Reminisce：一种仿生人类记忆机制的多模态移动嵌入系统

Nature Communications,2025

Presenter：Yuankun Feng

2025.10.20

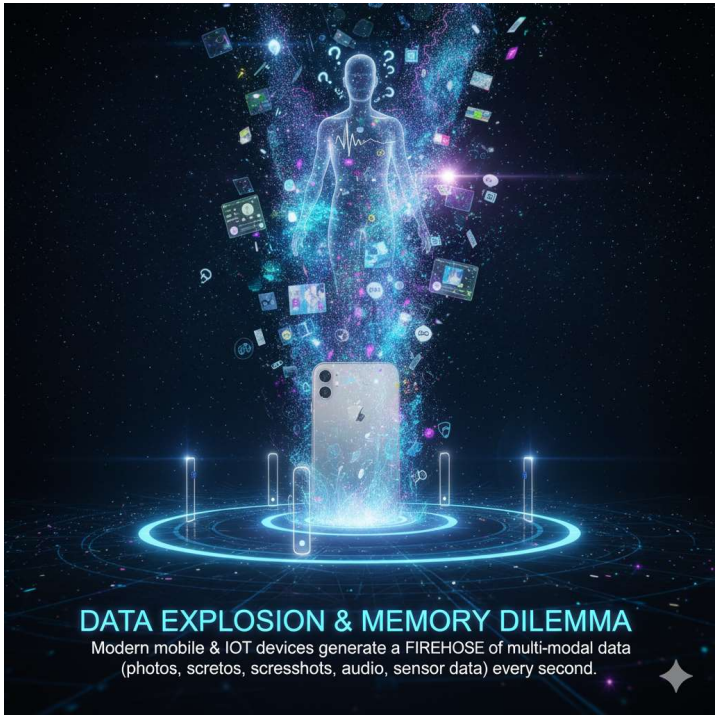
目录

目录

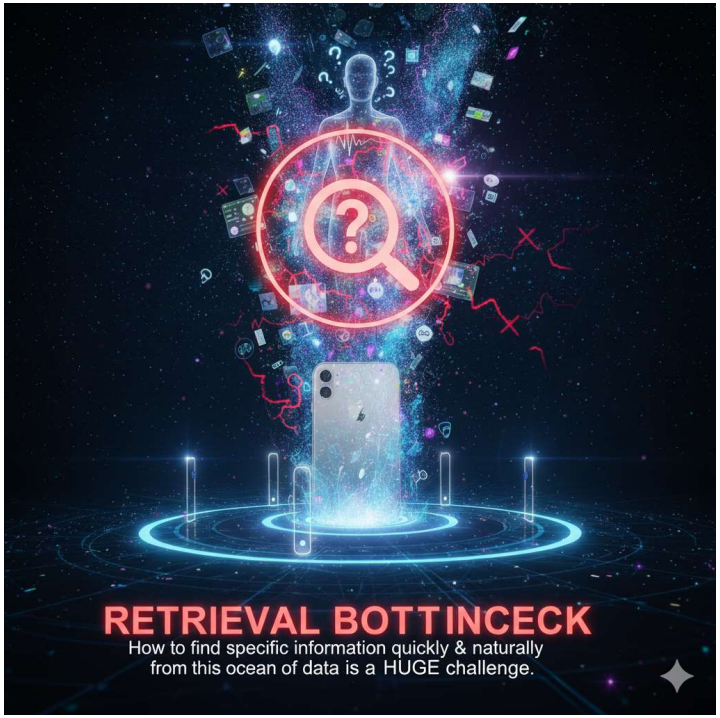
研究背景

问题场景

1. 移动设备无时无刻不在产生海量的多模态数据



2. 如何从这片数据的海洋中快速、准确、找到你想要的信息，是一个巨大的挑战



3. 本地数据放在云端进行处理，存在巨大的隐私泄露风险

研究背景

多模态嵌入模型（MEM）的崛起

➢ **技术基础**

多模态嵌入模型（如CLIP、ImageBind）能够将不同模态（文本、图像、音频等）的数据映射到同一个向量空间（Unified Embedding Space）。

