潜空间中的几何学：构建结构化的世界模型

从神经几何到等变表示

Press Space for next page →

目录

核心问题

学习到的潜变量动作（Latent Action）应该栖息在什么几何空间?

挑战

- 若无归纳偏置 → 网络学“捷径”
- 潜空间高度非线性、纠缠、不可组合
- 长程模拟迅速失稳（Total Chaos）

灵感

- 大脑（网格细胞、环面拓扑）预置高维几何
- 既是“容器”也是“计算器”

目标（模型期望）

- **可乘组合性**
  - 泛化到未见动作组合
- **可解释性**
  - 动作之间的代数关系可读
- **稳定性**
  - 长程预测不崩溃

神经科学视角

核心结论

- 大脑通过**流形展平（Manifold Flattening）**使操作线性化，便于预测与规划。
- **后继表征（SR）+ 谱分析**：网格细胞可视为预测图谱的低维特征向量（Eigenvectors）。

Tolman–Eichenbaum Machine (TEM)

- 因式分解（状态向量 **g** 与循环权重矩阵）
- 操作表现为对状态的线性变换（path integration via weight matrices）。

启示

将非线性环境动力学映射到便于线性预测与组合的潜空间是生物系统的有效策略。

等变表示：基本思想

核心公式

希望存在编码器 f 及线性表示 M_g，使得

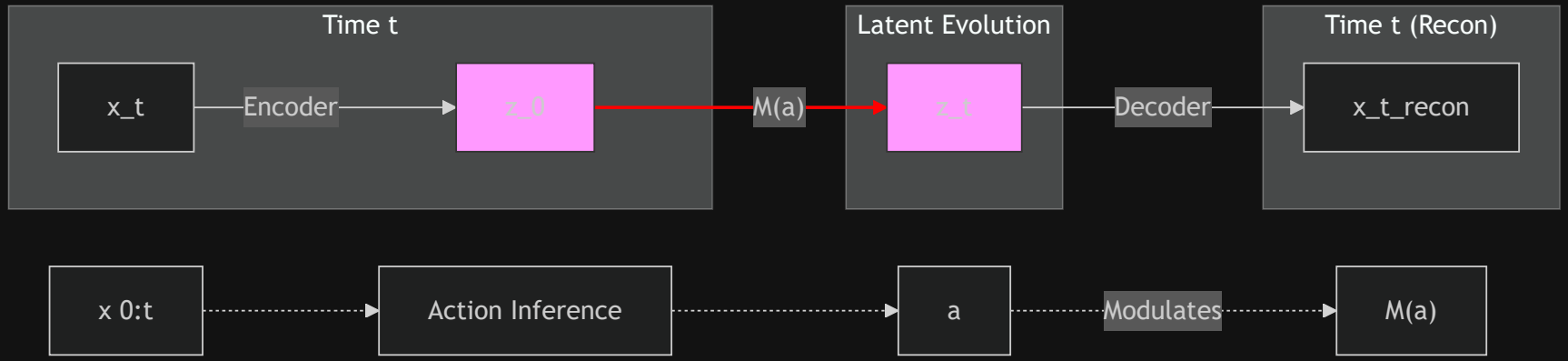
$$f(\mathcal{T}_g x) = M_g f(x)$$

其中 M_g 为群 G 的线性表示（等变性）。

含义

像素空间的复杂变换 $\mathcal{T}_g$ 在潜空间成为线性操作 M_g，便于组合与推断。

结构示意



$x_0 \xrightarrow{\text{Encoder}} z_0 \xrightarrow{M(a)} z_t \xrightarrow{\text{Decoder}} x_t$，动作 a 从观测序列推断，并调制算子 $M(a)$。

