2026.1.29组会

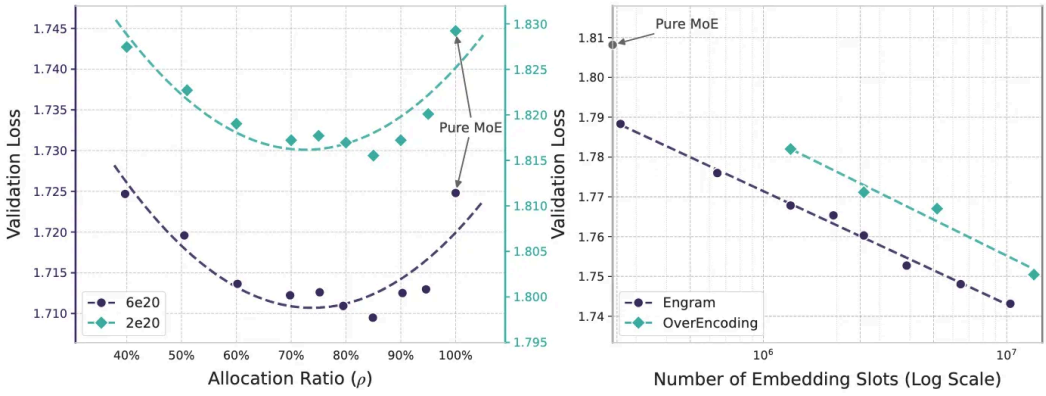


Figure 3 | **Sparsity allocation and Engram scaling.** Left: Validation loss across allocation ratios ρ. Two compute budgets are shown (2e20 and 6e20 FLOPs). Both regimes exhibit a U-shape, with hybrid allocation surpassing Pure MoE. Right: Scaling behavior in the infinite-memory regime. Validation loss exhibits a log-linear trend with respect to the number of embeddings.