➢ **应用潜力**

- 跨模态搜索
- 检索增强生成（RAG）

研究背景

现有挑战

1.庞大的MEM模型在移动设备上推理速度极慢

2.运行MEM所需的巨大算力会导致设备耗电剧增，续航下降

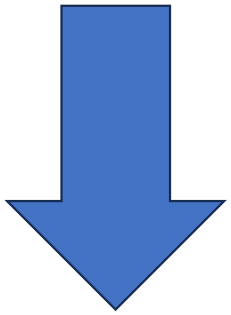
3.移动设备的CPU/GPU算力、内存容量均无法与云服务器相比

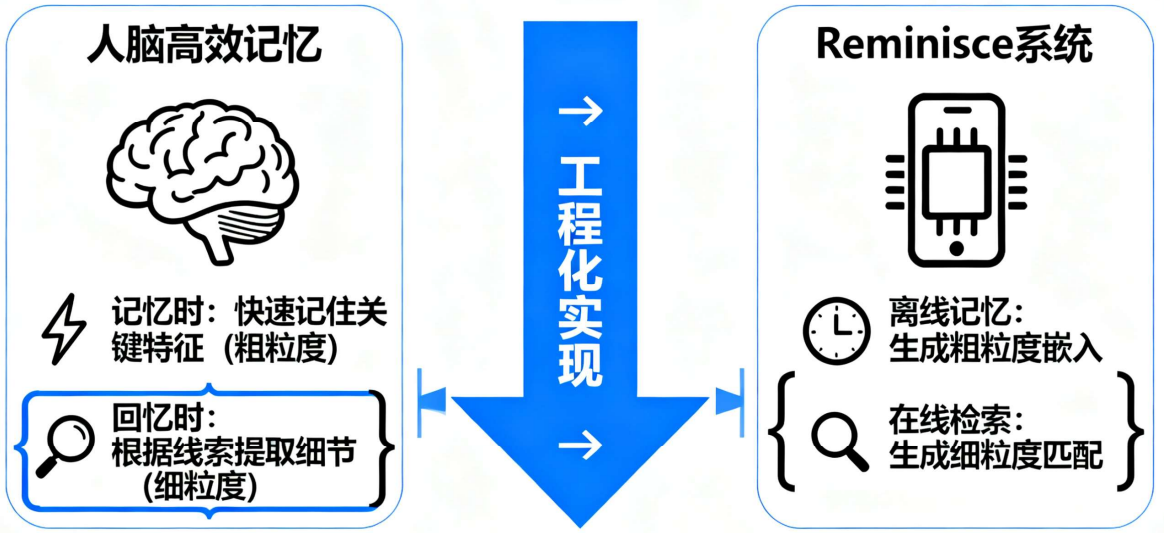
研究背景

灵感来源：模仿人脑的高效记忆机制

人脑启发：人脑不会事无巨细地记住所有细节

- **记忆时：先快速记住关键特征（粗粒度）**
- **回忆时：再根据线索提取细节（细粒度）**





将MEM高计算成本拆分为两阶段，适配端侧有限资源

将 MEM 的高计算成本 “拆分到离线记忆 + 在线检索”两阶段，适配端侧有限资源。

目录

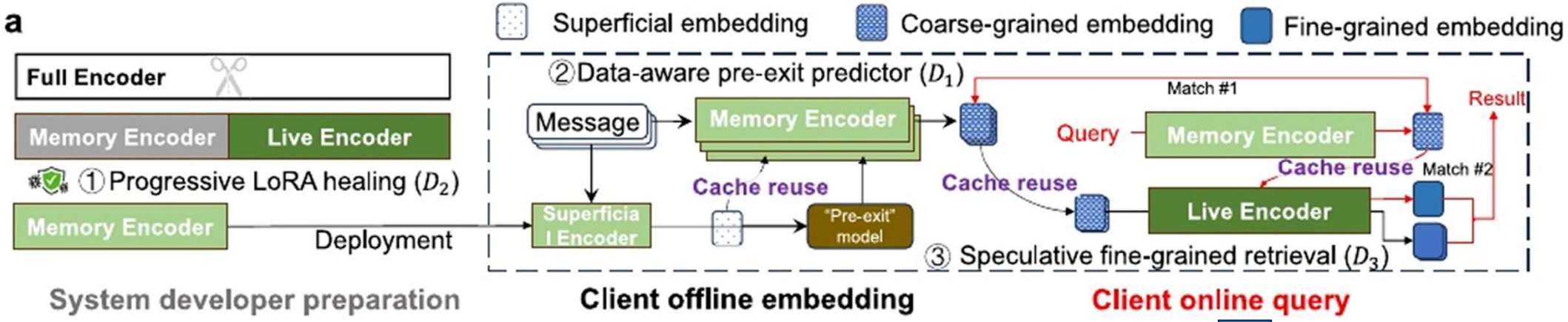
Reminisce

系统两个阶段

阶段	功能描述
离线嵌入	持续检测手机新生成数据，早退生成粗粒度嵌入（不跑全模型），本地存储。
在线查询	接收用户查询，先匹配粗粒度嵌入筛选候选，再用模型剩余层生成细粒度嵌入精准匹配。

Reminisce

Reminisce系统架构



1. Memory Encoder：用于离线生成粗粒度嵌入
2. Live Encoder：用于在线查询时生成细粒度嵌入

早退生成粗粒度嵌入

1.生成细粒度嵌入
2. 进行查询匹配

Reminisce

早期退出（Early Exit）

为什么需要早退?

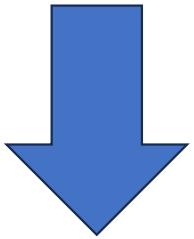
- **在移动端，多模态大模型推理成本高：**
  - **计算慢（延迟大）**
  - **耗电多（能耗高）**
  - **存储占用大（嵌入太长、太多）**
- **但 不是所有输入都需要完整模型：**
  - **简单样本在浅层就有足够语义信息**
  - **深层计算对它们的收益很小**

Reminisce

数据感知的预退出预测器（ Data-aware pre-exit predictor）

早期退出想法很好，但传统的实现方式在移动设备上效率极低。

传统做法（运行时决策）：每计算完一层，才判断"现在能退出吗?



✕ 批量处理失效：一个批次里的输入，退出时间各不相同，快的要等慢的，硬件并行能力被浪费。

✕ 内存碎片化：计算路径长短不一，内存分配混乱。

✕ 加载时间无法隐藏：无法预加载下一层参数。

我们能否在计算开始前，就预测出每个输入需要"算多深"?

