MemoryTransit：一种基于可逆压缩与可学习预测的神经记忆存储模型

第一作者：WeinuoOu

第二作者：BaolinLiao

编辑：BaolinLiao和WeinuoOu

摘要

当前大型语言模型（LLM）等人工智能系统普遍缺乏有效的"运行时"记忆机制，难以适应动态与个性化的交互需求。为解决这一问题，本文提出 **MemoryTransit——一种新型神经记忆存储架构**，其核心融合了 **可逆维度压缩（Invertible Dimensionality Reduction）** 与 **可学习的辅助信息预测器（Learnable Auxiliary Predictor）**。通过可逆神经网络将输入数据无损映射至低维潜在空间，并分解为存储用的**压缩表示**（`z_comp`）与可丢弃的**辅助表示**（`z_aux`）。关键创新在于引入轻量级预测网络，从 `z_comp` 中预测 `z_aux`，进而在重建阶段通过逆变换高保真恢复原始数据。在此基础上，本工作构建了槽式全局记忆库（Memory Bank），并设计了基于余弦相似度的读取机制与基于访问频率的写入策略。实验表明，MemoryTransit 在保持与主成分分析（PCA）相近压缩效率的同时，展现出更强的非线性建模能力，为LLM等系统提供了一种灵活、高效且可学习的运行时记忆解决方案。

关键词：运行时记忆，可逆神经网络，流模型，维度压缩，记忆库，大型语言模型

---

1. 引言

人工智能尤其是大型语言模型（LLM）的快速发展，使得计算与存储资源成为制约其能力进一步扩展的关键瓶颈。当前主流LLM本质为"无状态"系统，其知识完全固化于训练所得参数中，即"训练时记忆"。这种范式在应对需长期上下文理解、个性化交互或动态知识更新的任务时表现局限。人类智能的核心特征之一是具备强大的"运行时记忆"，能够即时存储、检索并利用新信息。因此，为AI系统赋予类似的记忆能力，构建高效、可学习的外部记忆模块，成为极具价值的研究方向。

传统数据压缩方法如主成分分析（PCA）虽能有效降维，但其线性假设限制了其对复杂数据分布的建模能力，且通常为有损压缩，缺乏通过可学习机制优化重建过程的能力。

针对上述挑战，本文提出 **MemoryTransit** 模型，核心思想在于**将记忆的存储与重建解耦，并通过可学习预测器连接二者。**具体采用基于耦合层（Coupling Layer）的可逆神经网络作为编码器，确保变换的精确可逆性。输入数据经编码后切分为 `z_comp`（压缩存储部分）与 `z_aux`（辅助信息部分）。存储时仅保留 `z_comp`，并训练轻量级网络从 `z_comp` 预测 `z_aux`。重建时，将存储的 `z_comp` 与预测的 `z_aux` 拼接，经逆变换恢复原始数据。

进一步，本文构建全局记忆库，并设计基于余弦相似度的读取机制与基于访问频率的写入策略，实现动态记忆管理。本文主要贡献如下：

1. 提出MemoryTransit架构，融合可逆压缩与可学习预测，实现可优化的有损-重建记忆范式；
2. 设计完整的记忆读写机制，支持基于内容的检索与基于频率的更新；
3. 通过实验验证模型在非线性数据上的重建性能优于传统线性压缩方法。