表示为 $E(x_{t+1}) = M(a_t)E(x_t)$

两大数学流派概览

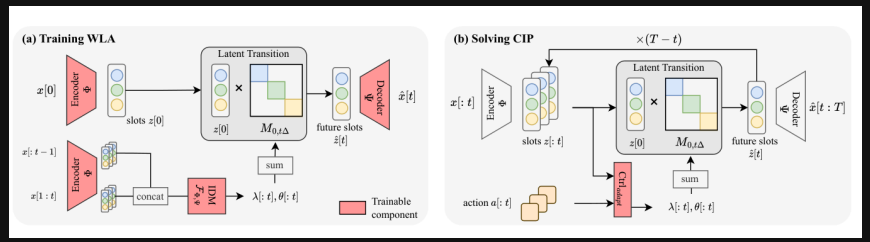
流派一：群表示论（Group Theory）

- 关心数据固有对称性
- 将群元素 g 映射为可逆线性变换 $\rho(g)$ $\phi(g \cdot x) = \rho(g)\phi(x)$
- 同态性质 $\rho(g_1 g_2) = \rho(g_1)\rho(g_2)$
- **连续群**：用生成元（李代数）与指数映射 $\rho(g) = \exp(A \cdot t)$
- **优缺**：参数极度压缩、可解释；但难以处理不可逆/耗散过程与遮挡
- **与几何深度学习 (GDL) 的区别**： GDL 通常硬编码对称性，而我们试图**学习**未知环境的对称群结构。

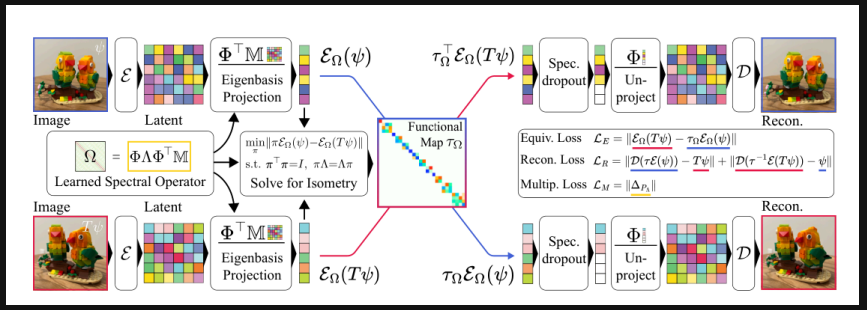
流派二：Koopman 算子理论

- 关注非线性动力系统的线性化（通过升维到可观测函数空间）
- 存在无限维线性算子 $\mathcal{K}$ 使 $g(x_{t+1}) = \mathcal{K} g(x_t)$
- **实践**：学习有限维观测函数 $\psi(x)$，使 $z_{t+1} \approx K z_t$（必要时引入 action 依赖的 $\mathcal{K}_t(a)$）
- **优缺**：强动力学导向；但可能需要升维与选择合适观测函数

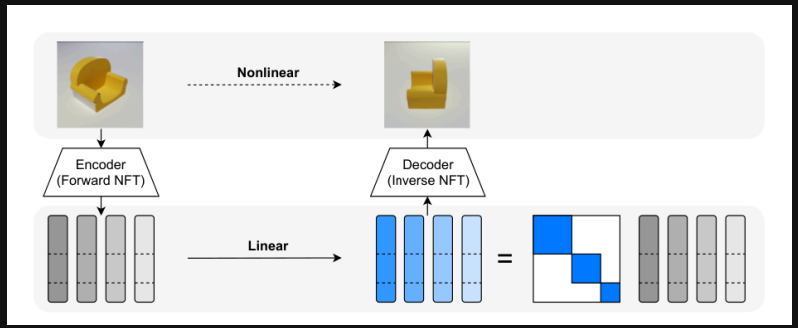
现有代表性工作



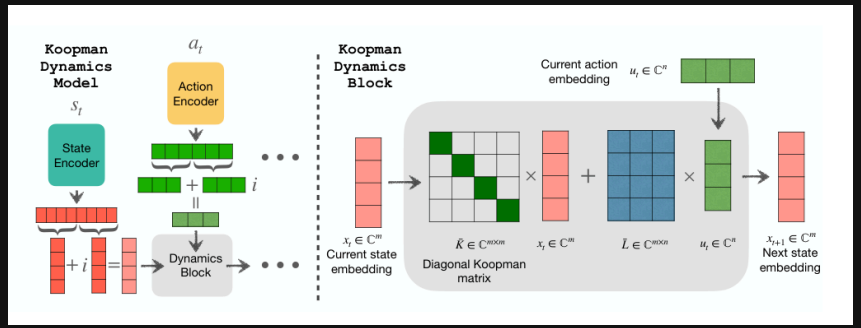
WLA (Hayashi et al., 2025)