Reminisce

数据感知的预退出预测器（ Data-aware pre-exit predictor）

- 输入：浅层嵌入（Superficial Embedding）

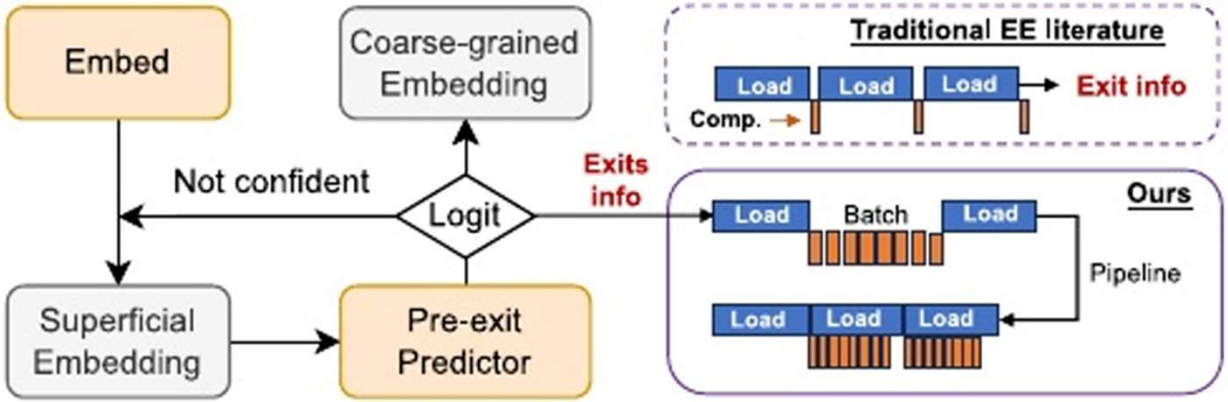
  来自模型前几层的中间激活，成本极低

- 预测：轻量级MLP

  用完整模型生成的细粒度嵌入作为基准，训练MLP预测最优退出点。

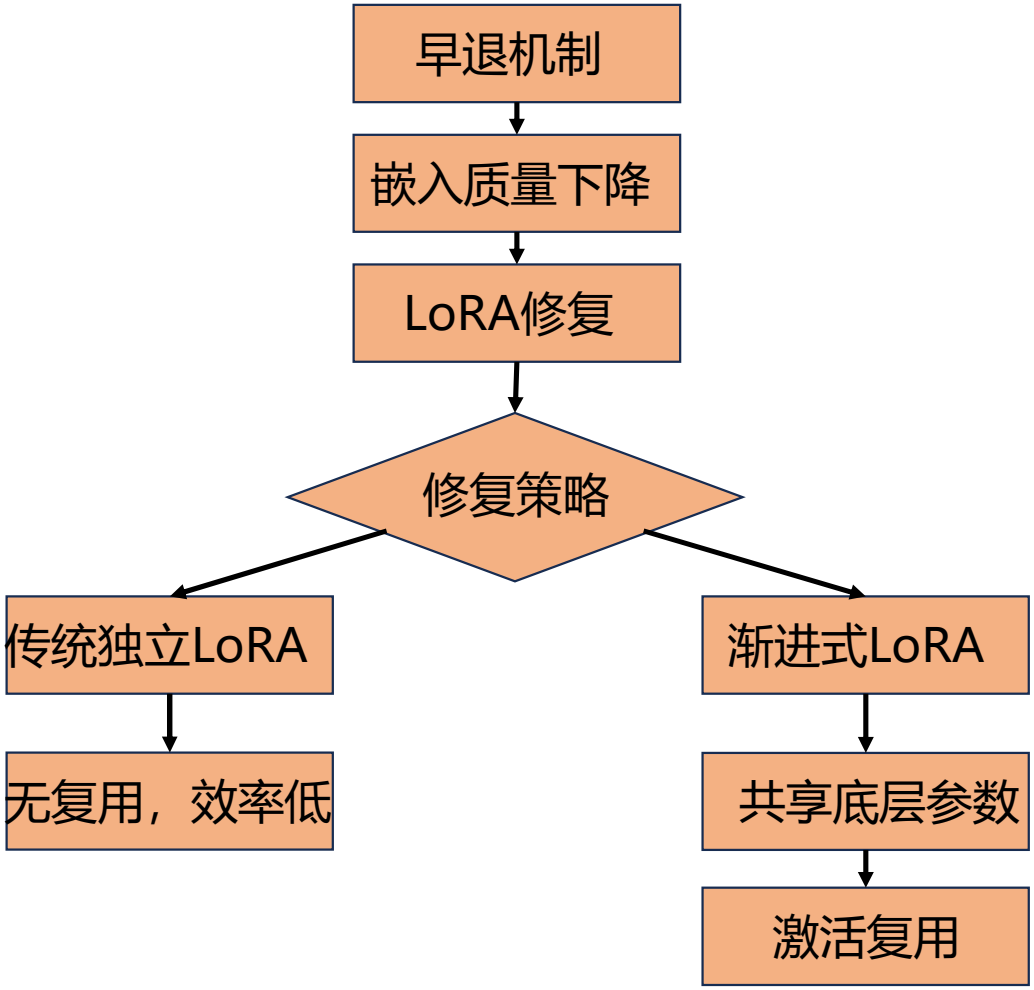
- 动态批处理

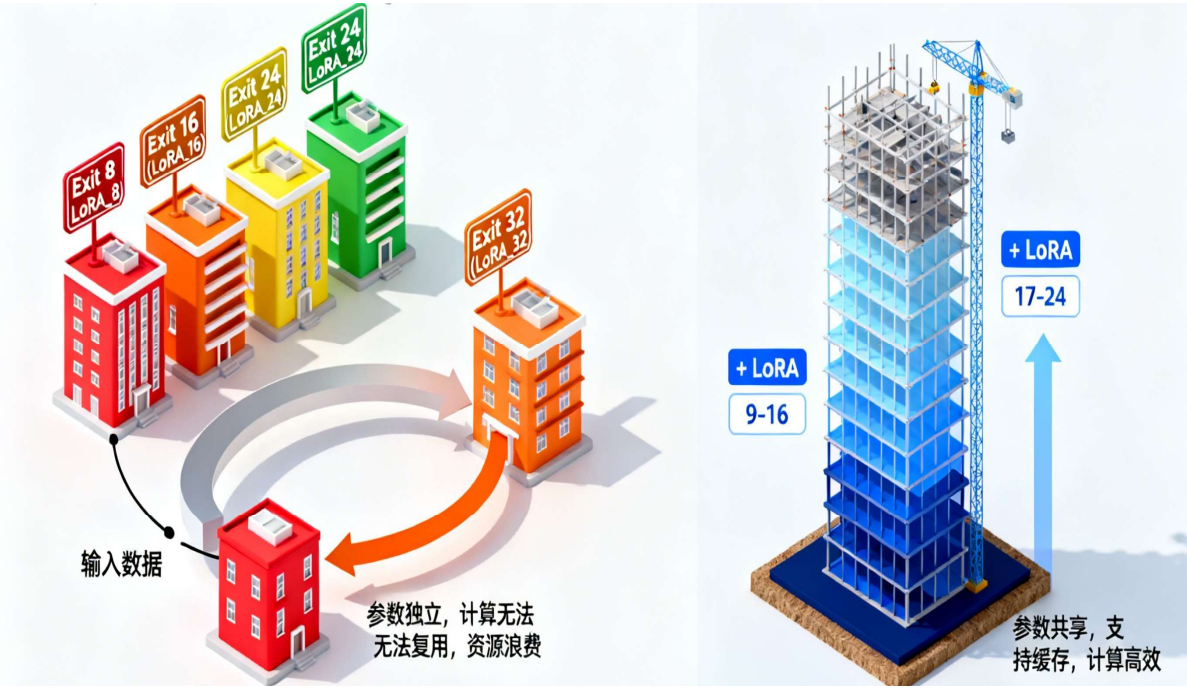
通过预判退出层，将相同退出层的样本批量处理，实现高效批处理和流水线执行



