、Y. Yang、Y. Zhang、C.
[61] Zheng、J. Feng、S. Wang、Z. Wang、S. Chen、B. Wang 、Y. Zhang 、X. Liu 、S. Cui 、和 Z. Li 、“FutureX:

通过潜在思维链世界模型增强端到端自动驾驶，”2025年12月，arXiv:2512.11226 [cs]。 [在线]。可访问：http://arxiv.org/abs/2512.11226。S. Tan、K. Chitta、Y. Chen、R. Tian、Y. You、Y. Wang、W. Luo、Y. Cao、P. Krahenbuhl、M. Pavone 和 B. Ivanovic，“面向端到端驾驶的潜在思维链世界建模”，2025，2512.10226 年12月，[在线]。可访问：http://arxiv.org/abs/2512.10226。B. Sun、Y. Cao、Y. Wang、R. Wang、J. Shang、X. Feng、J. Lu、J. Shi、S. Yang、X. Yan 和 Z. Song，“MindDrive：连接世界模型与视觉语言模型、用于端到端自动驾驶的一体化框架”，2025年12月，arXiv:2512.04441 [cs]。 [在线]。可访问：http://arxiv.org/abs/2512.04441。Z. Zhang、H. Zheng、Y. Wang、L. Xu、T. Deng、X. Chen、Q. Chen、B. Zhang 和 W. Huang，“OmniDrive-R1：用于可信视觉语言自动驾驶的强化驱动交错多模态思维链”，2025，2512.14044 年12月，[在线]。可访问：http://arxiv.org/abs/2512.14044。J. Yao、B. Yang 和 X. Wang，“重建与生成：驾驭潜在扩散模型中的优化困境。”B. Feng、Z. Mei、B. Li、J. Ost、R. Girgis、A. Majumdar 和 F. Heide，“VERDI：用于自动驾驶的视觉语言模型嵌入式推理”，2025年5月，arXiv:2505.15925 [cs]。 [在线]。可访问：http://arxiv.org/abs/2505.15925。D. Shenkut 和 B. V. Kumar，“行人轨迹预测的视觉语言推理”，载于《2025年IEEE国际机器人与自动化会议》（ICRA），2025年5月，第771–778页。 [在线]。可访问：https://ieeexplore.ieee.org/document/11127538“VLA-R：面向开放世界端到端自动驾驶的视觉语言动作检索。” [在线]。可访问：https://arxiv.org/html/2511.12405v1?utm_source=chatgpt.com Z. Shao、P. Wang、Q. Zhu、R. Xu、J. Song、X. Bi、H. Zhang、M. Zhang、Y. K. Li、Y. Wu 和 D. Guo，“DeepSeekMath：突破开放语言模型数学推理的极限”，2024年4月，arXiv:2402.03300 [cs]。 [在线]。可访问：http://arxiv.org/abs/2402.03300。R. Rombach、A. Blattmann、D. Lorenz、P. Esser 和 B. Ommer，“利用潜在扩散模型进行高分辨率图像合成”，载于《2022年IEEE/CVF计算机视觉与模式识别会议》（CVPR）。IEEE，2022年，第10 674–10 685页。L. Zhang、A. Rao 和 M. Agrawala，“为文本到图像扩散模型添加条件控制”，2023年。 [在线]。可访问：https://arxiv.org/abs/2302.05543。A. van den Oord、O. Vinyals 和 K. Kavukcuoglu，“神经离散表示学习”，2018年。 [在线]。可访问：https://arxiv.org/abs/1711.00937。H. Chang、H. Zhang、L. Jiang、C. Liu 和 W. T. Freeman，“MaskGIT：掩码生成式图像变换器”，2022年。 [在线]。可访问：https://arxiv.org/abs/2202.04200。P. Dhariwal 和 A. Nichol，“扩散模型在图像合成方面胜过生成对抗网络”，2021 [在线]。可访问：https://arxiv.org/abs/2105.05233。D. Kahneman，《思考，快与慢》。美国纽约州纽约：法勒、斯特劳斯和吉鲁出版社，2011年。T. Unterthiner 等，“迈向精准的视频生成模型：一项新指标与挑战”，2019年。H. Caesar、V. Bankiti、A. H. Lang、S. Vora、V. E. Liong、Q. Xu、A. Krishnan、Y. Pan、G. Baldan 和 O. Beijbom，“nuScenes：用于自动驾驶的多模态数据集”，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2020年，第11 621–11 631页。