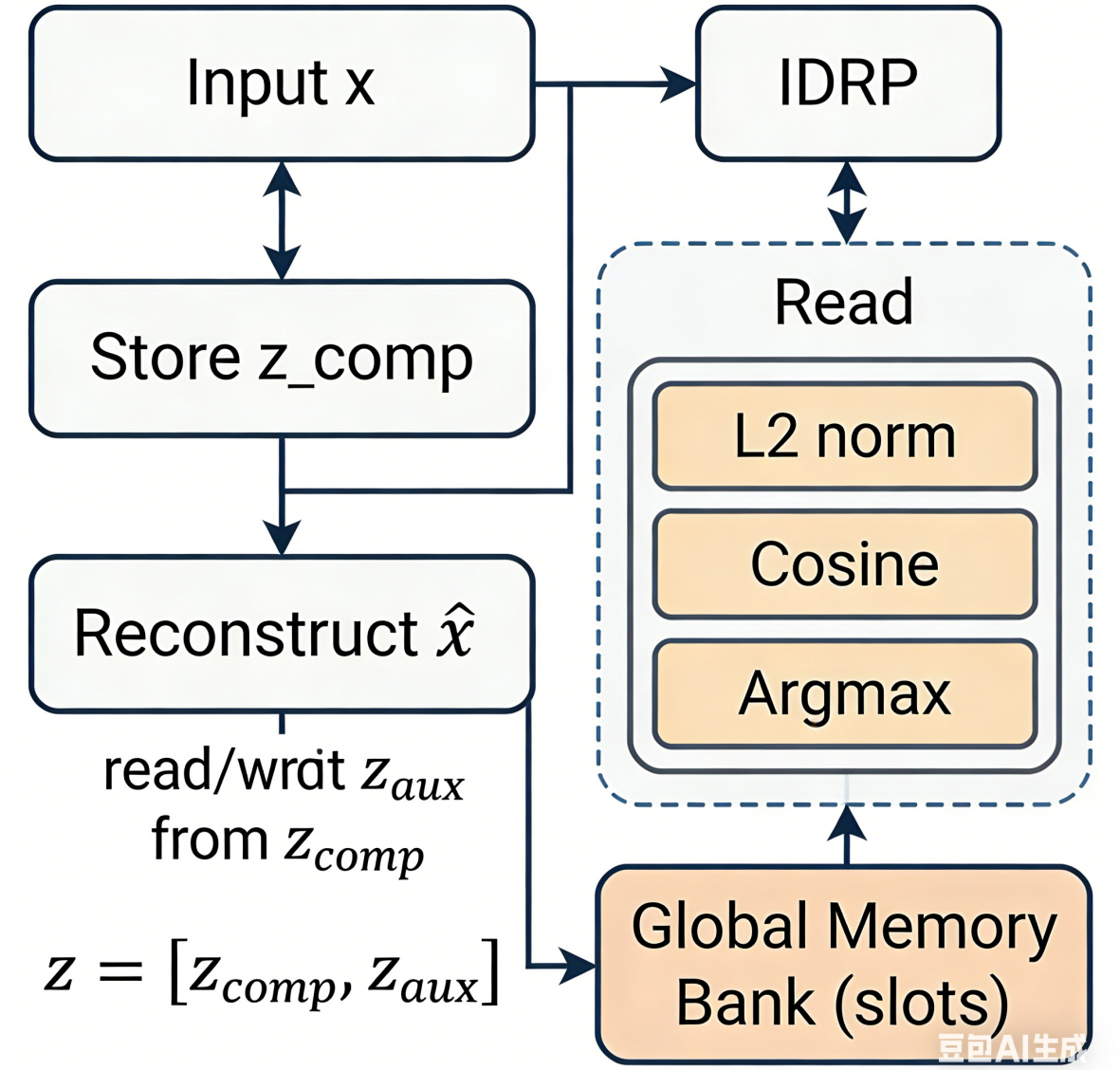
---

2. 模型架构

MemoryTransit 旨在构建可学习、高效、灵活的运行时记忆系统，其架构包含两大核心组件：**可逆维度压缩与预测器（IDRP）** 与 **记忆读写控制器**。

2.1 可逆维度压缩与预测器（IDRP）

IDRP 结合可逆变换与预测机制，实现高效压缩与高保真重建。



1. 可逆网络编码器
编码器由 N 层堆叠的**仿射耦合层**与**随机置换层**组成。

- **仿射耦合层**（`InvertibleCouplingLayer`）：输入 $\mathbf{x} \in \mathbb{R}^d$ 分为 $\mathbf{x}*1, \mathbf{x}*2 \in \mathbb{R}^{d/2}$。子网络 $\mathcal{N}$ 基于 $\mathbf{x}_1$ 生成缩放因子 $\mathbf{s}$ 与平移因子 $\mathbf{t}$：

  $$[\mathbf{s}, \mathbf{t}] = \mathcal{N}(\mathbf{x}_1)$$

  子网络由线性层、SwiGLU 激活函数与残差块构成。变换输出为：

  $$\mathbf{y}_1 = \mathbf{x}_1, \quad \mathbf{y}_2 = \mathbf{x}_2 \odot \exp(\mathbf{s}) + \mathbf{t}$$

