arXiv:2602.01536v1 [cs.RO] 2 Feb 2026

UniDWM：通过多方面表征学习迈向统一驾驶世界模型

Shuai Liu [1] **Siheng Ren** [1] **Xiaoyao Zhu** [1] **Quanmin Liang** [1] **Zefeng Li** [1] **Qiang Li** [2] **Xin Hu** [2] **Kai Huang** [1]

摘要

要在复杂驾驶环境中实现可靠且高效的规划，需要一个能够对场景几何结构、外观和动力学进行推理的模型。我们提出了**UniDWM**，一个**统一驾驶世界模型**，通过多方面表征学习推动自动驾驶发展。UniDWM构建了一种结构与动力学感知的潜在世界表示，作为具有物理基础的状态空间，使感知、预测和规划能够进行一致的推理。具体而言，联合重建路径学习恢复场景结构，包括几何形状和视觉纹理；而协同生成框架则利用条件扩散Transformer，在潜在空间中预测未来的世界演化。此外，我们证明UniDWM可视为变分自编码器的一种变体，为多方面表征学习提供理论指导。大量实验表明，UniDWM在轨迹规划、4D重建与生成方面效果显著，凸显了多方面世界表征作为统一驾驶智能基础的潜力。代码将公开发布于 https://github.com/Say2L/UniDWM。

1. 引言

自动驾驶中的可靠规划，取决于对物理世界的全面理解——这种理解不仅要捕捉场景的视觉外观，还要把握其内在几何结构及随时间推移的动态演化。近年来，自动驾驶领域取得了长足进展（Hu等人，2023；Jiang等人，2023；Chen等人，2024年；Liao 等人，2025年；Liu 等人，2025年）通过将传统的感知-预测-规划流水线重新构建为端到端范式，取得了显著进展，从而减少了各阶段间的误差累积。然而，为了学习能够联合编码外观、几何结构和场景动力学的表征，这些方法仍严重依赖成本高昂的感知标注，限制了其可扩展性。

人类通过在脑海中模拟不断演变的世界来做出驾驶决策——依据有限的观测，预测周围智能体、物体和自车的未来状态。这一过程隐式地构建出连贯的世界模型，在统一框架中紧密结合视觉感知、空间推理和动作规划。受此观察启发，我们认为，自主智能体应配备统一驾驶世界模型，能够根据多视角视觉输入等部分观测，联合重建并预测4D世界状态。这样的建模方式天然适用于自监督学习：从有限感官输入中推断出的潜在世界表征，可用于恢复未观测到的信息，包括场景几何和未来世界动态。

仅从视觉观测中学习统一驾驶世界模型面临若干根本性挑战。首先，建立 时空一致性 对于建模真实动力学至关重要。潜在表征不仅必须编码每帧的空间结构，还必须保持时间连续性，忠实反映场景和智能体在因果运动作用下随时间的演变。其次，统一驾驶世界模型的学习缺乏理论指导。现有大多数方法（Zhang 等人，2025年；Wang 等人，2025年；Liu 等人，2025年）都基于启发式设计，对于哪些原则对于联合建模动态物理世界至关重要缺乏深入理解，从而阻碍了系统性改进与泛化。

为应对这些挑战，我们提出了**UniDWM**，一种统一驾驶世界模型，能够学习结构感知且动态感知的潜在世界表征，并将其作为具有物理基础的状态空间。该表征支持在感知、预测和规划之间进行连贯推理。如

[1]中山大学计算机科学与工程学院 [2]小鹏汽车科技有限公司。通讯作者：Kai Huang <huangk36@mail.sysu.edu.cn>、Shuai Liu<liush376@mail2.sysu.edu.cn>。

预印本。*2026*年*2*月*3*日。

图 1所示，UniDWM由两个关键部分组成。首先，联合重建路径根据序列观测重建场景结构——包括几何、视觉外观和自车运动——从而得到统一的潜在世界表征。其次，协同生成模块利用条件扩散Transformer（DiT）预测未来的世界演化，从而将动力学感知信息隐式注入潜在空间。此外，我们证明，UniDWM可以表述为一种变分自编码器（VAE）的变体（Kingma 和 Welling，2013年；Zhao等人，2017年），为其统一表征学习范式提供理论基础。

通过在 NAVSIM 上进行涵盖轨迹规划、4D重建和4D生成的全面评估，我们证明 UniDWM 能够在单一潜在空间中有效统一感知、预测和规划，从而为规划带来切实收益。我们的主要贡献总结如下：

- 我们提出 UniDWM，这是一种统一驾驶世界模型，能够学习具有结构感知和动态感知能力的潜在表示，将其作为用于感知、预测和规划的物理基础状态空间。
- 我们从理论上将 UniDWM 解释为变分自编码器的一种变体，为多方面表征学习提供了有原则的指导。
- 我们在 NAVSIM 基准上对 UniDWM 进行了全面评估，涵盖轨迹规划、4D重建和4D生成，验证了我们方法的有效性。

2. 相关工作

驾驶世界模型。世界模型的最新进展推动自动驾驶研究转向数据驱动的生成式推理。现有方法大致可分为两类。第一类工作侧重于场景级生成，用于仿真和数据增强。VISTA（Gao 等人，2024年）和 GAIA-2（Russell 等人，2025年）等方法通过视角和语义条件扩散实现可控的场景合成，而 MagicDrive-V2（Gao 等人，2025年）和 Genesis（Guo 等人，2025年）则进一步融入结构化布局，以生成逼真的驾驶场景。第二类研究基于世界模型的驾驶控制（Zhang 等人，2025年；Zheng 等人，2025b；Shi 等人，2025年；Li 等人，2025b），将学习到的动力学和生成式预测整合进决策过程。具体而言，Epona（Zhang 等人，2025年）将基于扩散的预测与端到端规划相结合，而 DriveX（Shi 等人，2025年）和 WoTe（Li 等人，2025b）则将语义信息嵌入鸟瞰图潜在表征。相比之下，我们的 UniDWM 专注于学习统一的潜在世界表征，以联合建模空间结构与动力学，从而更贴近真实的底层世界，并提升在下游任务中的泛化能力。

端到端轨迹规划。端到端自动驾驶（E2E-AD）将传感器输入直接映射为控制输出，绕过模块化的感知—预测—规划流程。早期基于卷积神经网络的方法（Bojarski 等人，2016）根据摄像头图像预测转向，后续的条件模仿学习（CIL）（Codevilla 等人，2018）则将高层导航指令纳入其中。CILRS（Codevilla 等人，2019）通过残差架构和辅助目标提升了泛化能力。近期方法则探索联合优化和结构化表征。UniAD（Hu等人，2023）采用基于查询的鸟瞰视图设计进行联合优化，而 VAD（Jiang 等人，2023）和 VADv2（Chen 等人，2024）则采用向量化表示对驾驶场景建模，以提升效率。生成式方法，例如扩散条件轨迹生成器（Xing 等人，2025；Liao 等人，2025）和动量感知规划（Song 等人，2025），进一步提升了鲁棒性和不确定性处理能力。尽管取得了这些进展，大多数端到端自动驾驶方法仍高度依赖成本高昂的感知标注，因为仅基于轨迹的监督往往会导致表示坍塌或捷径学习行为（Prakash 等人，2021；Dauner 等人，2024）。相比之下，UniDWM采用自监督学习范式来学习统一的潜在世界表征，并通过多种重建和正则化目标促使其变得平滑且结构良好。