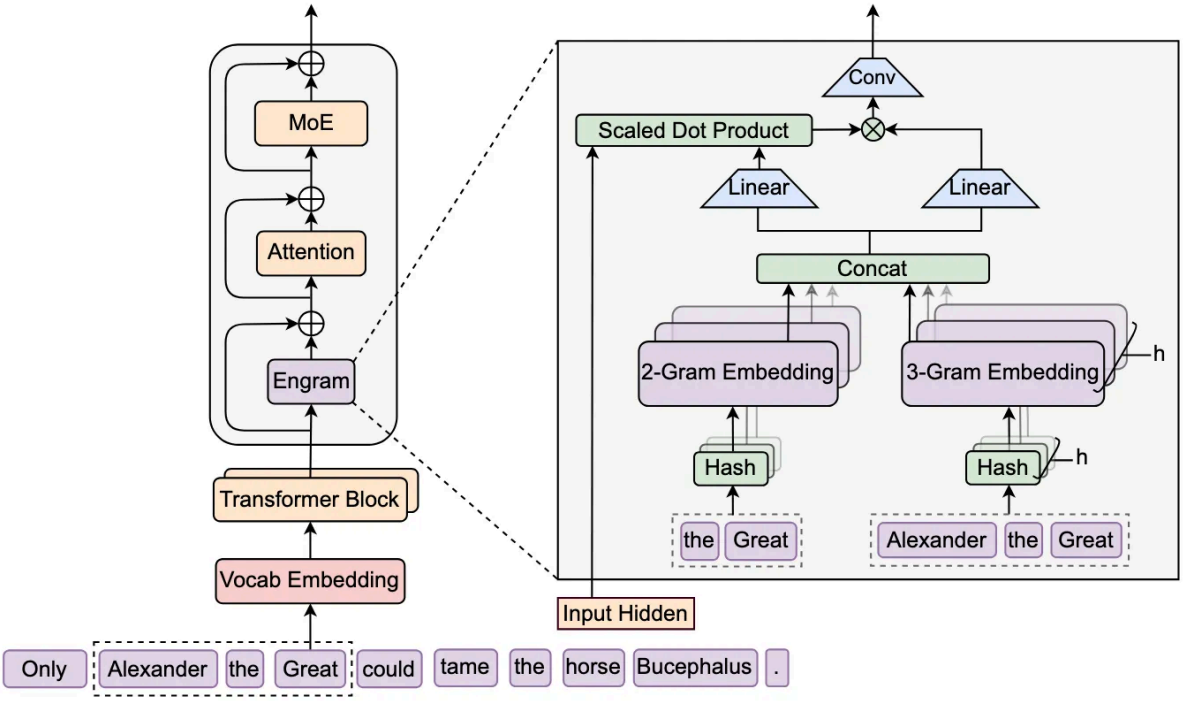
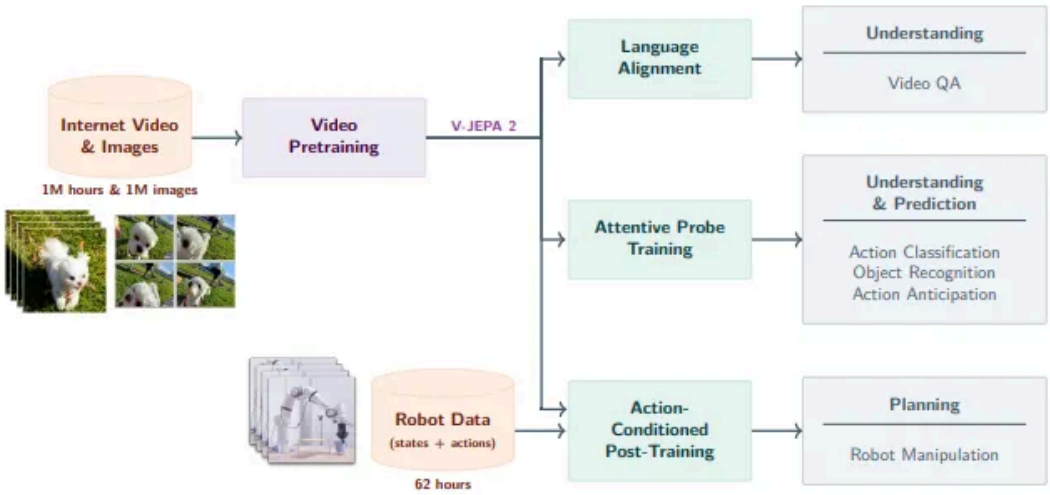
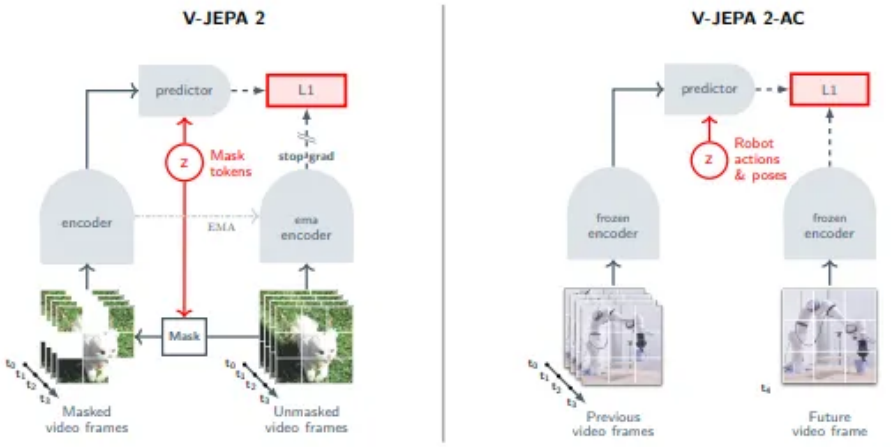


Figure 6 | **Visualization of the gating mechanism of Engram.** The heatmap intensity corresponds to the magnitude of the gating scalar $\alpha_t \in [0, 1]$, where darker red indicates stronger activation. Because Engram operates on suffix N-grams (here $N = 3$), a high activation on a specific token x_t implies that the preceding tokens culminating in that token (e.g., the phrase ending at t) are recognized as a static pattern effectively retrieved from memory.





1. Motivation: The "Reasoning" vs. "Retrieval" Dilemma

当前基于联合嵌入预测架构（JEPA）的世界模型在物理理解上取得了巨大突破，但仍受限于**自回归生成的本质缺陷**。

- **Parametric Reasoner (参数化推理的局限):** V-JEPA 2 本质上是一个自回归模型。在进行长时程（Long-horizon）预测时，模型完全依赖神经网络的权重进行连续推理。
- **数学困境：** 假设单步预测误差为 （即使 极小），在 步之后的累积误差将接近 （线性）甚至 （在非线性动力系统中呈指数级）。
- **物理后果：** 这种**灾难性漂移 (Catastrophic Drift)** 会导致预测出的特征向量 逐渐偏离真实的物理流形（Physical Manifold），表现为物体凭空消失、穿模或运动轨迹违背牛顿定律。
- **Non-Parametric Memorizer (非参数化记忆的机遇):** 物理世界具有极高的**时空冗余性 (Spatiotemporal Redundancy)**。
- **直觉：** “杯子从桌上掉落”这个物理过程，在数百万小时的视频中出现了无数次。对于这种高频事件，模型不需要每次都费力去“推理”微分方程，只需要“检索”过去见过的标准轨迹即可。
- **Engram 的启发：** DeepSeek 的 Engram 论文证明，将部分计算负担转移给静态的查表（Lookup）机制，可以在不增加 FLOPs 的前提下显著提升模型的困惑度表现。