  该变换具有精确可逆性，逆变换为 $\mathbf{x}*2 = (\mathbf{y}*2 - \mathbf{t}) \odot \exp(-\mathbf{s})$。
- **随机置换层**（`PermuteLayer`）：在耦合层后引入固定随机置换，确保维度充分交互。

经 N 层变换，输入 $\mathbf{x}$ 映射为潜在表示 $\mathbf{z} = f_{\theta}(\text{flatten}(\mathbf{x}))$，其中 f_θ 为可逆网络。

2. 潜在空间分解与预测
将 $\mathbf{z}$ 切分：

$$\mathbf{z} = [\mathbf{z}_{comp}, \mathbf{z}_{aux}]$$

$\mathbf{z}_{comp} \in \mathbb{R}^m$ *用于存储，*$\mathbf{z}_{aux}$ 被丢弃。引入辅助信息预测器 g_{ϕ} *(MLP结构：* `Linear → SwiGLU → Linear`*)，从* $\mathbf{z}_{comp}$ 预测 $\mathbf{z}_{aux}$：

$$\hat{\mathbf{z}}_{aux} = g_\phi(\mathbf{z}_{comp})$$

3. 工作流程

- **编码**：$\mathbf{x} \xrightarrow{f_{\theta}} \mathbf{z} \rightarrow (\mathbf{z}_{comp}, \mathbf{z}_{aux\_true})$
- **压缩**：输出 $\mathbf{z}_{comp}$
- **重建**：$\mathbf{z}_{comp} \xrightarrow{g_\phi} \hat{\mathbf{z}}_{aux} \rightarrow \hat{\mathbf{z}} = [\mathbf{z}_{comp}, \hat{\mathbf{z}}_{aux}] \xrightarrow{f_\theta^{-1}} \hat{\mathbf{x}}$

通过联合优化 f_{θ} *与* g_ϕ（如最小化重建损失），模型学习使 $\mathbf{z}_{comp}$ *蕴含足够信息以准确预测* $\mathbf{z}_{aux}$。

2.2 记忆读写控制器

基于IDRP构建全局记忆库 $\mathcal{M} \in \mathbb{R}^{\text{max\_mem} \times m}$，每行存储一个 $\mathbf{z}_{comp}$。

1. 读取机制

- 将查询 $\mathbf{x}$ 编码为 $\mathbf{q} = \mathbf{z}_{comp}$
- 计算 $\mathbf{q}$ 与 $\mathcal{M}$ 中各向量的余弦相似度：

$$\text{sim}_i = \frac{\mathbf{q}^\top \mathcal{M}_i}{\|\mathbf{q}\|_2 \|\mathcal{M}_i\|_2}$$

- 取最高相似度对应槽位 $\mathcal{M}_{i^*}$，*通过IDRP重建* $\hat{\mathbf{x}}_{\text{mem}}$ 并返回，更新访问频率。

2. 写入机制

- 对批次输入 $\{\mathbf{x}_j\}$ *编码得* $\{\mathbf{z}_{\text{comp}}^{(j)}\}$，计算均值 $\bar{\mathbf{z}}$ 作为写入向量
- 采用"先空闲，后最少使用"策略选择槽位：
    1. 优先选择未使用槽位（`AFF_ctrl[i] == 0`）
    2. 否则覆盖访问频率最低的槽位
- 将 $\bar{\mathbf{z}}$ 写入选定槽位，重置其访问计数

---

3. 运作机制与原理

MemoryTransit 模拟人脑记忆的编码、存储、检索与重建过程：

1. **编码与分离**：可逆网络作为非线性编码器，将输入重组为潜在表示 $\mathbf{z}$。通过切分 $\mathbf{z}$，迫使模型学习一种表示，使 $\mathbf{z}_{comp}$ *包含足以推断* $\mathbf{z}_{aux}$ 的关键信息。