NIso (Mitchel et al., 2024)



NFT (Koyama et al., 2023)



Koopman (Mondal et al., 2024)

五类核心约束

为防止网络学出平凡解，现有工作引入五类约束：

1. **构建类 (Parameterization / Architecture)**
2. **稀疏类 (Sparsity / Block Structure)**
3. **范数类 (Norm / Stability)**
4. **等变性误差 (Equivariance Loss / Consistency)**
5. **对齐类 (Alignment / Inverse-Consistency / Triplet)**

构建类

目的：解决“如何构造/计算这个矩阵”的问题，决定了模型的计算图结构。

论文	实现方式	描述
Koopman	**特征值参数化**	通过学习复数特征值 $\lambda = e^{\mu+i\omega}$，利用范德蒙德矩阵将序列预测转化为卷积。
NFT	**解析解求解器**	不反向传播 M 的梯度，直接在前向过程中用 `lstsq` 或伪逆算子计算 M。
NIso	**谱分解重构**	不直接存 $N \times N$ 矩阵，而是存 Φ (特征向量) 和 Λ (特征值)，通过 $\Omega = \Phi\Lambda\Phi^\top$ 重构算子。
WLA	**连续时间积分**	利用指数映射 $\exp(\cdot)$ 将李代数（切空间）映射回李群（流形），实现连续时间建模。

稀疏类

目的：降低自由度，解耦维度，赋予潜变量独立的物理/频率意义。同时工程上方便并行, 减少显存占用

论文	实现方式 (硬约束/软约束)	数学表达式 / 形式
Koopman	**[硬] 对角化** 强制矩阵 K 只有对角线元素。	$K = \text{diag}(\lambda_1, \dots, \lambda_m), \quad \lambda \in \mathbb{C}$
NFT	**[硬] 块对角化 (Irreps)** 根据群论先验，强制 M 为不可约表示的直和,。	$M(g) = \bigoplus_k R_k(g), \quad R_k \in \mathbb{R}^{d_k \times d_k}$
NIso	**[软] 重数损失 (Multiplicity Loss)** 惩罚特征值 λ 之间的相似度。若特征值互异，则与对角阵交换的矩阵必须也是对角阵。	$\mathcal{L}_M = \lvert\Delta_{P_\Lambda}\rvert = \lvert\text{diag}(P_\Lambda \mathbf{1}) - P_\Lambda\rvert$ 其中 $(P_\Lambda)_{ij} = \exp(-\lvert\lambda_i - \lambda_j\rvert)$
WLA	**[硬] 李代数块对角化** 强制生成元 $A(t)$ 为 2×2 块对角结构（缩放+旋转）。	$A_k(t) = \begin{pmatrix} \lambda_k(t) & -\theta_k(t) \\ \theta_k(t) & \lambda_k(t) \end{pmatrix}$

范数类

目的：控制算子的能量（稳定性），防止梯度爆炸/消失，或诱导特定的参数分布。相当于矩阵空间中的lasso正则化

论文	实现方式 (硬约束/软约束)	数学表达式 / 形式
Koopman	**[硬] 实部剪裁 (Real Part Clamping)** 限制特征值实部 $\mu \leq 0$，防止长程预测数值爆炸。μ 是可学习参数	$\mathrm{Re}(\lambda_j) \in [-0.3, 0]$ (Clip or fixed)
NFT	**[硬] 幺正性 (Unitarity)** 对于紧致群，假设 M 是幺正/正交的（即保留模长）。	$M^\top M = I$ (Implied via Irreps definition)
NIso	**[硬] 正交投影 (Procrustes)** 强制潜空间变换 τ 为严格的正交矩阵。	$\tau^* = \mathcal{K}(M) = UV^\top, \quad \text{where } M = U\Sigma V^\top$
WLA	**[软] L1 稀疏惩罚** 惩罚李代数参数的绝对值总和，促使瞬时动作仅由少量	$\mathcal{L}_1 = \sum_t (\lvert\lambda(t)\rvert_1 + \lvert\theta(t)\rvert_1)$