Core Hypothesis (核心假设): 通过引入非参数化记忆模块（Engram），我们可以将 V-JEPA 的长时程预测问题，转化为“推理”与“检索”的混合问题，从而利用记忆的**Ground-truth**性质来锚定（Anchor）推理的漂移。

2. From Latent Space to Memory Space

构建一个混合架构 **Mem-JEPA**。核心设计原则是“冻结骨干，训练索引策略”。

Step 1: Latent Quantization & Tokenization

V-JEPA 的 Encoder 输出是连续的高维向量 。为了利用 Engram 的高效哈希查表，必须将其离散化。

- **技术方案：** 引入一个轻量级的 **Residual Vector Quantizer (RVQ)** 或 **Product Quantization (PQ)** 模块。

- **流程：**

1. **输入：** 接收 V-JEPA Encoder 提取的连续特征 。
2. **量化：** 将 映射到最近的 Codebook 向量索引上。
3. **序列化：** 将连续的视频流转化为离散的 Token 序列（Visual N-grams）。

- *Example:* 物理状态 ，其中 是量化后的 Codebook ID。

- **训练目标：** 只训练 Quantizer 的 Codebook，使其重构误差最小化，确保离散 Token 能保留 V-JEPA 的语义信息。

Step 2: Sparse Retrieval via Engram

这是系统的“海马体”。我们构建一个基于哈希的键值对存储（Key-Value Store）。

- **Key (Query):** 过去 帧的离散 Token 序列（Visual N-gram）。
- **Value (Target):** 对应未来的真实状态向量 （来自 Teacher 模型或训练数据）。
- **O(1) 查表机制：**
- 利用 Engram 的哈希函数 直接定位内存地址。
- **稀疏性策略 (Sparsity Policy):** 并非每一步都查表。
- 设置一个**置信度阈值**。只有当当前场景在记忆库中出现频率极高（High Frequency Count）时，才触发检索。
- **Hit (命中):** 直接输出记忆向量 ， 跳过 V-JEPA Predictor 的计算（加速推理）。
- **Miss (未命中):** 回退到 V-JEPA Predictor 进行常规推理。

Step 3: Drift Rectification & Manifold Projection

这是解决“幻觉”的关键。当 V-JEPA Predictor 在工作时，利用 Engram 作为“纠错器”。

- **问题定义：** V-JEPA 的预测结果 可能带有累积误差。
- **流形投影：** Engram 返回的向量 代表了训练数据中真实存在的物理状态，它天然位于真实物理流形 上。
- **门控残差融合 (Gated Residual Fusion):**
  设计一个**自适应融合层 (Adapter Layer)** 来结合两者的输出：
- V-JEPA 的自回归预测（含有漂移）。
- Engram 检索到的历史真实状态（无漂移，但可能有场景偏差）。
- 这是一个可学习的门控系数（Gating Factor, ）。模型会自动学习： **“什么时候该信自己的推理，什么时候该信记忆的经验”**。