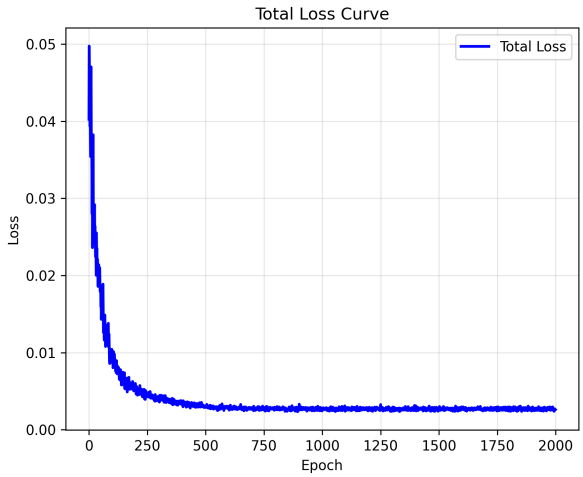
2. **有损存储与智能重建**：仅存储 $\mathbf{z}_{comp}$ 实现压缩。重建质量取决于预测器性能，通过端到端训练优化，实现超越传统线性方法的有损重建。
3. **记忆交互**：读取实现基于内容的关联检索；写入遵循"用进废退"原则，保持记忆库的动态更新与有效利用。