等变性误差

目的：核心损失函数，强制“观测空间的变换”等价于“潜空间的线性变换”。

论文	实现方式	数学表达式
Koopman	**一致性损失** 多步预测后的潜变量/解码结果需与真值一致。	$\mathcal{L}_{\text{cons}} = \sum_k \lvert\hat{x}_{t+k} - x_{t+k}\rvert^2 + \lvert\Psi(K^k\Phi(x_t)) - x_{t+k}\rvert^2$
NFT	**最小二乘误差 (MSP)** 既然 M 是算出来的最优解，那么该误差衡量编码器是否提取了线性特征。	$\mathcal{L} = \mathbb{E}\lvert x_2 - \Psi(M^*\Phi(x_1))\rvert^2, \quad M^* = \Phi(x_2)\Phi(x_1)^\dagger$
NIso	**重建损失** 变换后的特征与特征变换后的结果需一致。	$\lvert D(\tau E(\psi)) - T\psi\rvert$
WLA	**前向预测损失** 积分后的状态需匹配未来的观测。	$\mathcal{L}_{\text{fwd}} = \sum_t \lvert x[t] - \hat{x}_f[t]\rvert^2, \quad \hat{x}_f[t] = \Psi(\exp(\int A)z_0)$

对齐类

目的：利用群的逆元性质或双向一致性，强约束空间的几何结构，防止坍缩或退化。

论文	实现方式	数学表达式 / 机制
NIso	**等变损失** 使用 (x, Tx, T^2x) 三元组保证结合律	$\mathcal{L}_B = \lvert\mathcal{E}_\Omega(\psi) - \tau_\Omega^\top \mathcal{E}_\Omega(T\psi)\rvert^2 + \mathcal{L}_E = \lvert\tau_\Omega \mathcal{E}_\Omega(\psi) - \mathcal{E}_\Omega(T\psi)\rvert^2$
WLA	**反向预测损失** 既然 $z(t+\delta) = \exp(A\delta)z(t)$，那么必须满足 $z(t) = \exp(-A\delta)z(t+\delta)$。这强制了正向动作和反向动作在代数上的对称性。	$\mathcal{L}_{\text{bwd}} = \sum_t \lvert x[t] - \hat{x}_b[t]\rvert^2$ $\hat{x}_b[t] = \Psi\left(\exp\left(-\Delta \sum_{k=t}^{T} A[k]\right) z[T]\right)$
NIso	**Triplet Loss** 使用 (x, Tx, T^2x) 三元组保证结合律, 强化一致性	$\lvert\tau_{2,\Omega}\tau_{1,\Omega} E_\Omega(\psi) - E_\Omega(T^2\psi)\rvert$

实验验证

检验“学到的几何”

验证重点（不仅看 MSE）

1. **潜空间表示是否对应真实几何基底?**
   - 频率 / 球谐 / 拉普拉斯本征等
2. **潜空间算子是否呈现块对角、对称、可加性等结构?**
3. **是否能通过潜空间代数操作实现组合/插值/反向推断?**

实验 1：Neural Fourier Transform (NFT)

逻辑

- 模型能否提取真实频率表示（类似傅里叶变换）？

实验设置

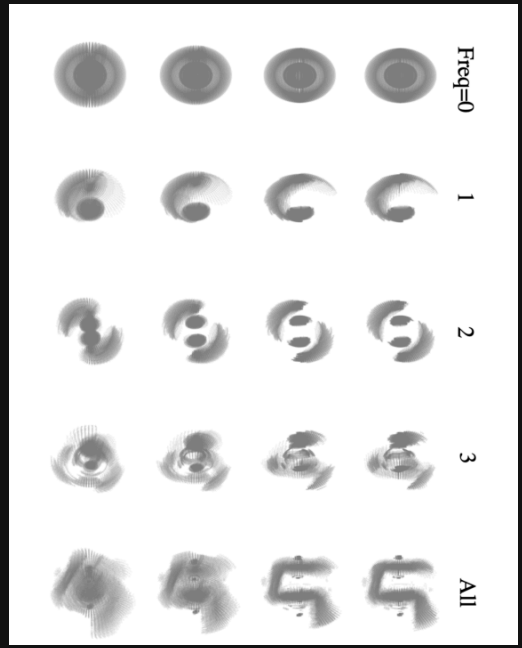
- 3D Structure from 2D（ModelNet）
- 测试从未见角度的 2D 视图预测