自动驾驶表示学习。学习有效的世界表征对于感知、预测和规划推理至关重要。基于鸟瞰视图的方法（Li 等人，2024b；Liang 等人，2022；Jia 等人，2023）提供了结构化的空间对齐，支持联合检测、建图和运动预测，但会因投影而丢失信息。以图像为中心的多视图表示（Wang 等人，2022；Liu 等人，2022；Jia 等人，2025）保留了用于细粒度感知的透视线索，但会引入深度歧义。基于高斯的场景表征（Kerbl 等人，2023；2024；Huang 等人，2024；Liu 等人，2025）能够连续地建模几何与外观，并实现实时渲染，但通常需要离线优化。总体而言，鸟瞰视图方法依赖较强的归纳偏置，以图像为中心的方法缺乏显式几何建模，而基于高斯的表征很少应用于轨迹规划。UniDWM 旨在学习具有强泛化能力的统一潜在世界表征，以最小的归纳偏置支持多个下游任务。UniDWM 通过多方面表征学习实现这一目标，联合捕获几何、外观和演化动力学，从而为基于物理的世界建模奠定基础。

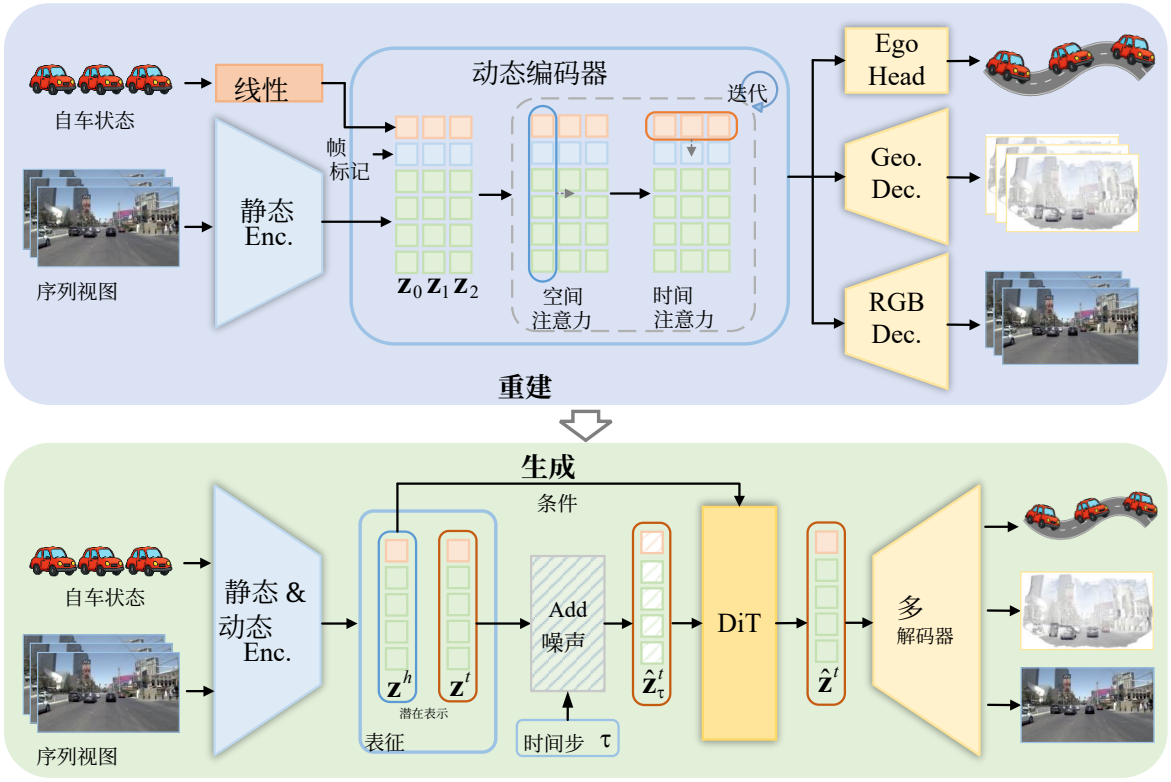


图*1.* **UniDWM 的整体框架**。 UniDWM 由重建阶段和生成阶段组成。给定一系列视觉观测作为输入，UniDWM 使用静态编码器和动态编码器将其投影到统一特征空间中。随后，重建解码器和生成解码器可分别用于重建观测到的场景和生成动态演化。

3. 方法

3.1. 预备知识

变分自编码器（VAE）（Kingma 和 Welling, 2013年）是一种概率生成框架，它引入潜在变量 $z \in \mathcal{Z}$ 来解释观测$\mathbf{x} \in \mathcal{X}$。该模型假设潜在空间上存在一个简单的先验分布 p（$\mathbf{z}$），通常选用高斯分布，并定义一个由神经网络参数化的条件数据分布 p_θ（$\mathbf{x}|\mathbf{z}$）。给定由有限训练集近似的数据生成分布 $p_\mathcal{D}$（$\mathbf{x}$），学习过程自然可表述为最大化期望边际对数似然：

$$\mathbb{E}_{p_\mathcal{D}(\mathbf{x})}\big[\log p_\theta(\mathbf{x})\big] = \mathbb{E}_{p_\mathcal{D}(\mathbf{x})}\Big[\log \mathbb{E}_{p(\mathbf{z})}\big[p_\theta(\mathbf{x} \mid \mathbf{z})\big]\Big]. \quad (1)$$

由于对潜变量进行积分难以处理，直接优化这一目标通常不可行。为克服这一困难，变分自编码器引入变分后验 $q_\phi(\mathbf{z}|\mathbf{x})$，并最大化对数似然的一个可处理下界，如下所示：

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathbf{x} \mid \mathbf{z})\big] - \mathrm{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) \,\|\, p(\mathbf{z})), \quad (2)$$

其满足 $\mathcal{L}_{\text{VAE}} \le \log p_\theta(\mathbf{x})$。

尽管标准变分自编码器目标在实践中效果显著，但它会隐式限制观测与潜变量之间的互信息，常导致表征的信息性较弱。InfoVAE（Zhao等人，2017年）通过将信息保持与先验正则化解耦，缓解了这一问题。由此得到的目标函数表示为：

$$\begin{aligned}\mathcal{L}_{\text{InfoVAE}} \equiv\ & \mathbb{E}_{p_\mathcal{D}(\mathbf{x})}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathbf{x} \mid \mathbf{z})\big] \\ & - (1-\alpha)\,\mathbb{E}_{p_\mathcal{D}(\mathbf{x})}\mathrm{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) \,\|\, p(\mathbf{z})) \\ & - (\alpha + \lambda - 1)\,\mathrm{D}(q_\phi(\mathbf{z}) \,\|\, p(\mathbf{z})),\end{aligned} \quad (3)$$

其中，$q_\phi(z) = \mathbb{E}_{p_\mathcal{D}(x)} q_\phi(z|x)$ 表示聚合后验。这里，$D(\cdot\|\cdot)$ 表示一种散度度量，例如 KL 散度或最大均值差异，而超参数 α 和 λ 用于平衡重建保真度、先验匹配和信息保持。

3.2. 统一驾驶世界模型

我们考虑一个由 $\mathcal{D} = \{\mathbf{x}_i \in \mathcal{X}\}_{i=1}^N$ 个独立同分布场景组成的驾驶数据集，每个场景都是由来自不同视角的 M 个观测组成的集合。我们将具有单一模态的一帧视为一个观测视角。如果数据集中的一个场景由 l 帧和 m 种模态（例如点、图像、自车位姿等）组成，那么 M 等于 $l \times m$。给定场景的一些局部观测 $\mathbf{x}^{local} \in \mathbf{x}$

，UniDWM 的目标是重建全局观测 $\mathbf{x}$。UniDWM 的证据下界（ELBO）可以写为：

$$\mathcal{L}_{\text{UniDWM0}} \equiv \sum_{j=1}^{M} \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^{local})} \left[\log p_\theta(\boldsymbol{x}^{(j)} \mid \mathbf{z}) \right] - \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})} \mathrm{KL}\left(q_\phi(\mathbf{z} \mid \mathbf{x}^{local}) \,\|\, p(\mathbf{z}) \right), \tag{4}$$