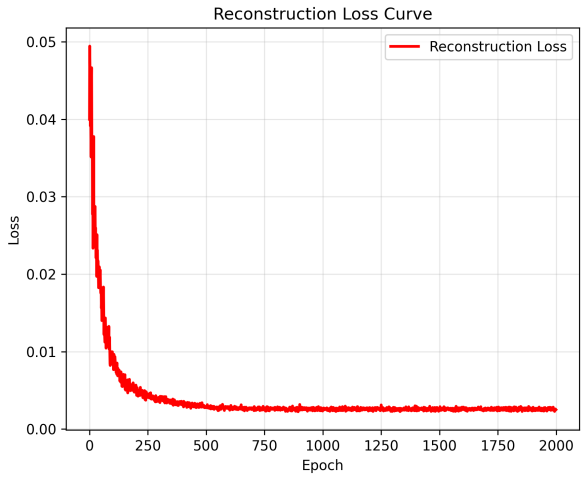
---

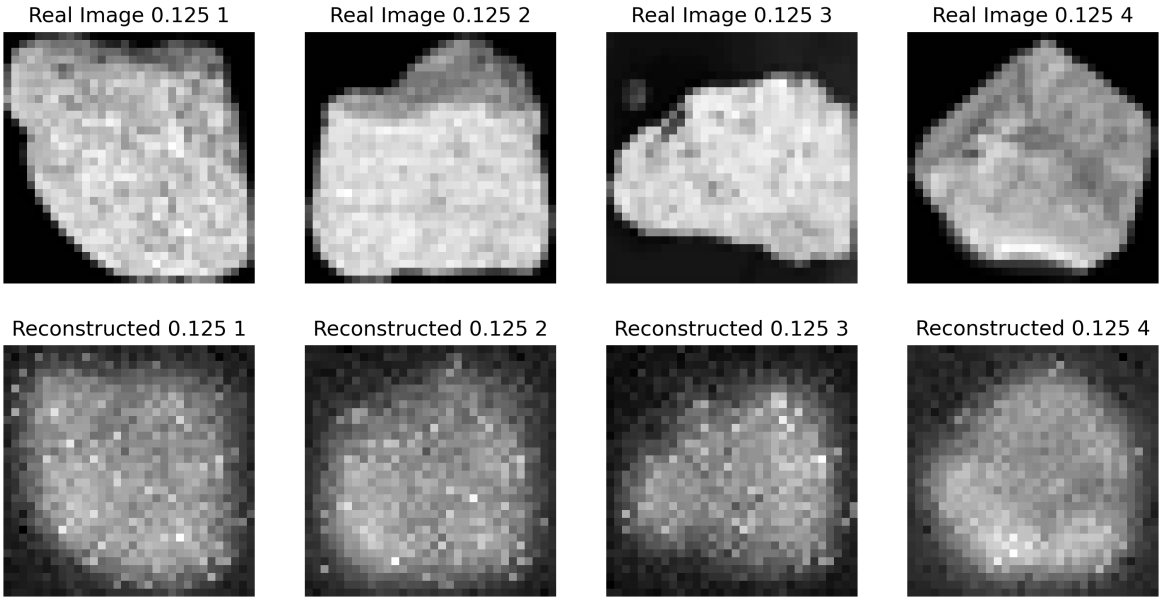
4. 实验与数据分析

4.1 合成数据训练（PCA为拟合）、真实数据测试

IDRP（预测网络预训练2000轮）：

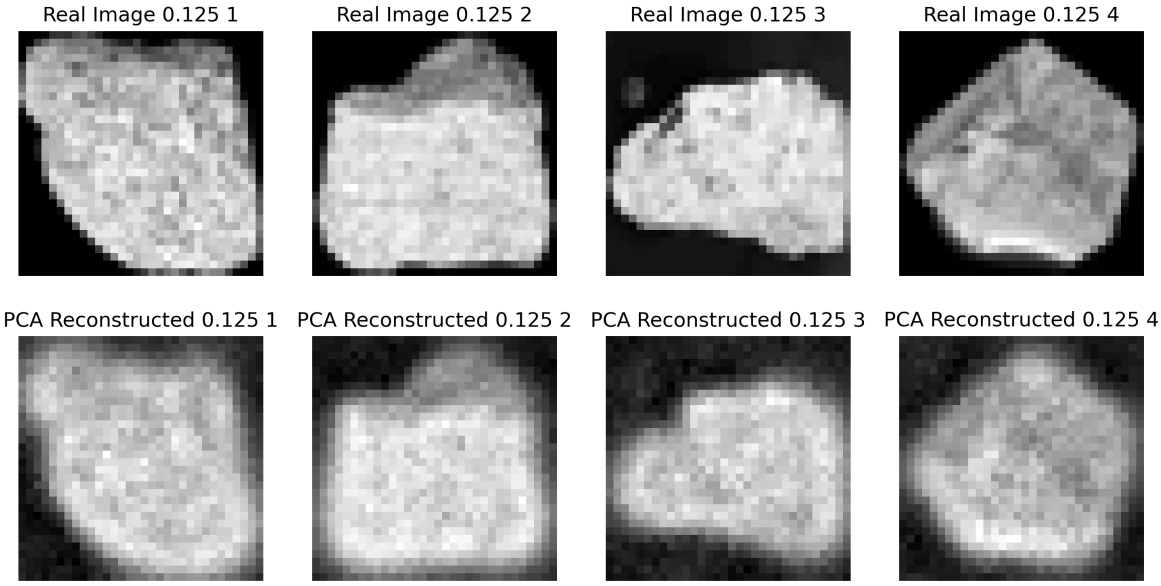







```
图像 1: PSNR = 15.40 dB, MAE = 0.128622, MSE = 0.028872
图像 2: PSNR = 14.94 dB, MAE = 0.135155, MSE = 0.032035
图像 3: PSNR = 13.54 dB, MAE = 0.154129, MSE = 0.044223
图像 4: PSNR = 17.87 dB, MAE = 0.100705, MSE = 0.016349
```