Reminisce

渐进式 LoRA 修复（Progressive LoRA healing）



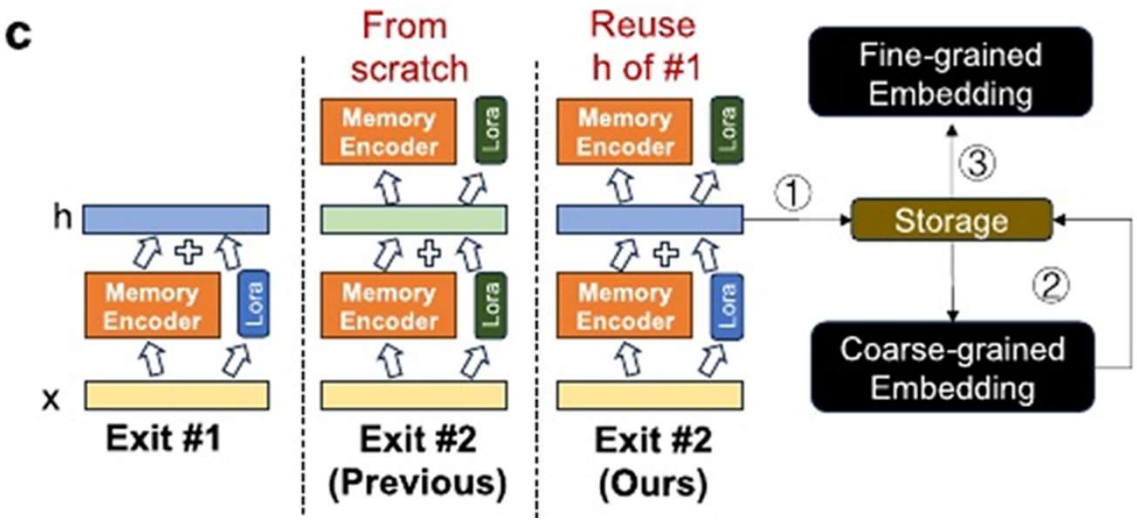


Reminisce

渐进式 LoRA 修复（Progressive LoRA healing）

- 什么是渐进式 LoRA 修复?

模块	计算路径
Exit #1	x→Memory Encoder+LoRA1→h1
传统 Exit #2	x→Memory Encoder+LoRA1 ' →Memory Encoder+LoRA2→h2
Ours，渐进式修复	h1→Memory Encoder+LoRA2→h2