其中 $z \in \mathcal{Z}$ 表示潜变量，$p_\theta(\boldsymbol{x}^{(j)}|\mathbf{z})$表示单模态条件分布，而$q_\phi(z|\mathbf{x}^{local})$ 是给定局部观测的近似后验。证明见附录A。然而，变分自编码器风格的证据下界可能无法学习到有意义的潜在表征，尤其是在数据空间 $\mathcal{X}$ 复杂且维度较高时。因此，我们进一步提出一个受 InfoVAE 启发的证据下界目标：

$$\mathcal{L}_{\text{UniDWM}} = \sum_{j=1}^{M} \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^{local})} \left[\log p_\theta(\boldsymbol{x}^{(j)} \mid \mathbf{z}) \right] - \lambda D(q_\phi(\mathbf{z}) \,\|\, p(\mathbf{z})), \tag{5}$$

该目标与式 3 类似，但采用多观测重建，并令 $a = 1$。证明见附录 B。实践中，我们采用 SIGReg（Balestriero & LeCun，2025年）作为正则项 $D(\cdot\|\cdot)$。SIGReg 使用 Epps-Pulley 检验来衡量分布差异，满足 InfoVAE 的要求（$D(q_\phi(z)\|\, p(z)) = 0$ 当且仅当 $q_\phi(z) = p(z)$）。对数似然项通过负重建损失和生成损失 $\mathcal{L}_{recon}$ 及 $\mathcal{L}_{gen}$ 实现，详情将在第 3.4 节和第 3.5 节中介绍。

3.3. UniDWM的框架

UniDWM的整体框架如图1所示。UniDWM采用时空编码器–解码器架构：编码器将视觉输入压缩为紧凑的潜在特征场，而一系列解码器则从该表征中重建驾驶场景的不同物理方面。为实现如此全面的表征学习，UniDWM首先引入联合重建（第3.4节），通过从局部观测中联合重建几何、外观和自车运动，引导编码器感知空间结构。随后，预训练模型在协同生成阶段（第3.5节）进行微调，以进一步将动态线索编码到潜在世界表征中。这些组件共同使UniDWM不仅能够作为面向世界建模的自监督基础模型，也能够作为统一框架，支持广泛的下游任务，包括对几何与视觉结构的推理（感知）、对周围交通参与者动力学的建模（预测），以及捕捉自运动演变（规划）。

3.4. 联合重建

为了学习具有物理依据的潜在世界表征，我们通过联合4D

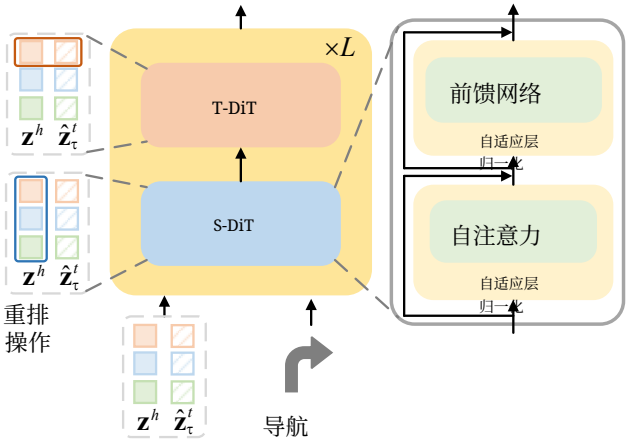


图2. **DiT 块的架构**，由空间 DiT（S-DiT）和时间 DiT（T-DiT）模块交替排列组成。

场景重建路径构建世界表征，以重建随时间变化的三维几何、视觉外观和自车运动。具体而言，我们采用变分自编码架构，其中共享的时空编码器 q_ϕ 将一系列视觉观测 $I_{1:n}$ 和自车状态 $s_{1:n}$ 压缩为统一的潜在嵌入 $\mathbf{z}$：

$$\mathbf{z} \sim q_\phi(\mathbf{z} \mid I_{1:n}, s_{1:n}), \tag{6}$$

其中，$\mathbf{z}$ 旨在实现模态无关 和时间连续，使其能够在共享潜在空间中联合编码三维结构先验、时间运动线索和视觉外观信息。共享时空编码器由两个互补组件构成：静态编码器和动态编码器。

静态编码器。 为降低训练开销并稳定表征学习，我们复用一个预训练编码器（Chen 等，2025年；Sim´eoni 等，2025年）作为静态编码器，负责从单帧中提取视角一致特征。训练期间，该静态编码器的参数保持冻结，以保留通用的低层表征。自车状态由多层感知机编码。随后得到一个经过深度空间压缩的潜在状态$\mathbf{z} \in \mathbb{R}^{n\times(l+1)\times c}$：

$$\begin{aligned} \mathbf{z}_{rgb} &= \text{Encoder}(I_{1:n}) \in \mathbb{R}^{n\times l\times c}, \\ \mathbf{z}_{ego} &= \text{MLP}(s_{1:n}) \in \mathbb{R}^{n\times 1\times c}, \\ \mathbf{z} &= \text{Concatenate}([\mathbf{z}_{rgb}, \mathbf{z}_{ego}]), \end{aligned} \tag{7}$$

其中，c 表示通道数，l 表示每帧词元的长度。

动态编码器。 动态编码器旨在对场景演变和自车动态随时间的变化进行建模。受时空建模与几何建模领域近期进展的启发（Hu等人，2024年；Zhang 等人，2025年；Wang et al.，2025年），

我们通过交替堆叠空间注意力层和时间注意力层来构建动态编码器。动态编码器的架构如图 1所示。空间注意力聚合每帧内的几何和上下文特征，而时间注意力则在潜在域中捕捉帧间运动与连续性。这种交替结构能够有效实现跨帧交互和流畅的时间推理，确保生成的潜在表示 L z 反映对驾驶场景一致的四维理解。具体而言，我们采用由 L 层时空变换器组成的堆栈，逐步捕捉驾驶场景的结构信息：

$$\begin{aligned}\mathbf{z} &\leftarrow \text{SpatialAttn}(\mathbf{z}),\\ \mathbf{z} &\leftarrow \text{TemporalAttn}(\mathbf{z}, \mathcal{M}),\end{aligned} \tag{8}$$

其中， $\mathcal{M} \in \mathbb{R}^{n\times n}$ 是因果掩码，用于确保信息随时间单向流动。

解耦解码器。给定共享的潜在表示世界表征$\mathbf{z}$，目标是分别重建几何、视觉外观和自车运动。由于这些模态位于不同的特征域中，我们设计了解耦解码器，将 $\mathbf{z}$ 投影到各自的特征空间：

$$\begin{aligned}\{\hat{\mathbf{D}}, \hat{\mathbf{P}}, \Sigma_d, \Sigma_p\} &= \mathcal{D}^{\text{geo}}(\mathbf{z}),\\ \hat{\mathbf{C}} = \mathcal{D}^{\text{rgb}}(\mathbf{z}), \hat{\mathbf{e}} &= \mathcal{D}^{\text{ego}}(\mathbf{z}),\end{aligned} \tag{9}$$

其中， $\{\hat{}\mathbf{D} \in \mathbb{R}^{n\times H\times W\times 3}$、$\hat{}\mathbf{P} \in \mathbb{R}^{n\times H\times W\times 1}\}$表示重建几何属性（点和深度）。$\{\Sigma_d, \Sigma_p\}$ 表示深度和点预测的偶然不确定性（Kendall &Cipolla，2016；Novotny et al.，2018年）。$\hat{}\mathbf{C} \in \mathbb{R}^{n\times H\times W\times 3}$和$\hat{}\mathbf{e} \in \mathbb{R}^{n\times 3}$ 分别表示视觉外观（RGB颜色）和自车位姿（自车中心和航向）。

对于几何与外观重建，我们遵循先验方法构建解码器（Chen 等人，2025年；Zheng 等人，2025a）。在实践中，$\mathcal{D}^{\text{rgb}}$ 由预训练解码器权重初始化（Chen 等人，2025年；Zheng 等人，2025a），并在训练期间冻结。对于自车运动解码器，遵循 VGGT（Wang 等人，2025年），我们采用一个轻量级的基于变换器的解码器，由四个自注意力层及随后的线性投影头组成。