结果

- 潜空间表示与 3D 空间的**球谐函数**分解一致

结论

- 模型仅凭 2D 视图自发学到 3D 几何基底



实验 2：Neural Isometries (NIso)

逻辑：检验算子 Ω 是否近似流形上的拉普拉斯-贝尔特拉米算子。

实验：环面 / 球面 平移/旋转 — 可视化 Ω 结构与潜空间变换矩阵。

结果要点

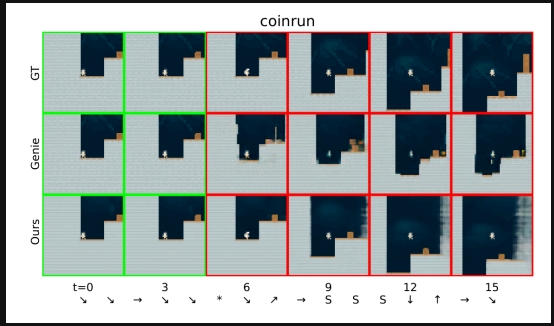
- Ω 结构类似离散拉普拉斯（如 5 点差分）
- 潜空间变换呈 **块对角**，块大小匹配 $SO(3)$ Wigner-D 矩阵结构（$1, 3, 5, \ldots$）

结论：模型在无监督下重发现了流形基底与群表示

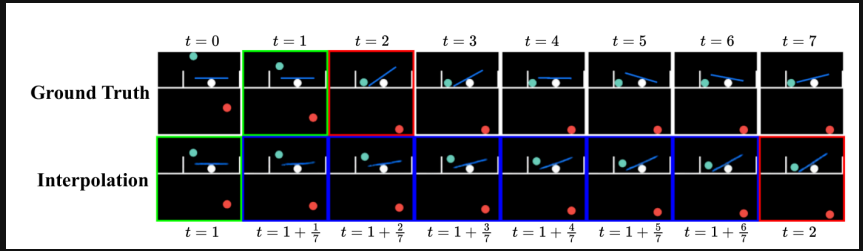
实验 3：World modeling through Lie Action (WLA)

逻辑：验证潜空间动作是否具备**组合性、连续性、可加性。**

操作：在潜空间代数相加 $\text{Action}_{total} = \text{Action}_{red} + \text{Action}_{blue}$，并对李代数参数插值以升频。



跨步预测显示合成效应



插值生成高帧率平滑动画

结论：潜变量可解释为物理上的可加力或速度向量。

关键结论

- **学到结构化潜空间能带来**
  - 更强的泛化、可解释性与长程稳定性。
- **两条可互补路径**
  - 学习群结构（低维、可解释）
  - Koopman 升维得到线性动力学（动力学导向）
- **五类约束**
  - 构建/稀疏/范数/等变误差/对齐 — 实务上需组合“软/硬约束”。
- **实验表明**
  - 不同方法能在无监督下自发重构物理或几何基底。

下一步工作与开放问题

- **模型融合**：将等变组件（群/李代数）与现有认知模型架构对接。
- **约束优化**：设计工程上更稳健的软硬约束组合。
- **结构探究**：深入评估对齐类损失对长程预测稳定性的贡献。
- **开放问题**：如何处理不可逆/耗散过程、遮挡与高维微观动力学的等变表示?

参考文献

Stachenfeld et al. (SR, 2017, Nature Neuroscience)

Whittington et al. (TEM, 2020, Cell)

Koyama et al. (NFT, ICLR 2023)

Mitchel et al. (NIso, NeurIPS 2024)

Mondal et al. (Koopman, ICLR 2024)

Hayashi et al. (WLA, arXiv 2025)

Thanks

Q & A