Reminisce

推测性细粒度检索（Speculative fine-grained retrieval）

设计动机

背景：离线存储的是粗粒度嵌入，而用户查询需要高精度结果。

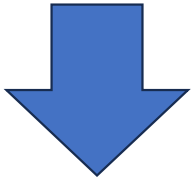
核心挑战：直接使用全模型生成的细粒度查询嵌入去匹配粗粒度嵌入库，会导致：

- **分布偏差**

  查询嵌入与粗粒度嵌入位于语义空间的不同区域，存在天然的“空间隔阂”。

- **对粗粒度嵌入的歧视**

  系统会系统性低估那些内容相关但表达粗糙的候选项。

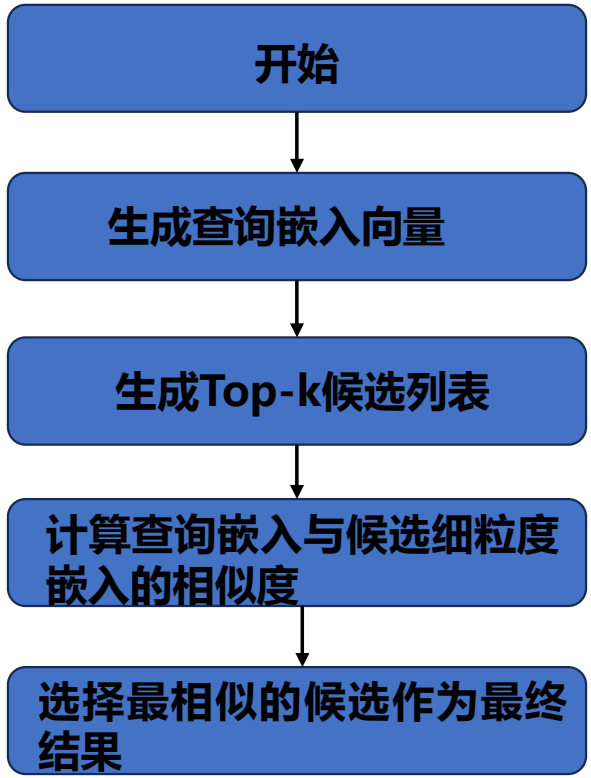


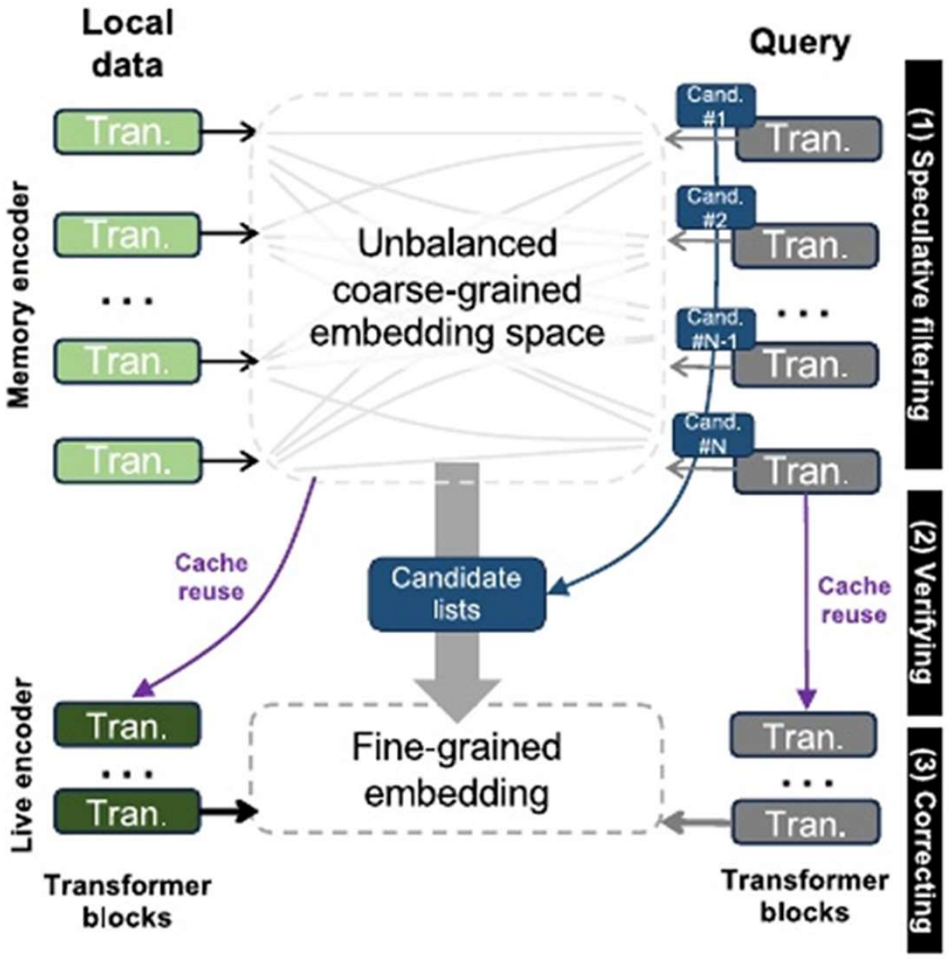
检索系统会漏掉大量相关的早期退出项，导致精度不达标。

Reminisce

推测性细粒度检索

■ 流程图

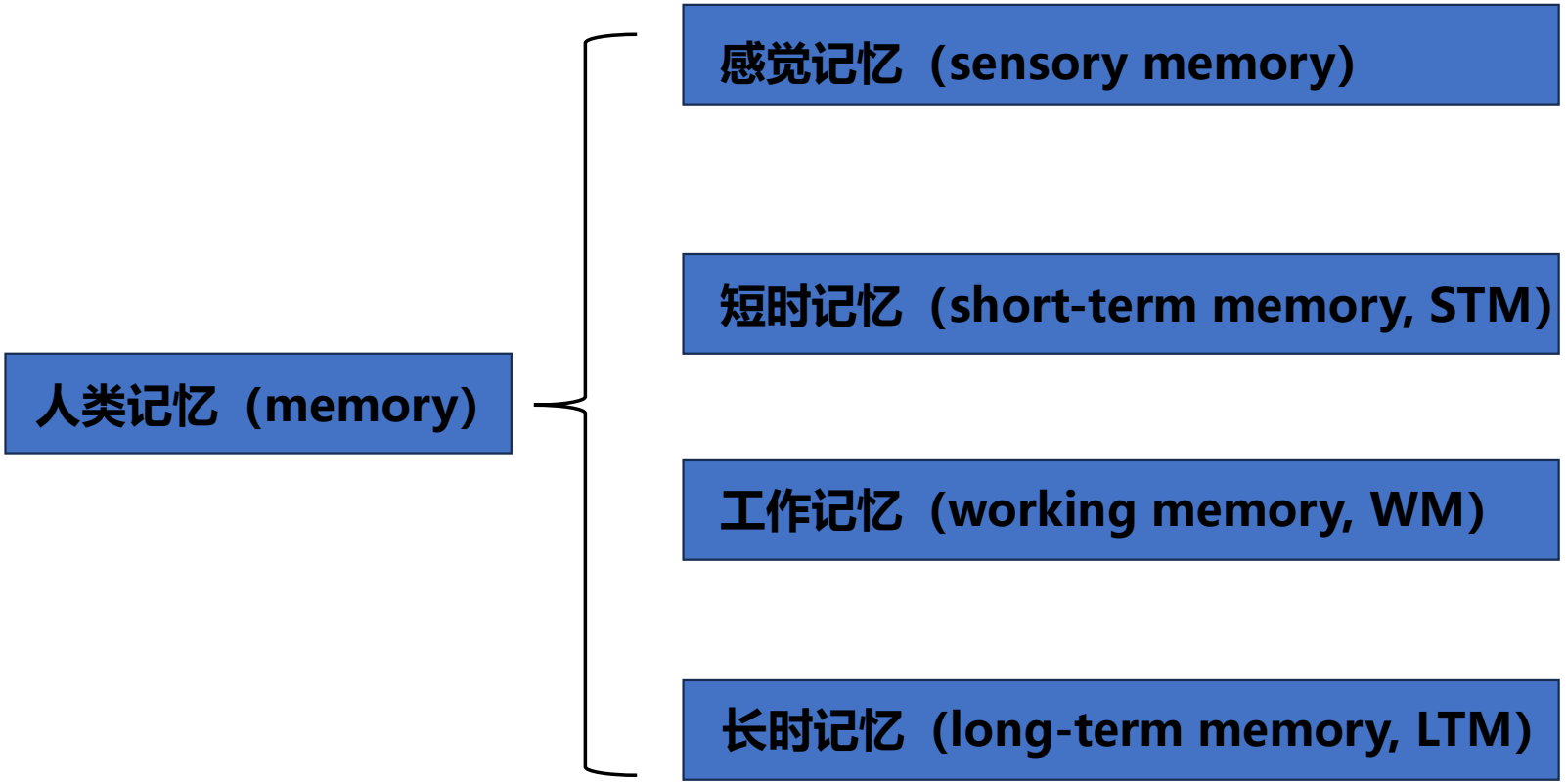