重建损失。我们采用三种损失来构建重建损失：

$$\mathcal{L}_{recon} = \mathcal{L}_{ego} + \mathcal{L}_{geo} + \mathcal{L}_{vis}, \tag{10}$$

其中 $\mathcal{L}_{ego}$、$\mathcal{L}_{geo}$和 $\mathcal{L}_{vis}$ 分别是自车运动、几何和视觉外观损失。 $\mathcal{L}_{ego}$ 通过比较预测自车位姿 $\hat{}\mathbf{e}$ 与真实自车位姿 $\mathbf{e}$之间的 L2距离计算得出： $\mathcal{L}_{ego} = \frac{1}{n}||\hat{}\mathbf{e} - \mathbf{e}||_2$。按照先前的研究（Wang 等人，2025年）， $\mathcal{L}_{geo}$ 计算如下 $^-$

如下：

$$\begin{aligned}\mathcal{L}_{geo} =& \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{point}},\\ \mathcal{L}_{\text{depth}} =& \frac{1}{n}\|\Sigma_d\left(\hat{\mathbf{D}} - \mathbf{D}\right)\| +\\ & \frac{1}{n}\|\Sigma_d \odot \left(\nabla\hat{\mathbf{D}} - \nabla\mathbf{D}\right)\| - a\log\Sigma_d,\\ \mathcal{L}_{\text{points}} =& \frac{1}{n}\|\Sigma_p\left(\hat{\mathbf{P}} - \mathbf{P}\right)\| +\\ & \frac{1}{n}\|\Sigma_p \odot \left(\nabla\hat{\mathbf{P}} - \nabla\mathbf{P}\right)\| - a\log\Sigma_p,\end{aligned} \tag{11}$$

其中 $\odot$ 表示通道广播的逐元素乘积， Σ_d 和 Σ_p 分别表示深度和点不确定性图。 $\mathcal{L}_{rgb}$ 的计算如下：

$$\mathcal{L}_{vis} = \|\hat{\mathbf{C}} - \mathbf{C}\| + \omega_l\text{LPIPS}(\hat{\mathbf{C}}, \mathbf{C}) + \omega_g\mathcal{L}_{GAN}, \tag{12}$$

其中 LPIPS$(\cdot)$ 表示 LPIPS（Zhang 等人，2018年）函数，$\mathcal{L}_{GAN}$为 GAN损失，如（Esser 等人，2021年；Zheng 等人，2025a）。 ω_l 和 ω_g 为 LPIPS 与 GAN损失的权重，并设为1.0 和0.75 ，实践中遵循（Zheng 等人，2025a）。

3.5. 协同生成

虽然联合4D场景重建使模型能够学习观测场景中具有物理基础的潜在表征，但自动驾驶不仅需要理解，还需要预测未来动态。为实现这一目标，我们引入了一种协同生成框架，将未来世界演化建模为潜在空间中的条件生成过程。通过以历史上下文为条件联合采样未来4D场景潜变量，模型能够在不同任务间实现连贯预测，从而统一想象未来场景与智能体行为。需要注意的是，生成模块也可被视为未来观测的解码器，使UniDWM具备动态想象能力。

具体而言，给定潜在世界表征$\mathbf{z}^h \in \mathbb{R}^{n^h\times(l+1)\times c}$，其由历史视觉观测和自车位姿编码得到，我们基于扩散的生成器旨在合成未来潜在世界表征$\mathbf{z}^t \in \mathbb{R}^{n^t\times(l+1)\times c}$，用于4D生成与规划。遵循先前工作（Kong 等人，2024年；Zhang 等人，2025年），生成过程采用条件DiT框架实现；详细架构如图2。DiT框架同样由空间注意力层和时间注意力层组成，与动态编码器类似，但使用AdaLN层以适应时间步和引导信息。

在实践中，我们将导航嵌入$\mathbf{g}_{\text{nav}}$ 作为条件引导注入DiT块。这使得生成器能够联合想象在语义上与导航意图对齐的合理未来潜在状态$\hat{}\mathbf{z}^t$ 。条件

表*1.* **在 NAVSIM *navtest* 划分上的性能。** '*'表示我们自行获得的结果。'感知标签'表示某方法所需的感知标签。'C'、'L'和'E'分别表示摄像头、激光雷达和自车状态输入。有感知标签与无感知标签的方法的最佳结果分别以**粗体**标出。

方法	输入	感知标签	NC ↑	DAC ↑	EP ↑	碰撞时间 ↑	PDMS ↑
UniAD（Hu等人，2023）	C	边界框、地图、占用	97.8	91.9	78.8	92.9	83.4
LTF（Prakash 等人，2021年）	C	边界框、地图	97.4	92.8	79.0	92.4	83.8
TransFuser (Prakash 等人，2021年)	C 与 L	边界框，地图	97.7	92.8	79.2	93.0	84.0
Hydra-MDP(Li 等人，2024a)	C 与 L	边界框，地图	98.3	96.0	78.7	94.6	86.5
DiffusionDrive（Liao 等人，2025年）	C 与 L	边界框，地图	98.2	96.2	82.2	94.7	88.1
GoalFlow（Xing 等人，2025年）	C 与 L	边界框，地图	98.4	**98.3**	85.0	94.6	90.3
GaussianFusion（Liu 等人，2025年）	C 与 L	边界框，地图	**98.7**	98.1	**88.2**	**95.7**	**92.0**
Ego-MLP	E	None	93.0	77.3	62.8	83.6	65.6
LAW（Li 等人，2025a）	C	None	97.2	93.3	78.8	91.9	83.8
World4Drive（Zheng 等人，2025b）	C	None	97.4	94.3	79.9	92.8	85.1
DINOv3（B）*（Sim´eoni 等人，2025年）	C	None	97.8	94.4	80.2	93.0	85.4
Epona（Zhang 等人，2025年）	C	None	97.9	95.1	80.4	93.8	86.2
UniDWM（DCAE）	C	None	97.4	93.2	80.1	93.3	84.9
UniDWM（DINOv3-B）	C	None	**98.3**	**97.9**	**86.0**	**95.1**	**90.6**

生成动力学表述为：

$$\hat{\mathbf{z}}_{\tau-1}^{t} = \hat{\mathbf{z}}_{\tau}^{t} + \Delta\tau \cdot v_{\Theta}\left(\hat{\mathbf{z}}_{\tau}^{t}, \tau, [\mathbf{z}^{h}, \mathbf{g}_{\text{nav}}]\right), \quad (13)$$

其中，$\hat{\mathbf{z}}_{\tau}^{t}$ 表示扩散时间步τ处的带噪未来潜在表示，$\mathbf{z}^{h}$ 表示历史潜变量上下文，v_{Θ} 为基于DiT的速度场。生成的损失采用标准的速度预测目标：

$$\mathcal{L}_{\text{gen}} = \mathbb{E}_{\mathbf{z}^{t},\epsilon,\tau}\left\| v_{\Theta}\left(\hat{\mathbf{z}}_{\tau}^{t}, \tau, [\mathbf{z}^{h}, \mathbf{g}_{\text{nav}}]\right) - (\mathbf{z}^{t} - \epsilon)\right\|^{2}, \quad (14)$$