X. Wang、Z. Zhu、W. Xu、Y. Zhang、Y. Wei、X. Chi、Y. Ye、D. Du、J. Lu 和 X. Wang，“OpenOccupancy：用于周围环境语义占用感知的大规模基准测试”，载于《IEEE/CVF国际计算机视觉会议论文集》，2023年，第17 850–17 859页。P. Sun、H. Kretzschmar、X. Dotiwalla、A. Chouard、V. Patnaik、P. Tsui、J. Guo、Y. Zhou、Y. Chai、B. Caine、V. Vasudevan、W. Han、J. Ngiam、H. Zhao、A. Timofeev、S. Ettinger、M. Krivokon、A. Gao、A. Joshi、Y. Zhang、J. Shlens、Z. Chen 和 D. Anguelov，“自动驾驶感知的可扩展性：Waymo开放数据集”，载于《IEEE/CVF计算机视觉与模式识别会议论文集》（CVPR），2020年6月。 CARLA模拟器团队，“CARLA自动驾驶排行榜2.0”，2024年。 [在线]。可访问：https://leaderboard.carla.org/get_started_v2_0/ D. Dauner、M. Hallgarten、T. Li、X. Weng、Z. Huang、Z. Yang、H. Li、I. Gilitschenski、B. Ivanovic、M. Pavone 等，“NAVSIM：数据驱动的非反应式自动驾驶车辆仿真与基准评测，”神经信息处理系统进展，第37卷，第28 706–28719页，2024。 G. Rong、B. H. Shin、H. Tabatabaee、Q. Lu、S. Lemke、M. Možeiko、E. Boise、G. Uhm、M. Gerow、S. Mehta等，“LGSVL仿真器：面向自动驾驶的高保真仿真器”，载于*2020年IEEE第23届智能交通系统国际会议（ITSC）*。IEEE，2020，第1–6页。 S. Shah、D. Dey、C. Lovett和A. Kapoor，“AirSim：面向自动驾驶车辆的高保真视觉与物理仿真”，载于野外与服务机器人：第*11*届国际会议成果。Springer，2017，第621–635页。 Q. Li、Z. Peng、L. Feng、Q. Zhang、Z. Xue和B. Zhou，“MetaDrive：为可泛化强化学习构建多样化驾驶场景”，*IEEE*模式分析与机器智能汇刊，第45卷，第3期，第3461–3475页，2022。 C. Gulino、J. Fu、W. Luo、G. Tucker、E. Bronstein、Y. Lu、J. Harb、X. Pan、Y. Wang、X. Chen、J. D. Co-Reyes、R. Agarwal、R. Roelofs、Y. Lu、N. Montali、P. Mougin、Z. Yang、B. White、A. Faust、R. McAllister、D. Anguelov和B. Sapp，“Waymax：面向大规模自动驾驶研究的加速数据驱动仿真器”，载于第*37*届神经信息处理系统国际会议论文集，系列NIPS ’23。美国纽约州Red Hook：Curran Associates Inc.，2023。 A. Dosovitskiy等，“CARLA：开放城市驾驶仿真器”，载于*CoRL*，2017。 F. Codevilla、M. M¨uller、A. L´opez、V. Koltun和A. Dosovitskiy，“基于条件模仿学习的端到端驾驶”，载于*2018年IEEE*国际机器人与自动化会议（*ICRA*）。IEEE，2018，第4693–4700页。 P. Sun等，“自动驾驶感知的可扩展性：Waymo开放数据集”，载于*CVPR*，2020。 A. Kendall和Y. Gal，“计算机视觉中的贝叶斯深度学习需要哪些不确定性？”，载于*NeurIPS*，2017。 A. Nachkov、D. Pani Paudel、J.-N. Zaech、D. Scaramuzza和L. Van Gool，“梦想驾驶：使用解析世界模型的基于模型的车辆控制”，*arXiv*预印本，第arXiv–2502页，2025。 Y. Dong、C. Liu、Y. Wang和Z. Fu，“迈向理解全球隐式驾驶线索的跨文化差异：综述、比较分析与研究路线图”，载于*2024年IEEE第27届*智能交通系统国际会议（*ITSC*），2024，第1569–1575页。



Rongxiang Zeng 于2025年获得中国深圳深圳技术大学工学学士学位。他目前正在德国亚琛亚琛工业大学攻读机器人系统工程理学硕士学位。他的研究兴趣包括机器人学、自动驾驶、世界模型以及视觉-语言-动作模型。



Yongqi Dong 于2017年获得中国北京清华大学控制科学与工程理学硕士学位，并于2025年获得荷兰代尔夫特代尔夫特理工大学交通与规划博士学位。他目前在德国亚琛亚琛工业大学担任研究组负责人。他担任IEEE ITSS混合交通中的自动出行技术委员会主席。他的研究兴趣包括深度学习、自动驾驶、视觉-语言模型以及交通安全。他力求运用人工智能与跨学科研究作为工具，塑造一个更美好的世界。