目录

记忆系统

Reminisce 的设计如此巧妙，是因为它在工程上复现了人类认知系统千万年来演化出的高效机制。

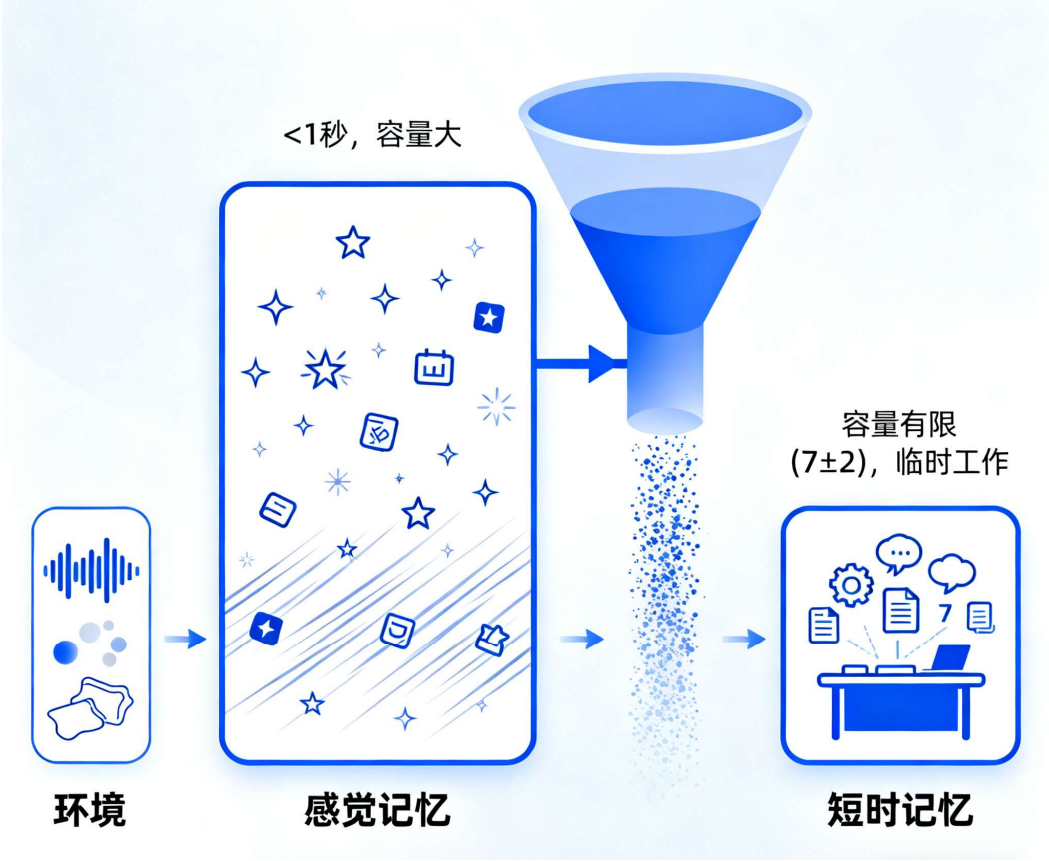


记忆系统

人类记忆的经典模型

- 感觉记忆（sensory memory）：感觉记忆是外界信息进入认知过程的初始存储系统，信息保存极短（视觉<1秒），容量大，但未加工 。

- 短时记忆 (Short-Term Memory)：容量有限（7±2个单元），是记忆系统中的临时工作台，用于暂时保存和处理当前正在使用的信息。



记忆系统

人类记忆的经典模型

- **工作记忆 (Working Memory)：现代认知心理学的核心。它不仅是存储仓库，更是一个主动的信息加工系统，包含存储和处理双重功能，是信息进入长时记忆的关键中转站 。**
- **长时记忆 (Long-Term Memory)：理论上容量无限，是知识、经验和技能的“永久档案馆”。**