其中，$\mathbf{z}^{t}$ 表示真实的未来潜在表示，$\epsilon \sim \mathcal{N}(0, I)$。

4. 实验

我们主要在 NAVSIM（Dauner 等人，2024年）数据集上评估我们的模型。数据集和评估指标的详细信息见附录C。

4.1. 实现细节

我们分别使用 DINOv3-B（Sim´eoni 等人，2025年）和 DCAE（Chen 等人，2025年）作为静态编码器。动态编码器由 12层时空 Transformer 组成，包含 170M 个参数。几何、外观和自车位姿的重建解码器彼此独立，并按照（Wang 等人，2025年）构建。我们采用一个包含 12层、具有 459M 个参数的下一帧预测扩散 Transformer，以及一个包含 16层、具有 33M 个参数的轨迹预测扩散 Transformer。点和深度图标注通过将激光雷达点投影到图像平面上获得。训练方案由两个阶段组成：1）重建训练，在 *navtrain* 划分的 NAVSIM 数据集上，以重建目标 $\mathcal{L}_{recon} + \lambda$SIGReg 进行从头训练；2）重建与生成联合训练，模型继承第一阶段训练得到的权重，并继续以目标$\mathcal{L}_{gen} + \mathcal{L}_{recon} + \lambda$SIGReg 进行训练。 λ 在所有实验中均设为 2×10^{-4}。所有输入图像均调整为224×384 像素的分辨率。第一阶段和第二阶段均训练 50 个 epoch，批量大小为 64。采用 AdamW优化器，学习率设为 1×10^{-4}，权重衰减设为 5×10^{-2}。在所有实验中，下一帧和轨迹的 DiT采样步数分别设置为 100 和 5。

4.2. 实验结果

轨迹规划评估。 我们评估 UniDWM 在 NAVSIM *navtest* 划分上的表现，并将其与多种具有代表性的轨迹规划方法进行比较，包括感知标签监督方法和无标签方法。如表1所示，利用密集感知标注的方法通常能取得较强的性能，其中 GaussianFusion（Liu 等人，2025年）在感知监督方法中取得了最佳总体结果。聚焦无标签设定，UniDWM 在不依赖任何感知标签的情况下展现出有竞争力的性能。特别是，采用 DINOv3-B 骨干网络的 UniDWM 在所有无标签方法中取得了最佳结果，显著优于原始DINOv3（Sim´eoni 等人，2025年）、Epona（Zhang 等人，2025年）和 World4Drive（Zheng 等人，2025b）等先验方法，且在所有评估指标上均如此。此外，比较采用不同编码器的 UniDWM 变体，凸显了具备足够容量的静态编码器的重要性。将 DCAE

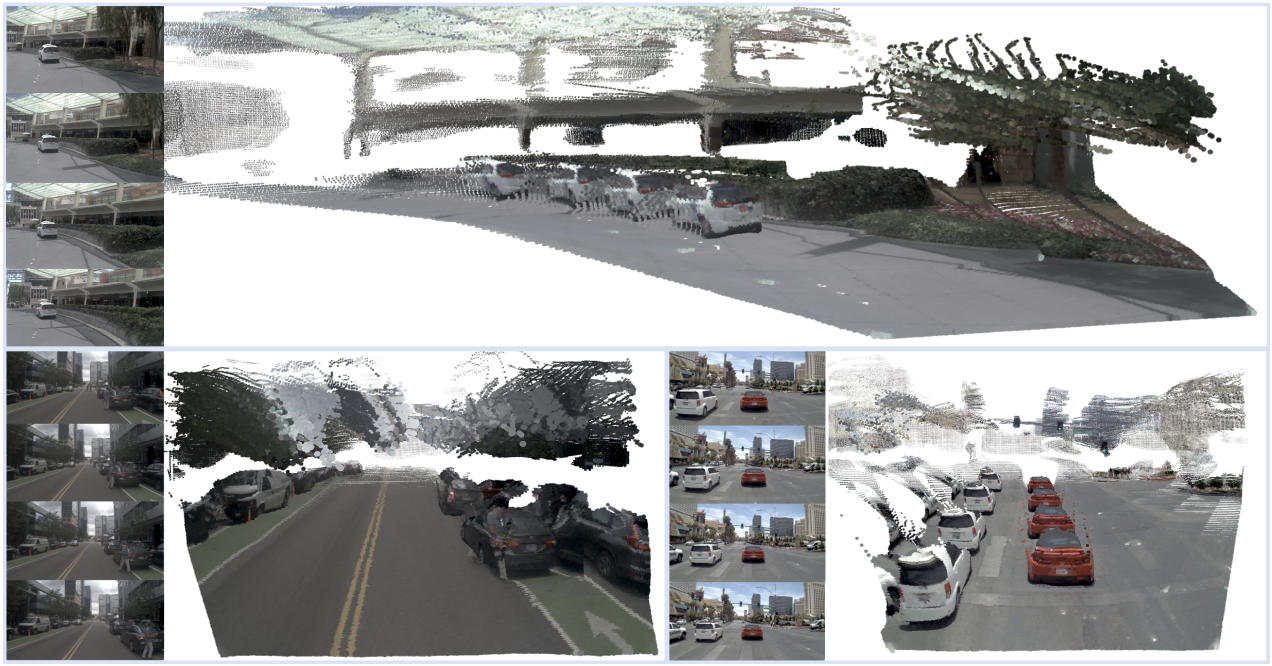


图 *3*。 **4D重建结果可视化**。 左侧图像为序列视觉输入，右侧可视化的点为重建几何。

表 *2*。 **NAVSIM *navtest* 划分上的4D重建结果**。

方法	准确率↓	完整性↓	总体↓
VGGT	3.764	2.236	3.000
Spann3R	2.028	2.202	2.115
UniDWM（我们的方法）	**1.878**	**1.575**	**1.727**

编码器替换为 DINOv3-B 可在所有指标上带来显著提升，表明 UniDWM 能够有效利用更强的编码器。总体而言，这些结果表明，UniDWM 为轨迹规划提供了一种可扩展且标注高效的替代方案，从而缩小无感知标签方法与全监督方法之间的性能差距。

4D重建评估。为评估UniDWM的4D重建能力，我们在NAVSIM Wang等人，2025年）和Spann3R（Wang与Agapito，2025年）的NAVSIM *navtest* 划分上，将其与VGGT（2所示，UniDWM在所有指标上均取得最佳性能，其总体倒角距离（1。727）、准确度（1。878）和完整度（1。575）均为最低。得益于联合4D重建（第3.4节），与VGGT相比，UniDWM将总体指标降低了**42**。4%，证明了其生成高保真且具有时间一致性的4D重建结果的有效性。图3 中的定性结果进一步验证了这些发现。UniDWM能够重建细粒度的几何细节——例如高架结构和地面标线，能够保持动态物体（例如车辆）较强的时间一致性，并实现良好的场景完整性（例如交通信号灯）。综上，这些结果证实了UniDWM用于4D重建的稳健性和有效性。

4D生成评估。图4 展示了 UniDWM 通过对图像和点云进行自回归下一帧预测所实现的 4D生成结果。相对于自车，UniDWM 能够准确捕捉静态场景元素的几何结构和外观，并使视觉内容与生成的几何结构良好对齐。对于动态智能体，模型能够保持合理的几何一致性，但由于自回归推理中的误差累积，长时间跨度下可能出现视觉伪影。这一问题源于训练–推理不匹配，可通过前向链训练等技术缓解（Zhang 等人，2025年）。然而，由于本文主要关注驾驶场景的表征学习，而非生成质量优化，我们将这类技术的探索留待未来工作。

4.3. 消融研究

为了解 UniDWM 中各项设计选择的影响，我们在轨迹规划任务上进行受控消融研究，通过逐步添加自车位姿重建、外观重建、几何结构重建和动态生成解码器，并使用共享的 DCAE 编码器（Chen 等人，2025年）。结果如表3所示。当



图*4.* **4的可视化** **D生成结果**。给定连续的先验视觉帧，UniDWM生成未来视觉帧以及 h
相应的几何。我们 合并几何体并将其可视化为彩色点。

表*3*。 **消融研究**不同架构配置。结果在 NAVSIM *navtest* spl 上获得。 it.