3. Implementation & Feasibility

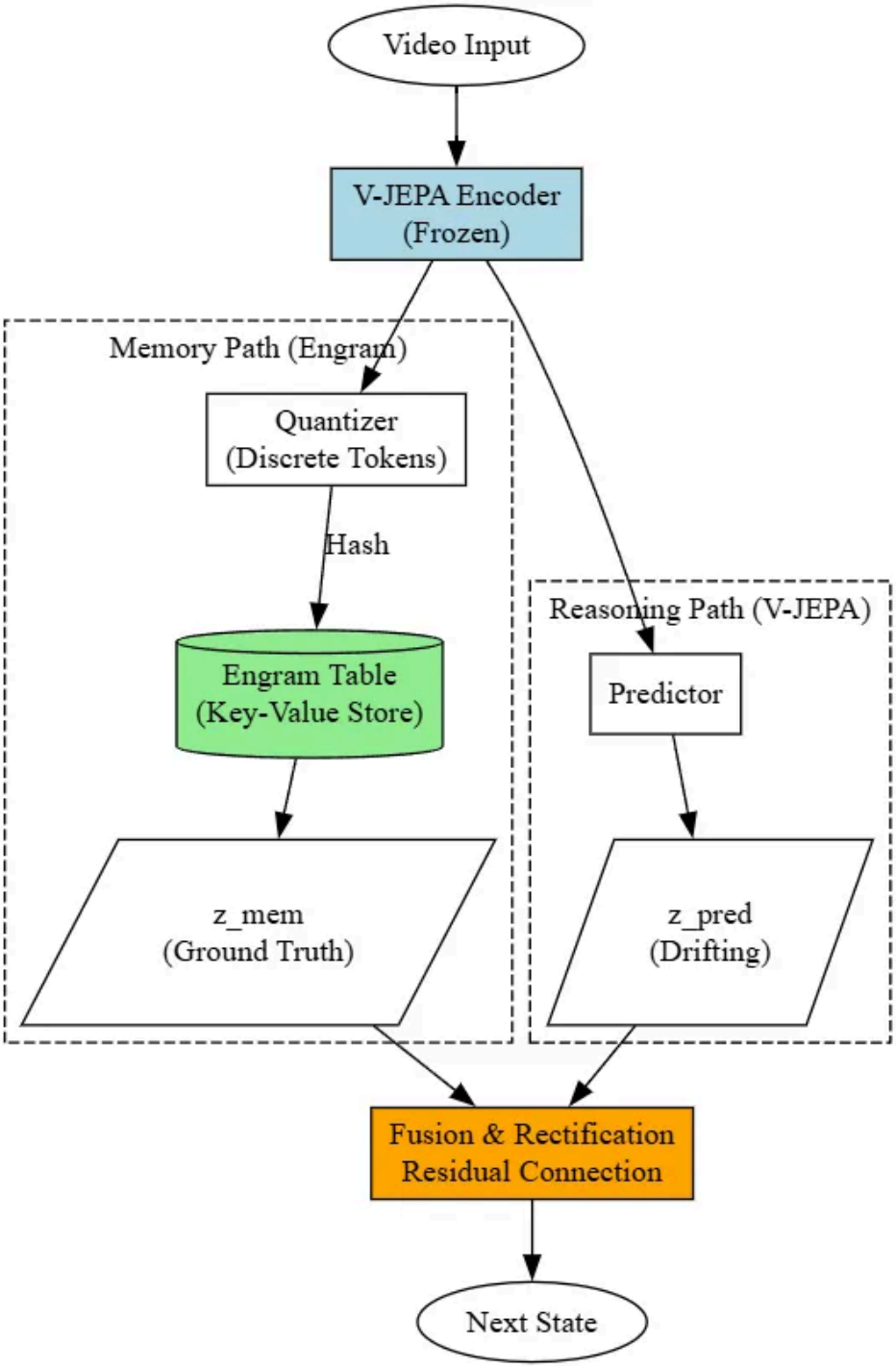
3.1 训练策略

为了避免高昂的计算成本，采用分阶段训练：

1. **Stage 1 (Freeze Backbone):** 冻结 V-JEPA 2 (ViT-H)，只使用其作为特征提取器。
2. **Stage 2 (Train Quantizer):** 在目标数据集（如 Something-Something v2）上训练 RVQ/PQ Codebook，确保离散化质量。
3. **Stage 3 (Train Engram & Adapter):**
   - 构建 Engram 表（这一步是统计过程，极快）。
   - 训练轻量级的 Adapter (MLP) 和 Gating Network，目标是最小化长时程预测误差 。

3.2 预期实验结果

- **长时程一致性 (Long-horizon Consistency):** 在预测未来 30s-60s 的任务中，Mem-JEPA 的特征漂移率（Drift Rate）应显著低于基线 V-JEPA 2。
- **计算效率 (Efficiency):** 在高频重复场景下，由于 Bypass 机制的存在，平均 FLOPs 应大幅下降，证明其在边缘设备上的潜力。
- **Video Scaling Law 验证:** 验证 Engram 论文中的 **U-shaped Law** 在视频领域同样适用——即在参数总量固定的情况下，分配一部分参数给非参数化记忆（Memory），比全部分配给计算（Compute）效果更好。


Video Input
V-JEPA Encoder
(Frozen)
Memory Path (Engram)
Quantizer
(Discrete Tokens)
Hash
Engram Table
(Key-Value Store)
z_mem
(Ground Truth)
Reasoning Path (V-JEPA)
Predictor
z_pred
(Drifting)
Fusion & Rectification
Residual Connection
Next State