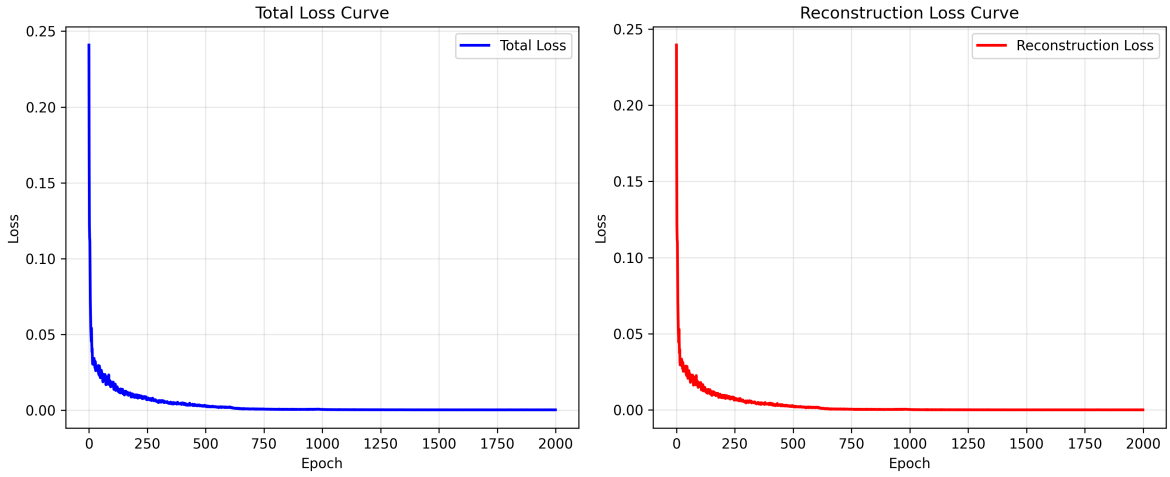
PCA：

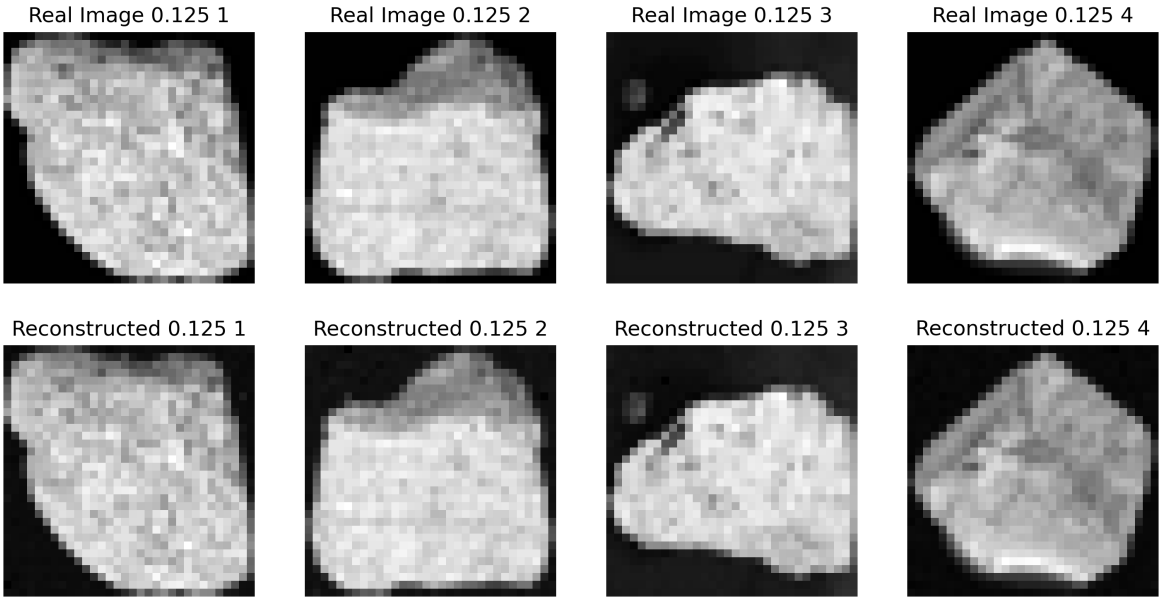



```
图像 1: PSNR = 27.68 dB, MAE = 0.032490, MSE = 0.001706
图像 2: PSNR = 27.22 dB, MAE = 0.034426, MSE = 0.001898
图像 3: PSNR = 29.60 dB, MAE = 0.026074, MSE = 0.001097
图像 4: PSNR = 27.90 dB, MAE = 0.031142, MSE = 0.001624
```

4.2 真实数据训练（PCA为拟合）、真实数据测试

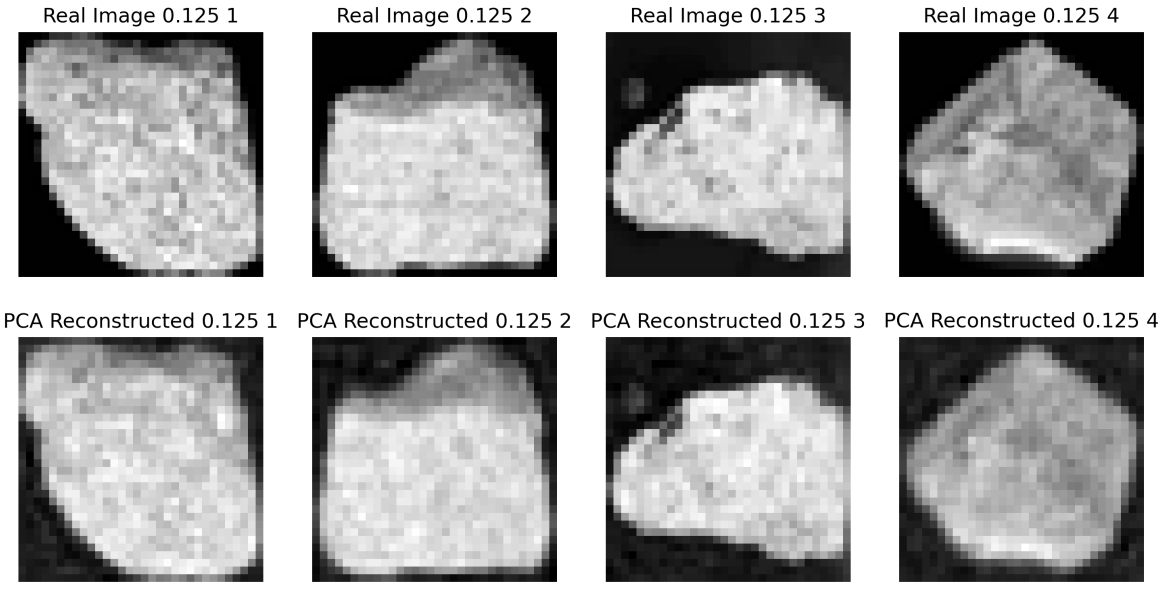
IDRP（预测网络预训练2000轮、6层、256隐藏维度）：






```
图像 1: PSNR = 40.60 dB, MAE = 0.004741, MSE = 0.000087
图像 2: PSNR = 39.04 dB, MAE = 0.005617, MSE = 0.000125
图像 3: PSNR = 44.34 dB, MAE = 0.002972, MSE = 0.000037
图像 4: PSNR = 42.05 dB, MAE = 0.004154, MSE = 0.000062
```