自车位姿重建	外观重建	几何重建	动态生成	NC↑	DAC↑	EP↑	TTC↑	PDMS↑
✓	✗	✗	✗	96.5	88.3	74.3	90.5	78.5
✓	✓	✗	✗	97.6	90.1	77.1	91.6	81.3
✓	✗	✓	✗	96.8	90.7	76.3	91.1	80.0
✓	✗	✗	✓	96.8	90.7	76.3	91.1	80.9
✓	✗	✓	✓	97.3	90.7	76.8	91.7	81.6
✓	✓	✓	✓	97.3	91.6	77.4	92.1	82.4

表 *4.* **使用 GRPO 进行额外微调**。 表中第一行的模型3 用作基线。

方法	NC↑	DAC↑	EP↑	TTC↑	PDMS↑
基线	96.5	88.3	74.3	90.5	78.5
使用 GRPO 的基线	96.6	90.4	76.6	91.8	81.2
UniDWM	97.3	91.6	77.4	92.1	82.4
UniDWM（含 GRPO）	**97.4**	**93.2**	**80.1**	**93.3**	**84.9**

仅使用自车位姿重建时，PDMS 为 78.5。引入外观重建、几何结构重建和动态生成后，PDMS 分别提升 +2.8、 +1.5 和 +2.4。结合几何结构重建与动态生成后，性能进一步提升至 81.6 PDMS。结合所有重建和生成解码器时，整体性能最佳。这些发现证实，我们的多方面表征学习策略能够有效捕捉驾驶场景的互补特征，其中外观、几何和动力学提供了相互增强的监督信号。

为进一步探究潜在世界表征中编码的、对规划有用的信息，我们冻结 UniDWM 编码器，并使用 GRPO 算法微调轨迹解码器（Li 等人，2025c）。如表4所示，与经 GRPO 增强的基线相比，未使用 GRPO 的 UniDWM 已取得具有竞争力甚至更优的性能。对 UniDWM 应用 GRPO 后，模型取得了整体最佳结果，这表明统一潜在世界建模与策略优化带来的益处互为补充，而非彼此冗余。这些结果凸显了高质量的潜在世界建模不仅能直接提升规划性能，还能实现有效的基于强化学习的微调。表征的更多平滑性分析见附录D。

5. 结论

我们提出了 UniDWM，这是一种统一驾驶世界模型，可学习兼顾结构与动力学的潜在表征，将其作为自动驾驶中符合物理规律的状态空间。UniDWM 根据视觉观测联合重建场景外观、几何和自车运动，并通过基于扩散的生成过程对未来的世界演化建模，从而在单一潜在空间中实现感知、预测和规划之间的连贯推理。我们将 UniDWM 构建为变分自编码器的一种变体，为超越启发式设计选择的多方面表征学习提供了严谨的理论基础。在 NAVSIM 基准上的广泛评估表明，UniDWM 持续提升了轨迹规划性能，同时在 4D 重建和生成方面表现出色，验证了统一世界建模对面向规划的自动驾驶的有效性。

影响声明

本文旨在促进世界模型的学习与应用。我们的工作可能带来许多潜在的社会影响，但我们认为没有哪一项需要在此特别强调。

参考文献

Balestriero, R. 和 LeCun, Y. Lejepa：一种无需启发式方法、可证明且可扩展的自监督学习方法。*arXiv* 预印本 *arXiv:2511.08544*，2025年。

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J. 等。面向自动驾驶汽车的端到端学习。*arXiv* 预印本 *arXiv:1604.07316*，2016年。

Chen, J., Cai, H., Chen, J., Xie, E., Yang, S., Tang, H., Li, M., Lu, Y. 和 Han, S. 用于高效高分辨率扩散模型的深度压缩自编码器。发表于 *ICLR*，2025年。

Chen, S., Jiang, B., Gao, H., Liao, B., Xu, Q., Zhang, Q., Huang, C., Liu, W., and Wang, X. VADv2：通过概率规划实现端到端向量化自动驾驶。*arXiv*预印本 *arXiv:2402.13243*，2024年。

Codevilla, F., M¨uller, M., L´opez, A., Koltun, V., and Doso- vitskiy, A. 通过条件模仿学习实现端到端驾驶。载于 *ICRA*，2018年。

Codevilla, F., Santana, E., L´opez, A. M., and Gaidon, A. 探索行为克隆在自动驾驶中的局限性。载于国际计算机视觉大会，2019年。

Dauner, D., Hallgarten, M., Li, T., Weng, X., Huang, Z., Yang, Z., Li, H., Gilitschenski, I., Ivanovic, B., Pavone, M., Geiger, A., and Chitta, K. NAVSIM：数据驱动的非反应式自动驾驶车辆仿真与基准测试。载于*NeurIPS*，2024年。

Esser, P., Rombach, R., 和 Ommer, B. 驯服变换器以实现高分辨率图像合成。发表于*CVPR*，2021年。

Gao, R., Chen, K., Xiao, B., Hong, L., Li, Z., 和 Xu, Q. Magicdrive-v2：面向自动驾驶的高分辨率长视频生成与自适应控制。发表于 国际计算机视觉大会，2025年。

Gao, S., Yang, J., Chen, L., Chitta, K., Qiu, Y., Geiger, A., Zhang, J., 和 Li, H. VISTA：一种具有高保真度和灵活可控性的通用驾驶世界模型。发表于*NeurIPS*，2024年。

Guo, X., Wu, Z., Xiong, K., Xu, Z., Zhou, L., Xu, G., Xu, S., Sun, H., Wang, B., Chen, G. 等。Genesis：多模态drivi场景生成 tion with spatio-t_{empora} l and 跨模态一致性。*arXiv*预印本 *arXiv:2506.07497*，2025年。Hu, X., Yin, W., Jia, M., Deng, J., Guo, X., Zhang, Q., Long, X., 和 Tan, P. Drivingworld：通过视频GPT构建自动驾驶世界模型。*arXiv*预印本 *arXiv:2412.19505*，2024年。Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., 等。面向规划的自动驾驶。发表于 *CVPR*，2023。Huang, Y., Zheng, W., Zhang, Y., Zhou, J., 和 Lu, J. Gaussianformer：用于基于视觉的3D语义占用预测的场景即高斯。发表于 *ECCV*，2024年。

Jia, X., Wu, P., Chen, L., Xie, J., He, C., Yan, J., and Li, H. 驾驶前请三思：面向端到端自动驾驶的可扩展解码器。发表于*CVPR*，2023年。

Jia, X., You, J., Zhang, Z., and Yan, J. DriveTransformer：用于可扩展端到端自动驾驶的统一Transformer。发表于*ICLR*，2025年。

Jiang, B., Chen, S., Xu, Q., Liao, B., Chen, J., Zhou, H., Zhang, Q., Liu, W., Huang, C., and Wang, X. VAD：用于高效自动驾驶的矢量化场景表征。发表于国际计算机视觉大会，2023年。

Kendall, A. and Cipolla, R. 深度学习中的不确定性建模及其在相机重定位中的应用。发表于*ICRA*，2016年。

Kerbl, B., Kopanas, G., Leimk¨uhler, T., andDrettakis, G. 用于实时辐射场渲染的三维高斯泼溅。*TOG*，2023年。

Kerbl, B.、Meuleman, A.、Kopanas, G.、Wimmer, M.、Lan- vin, A. 和 Drettakis, G. 面向超大型数据集实时渲染的分层 3D 高斯表示。*TOG*，2024年。

Kingma, D. P. 和 Welling, M. 自编码变分贝叶斯。*arXiv* 预印本 *arXiv:1312.6114*，2013年。

Kong, W.、Tian, Q.、Zhang, Z.、Min, R.、Dai, Z.、Zhou, J.、Xiong, J.、Li, X.、Wu, B.、Zhang, J. 等。混元视频：面向大型视频生成模型的系统性框架。*arXiv* 预印本 *arXiv:2412.03603*，2024年。

Li, Y.、Fan, L.、He, J.、Wang, Y.、Chen, Y.、Zhang, Z. 和 Tan, T. 借助潜在世界模型增强端到端自动驾驶。在 *ICLR*，2025a。

Li, Y.、Wang, Y.、Liu, Y.、He, J.、Fan, L. 和 Zhang, Z. 基于鸟瞰图世界模型，通过在线轨迹评估实现端到端驾驶。在国际计算机视觉大会，2025b。