记忆系统

工作记忆系统

■ 中枢执行系统（central executive）

负责协调和控制其他三个子系统的操作，类似于CPU，管理资源分配和任务执行。

■ 语音回路（phonological loop）

存储 + 加工 言语信息（如默念单词、记电话号码）

■ 视觉空间模板（visuospatial sketch pad）

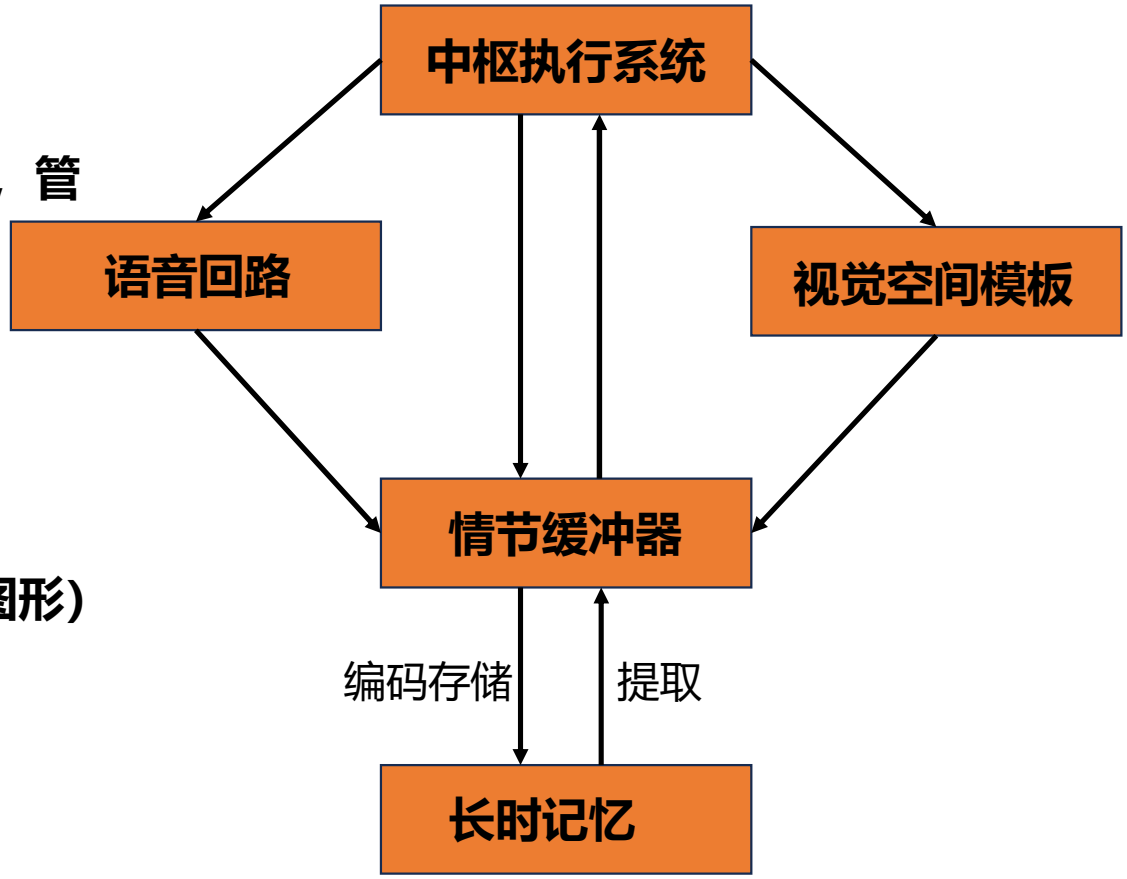
存储 + 加工 视觉 / 空间信息（如记地图方位、想象图形）

■ 情节缓冲器（episodic buffer）

① 将语音、视觉空间整合

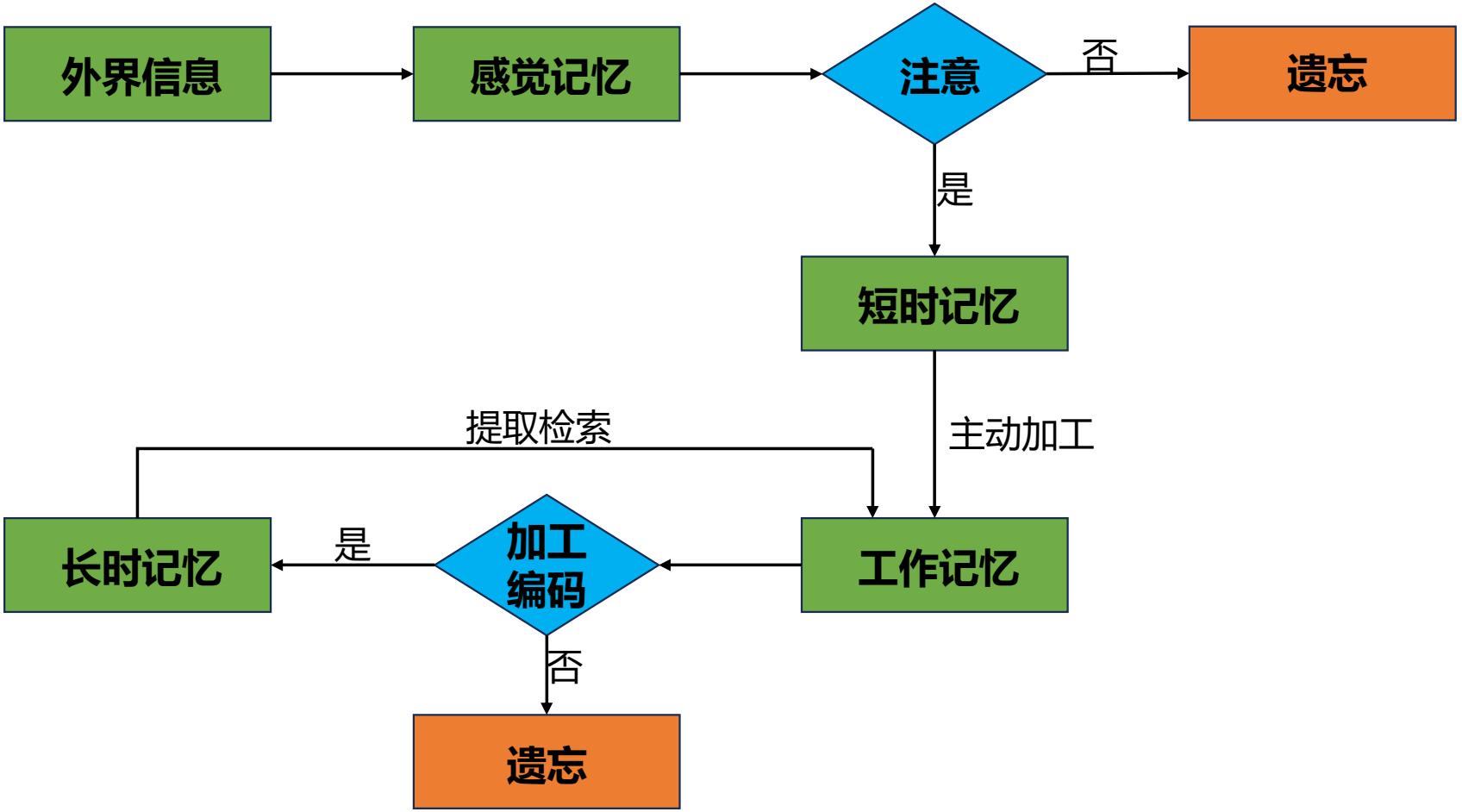
② 作为工作记忆与长时记忆的桥梁

③ 临时储存多模态信息



记忆系统

记忆系统结构



记忆系统

长时记忆编码与提取

- 编码（encoding）

语义关联越强，记忆越持久，如“鸟在天空飞” 比 “鸟在盒子里” 回忆率高 3 倍，因前者更符合语义逻辑。

- 提取（retrieval）

- 情境依存性线索：提取时的环境、场景与编码时越一致，记忆越清晰
- 状态依赖记忆（state-dependent memory）：生理或情绪状态一致有助于提取
- 组织与结构化：内容越组织化、图式化，越容易提取，如分类后的东西更容易记忆

遗忘（forgetting）

■ 遗忘的本质

① 信息仍存在在长时记忆中，只是你当下无法检索出来

② 遗忘本质上是一种“检索失败”，而非信息真正消失

■ 遗忘的核心理论

➢ 干扰（interference）

① 观点：一些信息可以“替换”其他信息，从而使得前面的信息难以获得提取。

② 类比：书架上，新放的书挡住旧书（新信息干扰旧信息），或旧书挡住新书（旧信息干扰新信息）

➢ 衰退（decay）

① 观点：记忆痕迹随时间自然消退。

② 类比：记住一串电话号码，不加以复述就会遗忘。

遗忘