PCA：




```
图像 1: PSNR = 27.68 dB, MAE = 0.032490, MSE = 0.001706
图像 2: PSNR = 27.22 dB, MAE = 0.034426, MSE = 0.001898
图像 3: PSNR = 29.60 dB, MAE = 0.026074, MSE = 0.001097
图像 4: PSNR = 27.90 dB, MAE = 0.031142, MSE = 0.001624
```

实验设置与发现：

- **基线**：PCA
- **指标**：MSE、PSNR、MAE
- **关键结论**：
  - 在非线性数据上，MemoryTransit 重建误差显著低于PCA，验证其非线性建模优势；
  - 对训练数据拟合能力较强，表明预测网络具备良好的信息补全能力；

---

5. 重建噪声分析

重建过程中对 $\mathbf{z}_{aux}$ 的预测误差会传导至输出，形成**结构性噪声**，其特性与预测器能力边界相关。当输入偏离训练分布时，重建质量可能下降。未来可探索更鲁棒的预测架构或引入不确定性量化机制。实验表明，网络深度与层数对噪声拟合存在最优区间。

---

6. 结论与展望

本文提出 MemoryTransit，一种基于可逆压缩与可学习预测的神经记忆存储模型，旨在为AI系统构建高效、可学习的运行时记忆模块。该模型通过可逆网络实现无损编码，借助轻量预测器实现有损重建，并结合记忆库实现动态存储与检索。实验表明，MemoryTransit 在非线性数据上优于传统线性压缩方法，表现出更强的建模能力与稳定的重建性能。

未来工作可从以下方向展开：

1. **系统集成**：将MemoryTransit嵌入LLM推理流程，探索其在长上下文建模、个性化对话等任务中的应用；
2. **预测器优化**：设计更鲁棒、高效的辅助信息预测网络，提升重建质量与泛化能力；
3. **记忆控制学习**：研究由模型自主控制记忆写入的机制，实现更类人的记忆管理与知识贯通；
4. **扩展应用场景**：尝试在持续学习、多模态记忆等任务中验证与拓展本框架的实用性。

MemoryTransit 为运行时记忆建模提供了新思路，其可学习、可扩展的特性有望推动具备记忆能力的AI系统向更高层次发展。

---

7. 致谢

感谢广东石油化工学院廖宝林同学在论文撰写中的贡献，感谢五邑大学刘思东教授的悉心指导。

---

8. 参考文献

[1] Dinh, L., Sohl-Dickstein, J., & Bengio, S. (2017). Density estimation using Real NVP. *ICLR*.
[2] Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., & Lakshminarayanan, B. (2021). Normalizing Flows for Probabilistic Modeling and Inference. *JMLR*.
[3] Shazeer, N. (2020). GLU Variants Improve Transformer. *arXiv preprint arXiv:2002.05202*.
[4] Graves, A., Wayne, G., & Danihelka, I. (2014). Neural Turing Machines. *arXiv preprint arXiv:1410.5401*.