Li, Y.、Xiong, K.、Guo, X.、Li, F.、Yan, S.、Xu, G.、Zhou, L.、Chen, L.、Sun, H.、Wang, B. 等。Recogdrive：面向端到端自动驾驶的强化认知框架。*arXiv* 预印本 *arXiv:2506.08052*，2025c。

Li, Z., Li, K., Wang, S., Lan, S., Yu, Z., Ji, Y., Li, Z., Zhu, Z., Kautz, J., Wu, Z., 等。Hydra-mdp：通过多目标 Hydra 蒸馏实现端到端多模态规划。*arXiv* 预印本 *arXiv:2406.06978*，2024a。

Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Yu, Q., andDai, J. BEVFormer：通过时空变换器从激光雷达-摄像头学习鸟瞰视图表征。*IEEE* 模式分析与机器智能汇刊，2024b。

Liang, T., Xie, H., Yu, K., Xia, Z., Lin, Z., Wang, Y., Tang, T., Wang, B., and Tang, Z. BEVFusion：一种简单而稳健的激光雷达-摄像头融合框架。发表于 *NeurIPS*，2022年。Liao, B., Chen, S., Yin, H., Ji ang, B., Wang, C., Yan, S., Zhang, X., Li, X., Z hang, Y., Zhang, Q., 等。DiffusionDrive：用于端到端自动驾驶的截断扩散模型。发表于*CVPR*，2025年。Liu, S., Liang, Q., Li, Z., Li, B., and Huang, K. GaussianFusion：用于端到端自动驾驶的基于高斯的多传感器融合。发表于*NeurIPS*，2025年。Liu, Y., Wang, T., Zhang, X., and Sun, J. PETR：用于多视角 3D 物体检测的位置嵌入变换。发表于 *ECCV*，2022年。Novotny, D., Larlus, D., and Vedaldi, A. 从视频监督中捕捉物体类别的几何特征。*IEEE* 模式分析与机器智能汇刊，2018年。Prakash, A., Chitta, K., and Geiger, A. 用于端到端自动驾驶的多模态融合 Transformer。发表于*CVPR*，2021年。Russell, L., Hu, A., Bertoni, L., Fedoseev, G., Shotton, J., Arani, E., and Corrado, G. GAIA-2：面向自动驾驶的可控多视角生成式世界模型。*arXiv* 预印本 *arXiv:2503.20523*，2025年。Shi, C., Shi, S., Sheng, K., Zhang, B., and Jiang, L. DriveX：用于学习自动驾驶可泛化世界知识的全方位场景建模。发表于国际计算机视觉大会，2025年。Sim´eoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Rama- monjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., J´egou, H., Labatut, P., and Bojanowski, P. DINOv3。*arXiv* 预印本 *arXiv: 2508.10104*，2025年。

Song, Z., Jia, C., Liu, L., Pan, H., Zhang, Y., Wang, J., Zhang, X., Xu, S., Yang, L., and Luo, Y. 别晃动方向盘：端到端自动驾驶中的动量感知规划。发表于*CVPR*，2025年。

Wang, H. and Agapito, L. 利用空间记忆进行三维重建。发表于*3D*视觉会议，2025年。

Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., and Novotny, D. VGGT：视觉几何基础变换器。发表于 *CVPR*，2025年。

Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., and Revaud, J. Dust3r：让几何三维视觉变得简单。发表于 *CVPR*，2024年。

Wang, Y.、Guizilini, V. C.、Zhang, T.、Wang, Y.、Zhao, H.和Solomon, J. DETR3D：通过三维到二维查询从多视角图像进行三维目标检测。见机器人学习会议，2022年。Xing, Z.、Zhang, X.、Hu, Y.、Jiang, B.、He, T.、Zhang, Q.、Long, X.和Yin, W. GoalFlow：面向端到端自动驾驶中多模态轨迹生成的目标驱动流匹配。见 *CVPR*，2025年。Zhang, K.、Tang, Z.、Hu, X.、Pan, X.、Guo, X.、Liu, Y.、Huang, J.、Yuan, L.、Zhang, Q.、Long, X.、Cao, X.和Yin, W. Epona：面向自动驾驶的自回归扩散世界模型。见 国际计算机视觉大会，2025年。Zhang, R.、Isola, P.、Efros, A. A.、Shechtman, E.和Wang, O. 深层特征作为感知度量的不合理有效性。见*CVPR*，2018年。Zhao, S.、Song, J.和Ermon, S. InfoVAE：信息最大化变分自编码器。*arXiv*预印本 *arXiv:1706.02262*，2017年。Zheng, B.、Ma, N.、Tong, S.和Xie, S. 基于表征自编码器的扩散 Transformer。*arXiv*预印本 *arXiv:2510.11690*，2025a。Zheng, Y.、Yang, P.、Xing, Z.、Zhang, Q.、Zheng, Y.、Gao, Y.、Li, P.、Zhang, T.、Xia, Z.、Jia, P.等。World4Drive：基于意图感知物理潜在世界模型的端到端自动驾驶。见 国际计算机视觉大会，2025b。

A. UniDWM 的变分自编码器式证据下界目标证明

A.1. 问题设定

我们考虑一个数据集

$$\mathcal{D} = \{\mathbf{x}_i \in \mathcal{X}\}_{i=1}^{N},$$

其中，每个场景$\mathbf{x}_i$ 由来自不同视角的 M 个观测组成，

$$\mathbf{x}_i = \{\boldsymbol{x}_i^{(j)}\}_{j=1}^{M}.$$

每个观测 $\boldsymbol{x}^{(j)}$ 对应一个帧–模态对。我们假设场景是从未知数据分布 $p_{\mathcal{D}}(\mathbf{x})$ 中独立同分布采样得到的。

给定局部观测$\mathbf{x}^{local} \subset \mathbf{x}$的一个子集，UniDWM 的目标是重建完整的全局观测集合$\mathbf{x}$。令$\mathbf{z} \in \mathcal{Z}$ 表示一个潜变量，用于捕获全局场景表征。

生成模型定义为

$$p_\theta(\mathbf{x}, \mathbf{z}) = p(\mathbf{z}) \prod_{j=1}^{M} p_\theta(\boldsymbol{x}^{(j)} \mid \mathbf{z}),$$

其中假设在给定$\mathbf{z}$ 的条件下，观测相互独立。近似后验表示为

$$q_\phi(\mathbf{z} \mid \mathbf{x}^{local}).$$

A.2. 变分下界

我们从完整观测的对数似然开始:

$$\log p_\theta(\mathbf{x}) = \log \int p_\theta(\mathbf{x}, \mathbf{z})\, d\mathbf{z}.$$

引入变分后验 $q_\phi(\mathbf{z} \mid \mathbf{x}^{local})$，我们将边际似然改写为

$$\log p_\theta(\mathbf{x}) = \log \int q_\phi(\mathbf{z} \mid \mathbf{x}^{local}) \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x}^{local})}\, d\mathbf{z}.$$

应用詹森不等式可得以下下界:

$$\log p_\theta(\mathbf{x}) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^{local})} \left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x}^{local})} \right].$$

代入因式分解的联合分布，我们得到

$$\log p_\theta(\mathbf{x}) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^{local})} \Bigg[\sum_{j=1}^{M} \log p_\theta(\boldsymbol{x}^{(j)} \mid \mathbf{z}) + \log p(\mathbf{z}) - \log q_\phi(\mathbf{z} \mid \mathbf{x}^{local}) \Bigg].$$

重新整理各项可得

$$\log p_\theta(\mathbf{x}) \geq \sum_{j=1}^{M} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^{local})} \big[\log p_\theta(\boldsymbol{x}^{(j)} \mid \mathbf{z}) \big] - \mathrm{KL}\big(q_\phi(\mathbf{z} \mid \mathbf{x}^{local}) \,\|\, p(\mathbf{z}) \big).$$