- 提取失败（retrieval failure)

① 观点：信息仍在记忆中，但缺乏有效提取线索。

② 现象：

- 舌尖现象（tip-of-the-tongue phenomenon）：知道明星名字，但说不出，给个提示（如 “演《甄嬛传》”）就想起来
- 编码特异性（encoding specificity）：在学习的教室考试（线索匹配），比在操场考得分高

- 动机性遗忘（motivated forgetting)
  - 观点：心理防御机制主动抑制痛苦记忆（如创伤事件）
  - 类比：把尴尬的照片锁进箱子底，假装没看见（潜意识主动压抑）

遗忘

■ 影响遗忘的因素

① 时间：艾宾浩斯曲线（Ebbinghaus curve）：遗忘先快后慢，越久越难提取。

② 加工深度：死记硬背（浅加工）比理解（深加工）更易忘。

③ 相似干扰：同时学多个相似内容（如同时学法语和西班牙语），词汇容易打架。

④ 负面情绪：焦虑、悲伤会 “冻结” 记忆提取（举例：考试紧张忘知识点）。

⑤ 脑损伤 / 疾病：阿尔茨海默病破坏海马体（情景记忆）和颞叶（语义记忆），导致不可逆遗忘。

■ 遗忘的意义

① 节省认知资源：大脑不会储存所有无用信息（就像缓存机制）

② 增强适应性：通过遗忘不必要或痛苦记忆，专注于重要信息

目录

总结与思考

总结

Reminisce 通过一种受大脑启发的协同计算架构，结合了离线粗粒度嵌入的高效率与在线细粒度精炼的高精度，提升了多模态嵌入模型在移动设备上的处理吞吐量与能效，满足了个人记忆增强应用对实时响应和长期运行的苛刻要求。

思考

缺乏对长期记忆演化的支持：

Reminisce 默认永久保存所有粗粒度嵌入，但未考虑记忆老化或语义冗余问题。随着使用时间增长，嵌入库会持续膨胀，不仅增加存储和检索开销，还可能因大量相似记忆（如重复截图、相似界面）干扰检索精度。未来可引入基于语义聚类的自动压缩或基于访问频率的遗忘机制，模拟人类“用进废退”的记忆特性。


Thank you!