A.3. 数据集级目标

对数据分布 $p_{\mathcal{D}}(\mathbf{x})$ 取期望后，我们得到数据集层面的证据下界：

$$\mathcal{L}_{\text{UniDWM0}} \equiv \sum_{j=1}^{M} \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^{local})} \big[\log p_\theta(\boldsymbol{x}^{(j)} \mid \mathbf{z})\big] - \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})} \mathrm{KL}\big(q_\phi(\mathbf{z} \mid \mathbf{x}^{local}) \,\|\, p(\mathbf{z})\big)\,,$$

这与方程4一致。

B. UniDWM 的 InfoVAE 风格目标的证明

在本节中，我们推导 UniDWM 中使用的受 InfoVAE 启发的目标，并说明它如何从标准变分自编码器证据下界的分解中得出。

B.1. 背景：变分自编码器证据下界的局限性

回顾附录A中推导的标准证据下界：

$$\mathcal{L}_{\text{UniDWM0}} = \sum_{j=1}^{M} \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^{local})} \big[\log p_\theta(\boldsymbol{x}^{(j)} \mid \mathbf{z})\big] - \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})} \mathrm{KL}\big(q_\phi(\mathbf{z} \mid \mathbf{x}^{local}) \,\|\, p(\mathbf{z})\big)\,.$$

当观测空间 $\mathcal{X}$ 复杂且维度很高时，KL正则化项可能会促使近似后验 q_ϕ（$\mathbf{z} \mid \mathbf{x}^{local}$）向先验 p（$\mathbf{z}$）坍塌，从而导致无信息的潜在表示。这种现象通常称为后验坍塌。

为缓解这一问题，InfoVAE（Zhao等人，2017年）提出了一种替代性的正则化方法，显式控制聚合后验 q_ϕ（$\mathbf{z}$），而非条件后验。

B.2. KL项的分解

我们将聚合后验定义为

$$q_\phi(\mathbf{z}) \triangleq \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})} \big[q_\phi(\mathbf{z} \mid \mathbf{x}^{local})\big],$$

其中，$\mathbf{x}^{local}$ 是通过固定的确定性采样策略从$\mathbf{x}$ 中获得的。

标准证据下界中的期望 KL 散度可以分解为

$$\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})} \mathrm{KL}\big(q_\phi(\mathbf{z} \mid \mathbf{x}^{local}) \,\|\, p(\mathbf{z})\big) = I_q(\mathbf{z}; \mathbf{x}^{local}) + \mathrm{KL}(q_\phi(\mathbf{z}) \,\|\, p(\mathbf{z}))\,,$$

其中

$$I_q(\mathbf{z}; \mathbf{x}^{local}) \triangleq \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})} \mathrm{KL}\big(q_\phi(\mathbf{z} \mid \mathbf{x}^{local}) \,\|\, q_\phi(\mathbf{z})\big)$$

是变分联合分布下 $\mathbf{z}$ 与 $\mathbf{x}^{local}$ 之间的互信息。

将此分解代入证据下界可得

$$\mathcal{L}_{\text{UniDWM0}} = \sum_{j=1}^{M} \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^{local})} \big[\log p_\theta(\boldsymbol{x}^{(j)} \mid \mathbf{z})\big] - I_q(\mathbf{z}; \mathbf{x}^{local}) - \mathrm{KL}(q_\phi(\mathbf{z}) \,\|\, p(\mathbf{z}))\,.$$

该形式表明，标准证据下界会隐式惩罚互信息 $I_q(\mathbf{z};\mathbf{x}^{local})$，从而抑制潜变量编码关于观测的有意义信息。

B.3. InfoVAE 式目标

InfoVAE（Zhao等人，2017年）提出了一个广义目标，允许显式控制互信息项与聚合后验正则化的相对重要性。一般的InfoVAE目标可以等价地写为

$$\mathcal{L}_{\text{InfoVAE}} = \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \big[\log p_\theta(\mathbf{x} \mid \mathbf{z})\big] - (1-\alpha)\, I_q(\mathbf{z}; \mathbf{x}) - \lambda D(q_\phi(\mathbf{z}) \,\|\, p(\mathbf{z}))\,,$$

其中 $D(\cdot\|\cdot)$是一个一般散度，满足 $D(q_\phi(\mathbf{z})\|p(\mathbf{z}))=0$ 当且仅当 $q_\phi(\mathbf{z})=p(\mathbf{z})$， α 和 λ 是缩放系数。

在 UniDWM 中，我们采用 InfoVAE 公式，并设置 $\alpha=1$。因此，互信息惩罚被完全移除，目标函数简化为

$$\mathcal{L}_{\text{UniDWM}}=\sum_{j=1}^{M}\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}^{local})}\big[\log p_\theta(\boldsymbol{x}^{(j)}\mid\mathbf{z})\big] - \lambda\, D(q_\phi(\mathbf{z})\,\|\,p(\mathbf{z}))\,.$$

因此，UniDWM 的目标函数可视为在多观测重建设置下，将 InfoVAE 直接实例化并扩展至 $\alpha=1$。

C. 数据集与评估指标

我们主要在 NAVSIM（Dauner 等人，2024年）数据集上评估模型。该数据集提供 120 小时的高难度驾驶数据，包含高分辨率摄像头图像和最长达 1.5 秒的激光雷达输入。该数据集会过滤掉简单的驾驶场景，以突出复杂的决策任务。对于轨迹规划任务，我们采用官方的预测驾驶员模型分数（PDMS）作为评估指标。对于三维重建任务，遵循先前的研究（Wang 等人，2024年；2025年），我们报告准确度（预测结果到真值的最小欧氏距离）、完整度（真值到预测结果的最小欧氏距离），以及二者的平均值 Overall（即 Chamfer距离）。

D. 学习表征的平滑性分析

我们对基线模型（与表4中相同）和 UniDWM 在 NAVSIM 测试划分上得到的潜在表示进行了平滑性分析。此项分析旨在评估学习到的表征有多平滑，并探究其与下游规划任务中的鲁棒性和泛化能力之间可能存在的相关性。对于每个模型，我们提取测试划分中所有样本的潜在表示，并进行逐特征标准化。随后，我们在表征空间中使用 k个近邻（实际使用$k=32$ ）构建局部邻域，并计算以下互补的平滑度指标:

- **k近邻平均距离**（kNN 距离）：到 k 个近邻的平均距离，用于衡量潜在空间中的局部敏感度。
- **局部PCA能量比**（PCA 比率）：每个局部邻域中由首要主成分解释的方差比例，反映局部线性。
- **图拉普拉斯平滑度**（拉普拉斯平滑度）：在 k近邻图上计算的图拉普拉斯平滑度，用于衡量潜在流形上的一致性。

基线模型和 UniDWM 的平滑度指标汇总于表 5。与基线相比，UniDWM 实现了：1）更低的 k近邻平均距离，表明潜在空间中的邻近样本距离更近，局部变化更小；2）更高的局部PCA能量比，说明局部邻域可以由低维线性子空间更好地近似；3）更低的图拉普拉斯平滑度，反映出潜在流形上的变化更加平滑。这些结果一致表明，UniDWM 比基线生成了更加平滑的潜在表征，预计有助于实现更稳定的潜在展开，并提升下游任务中的鲁棒性和泛化能力。

表 5. 平滑度指标 NAVSIM测试划分上的潜在表示。较低的 kNN 距离和拉普拉斯平滑度，以及 d 较高的 PCA 比率，表示表征更 平滑。

模型	kNN 距离。	PCA 比率	拉普拉斯平滑度
基线	130.578	0.554	328,596
UniDWM	117.238	0.622	263,981

E. 局限性

我们的框架主要基于序列单目视觉输入进行训练和评估，并未明确探索多视角视觉或多传感器输入，例如激光雷达和雷达。其次，尽管基于扩散的生成模块能够对未来世界动态进行富有表现力且灵活的建模，但其在训练过程中引入了不可忽视的计算开销。第三，UniDWM 目前在有限的时间范围内对未来演化进行建模（实践中采用下一帧预测）。我们将多模态、轻量化和长时域框架的探索留待未来工作。