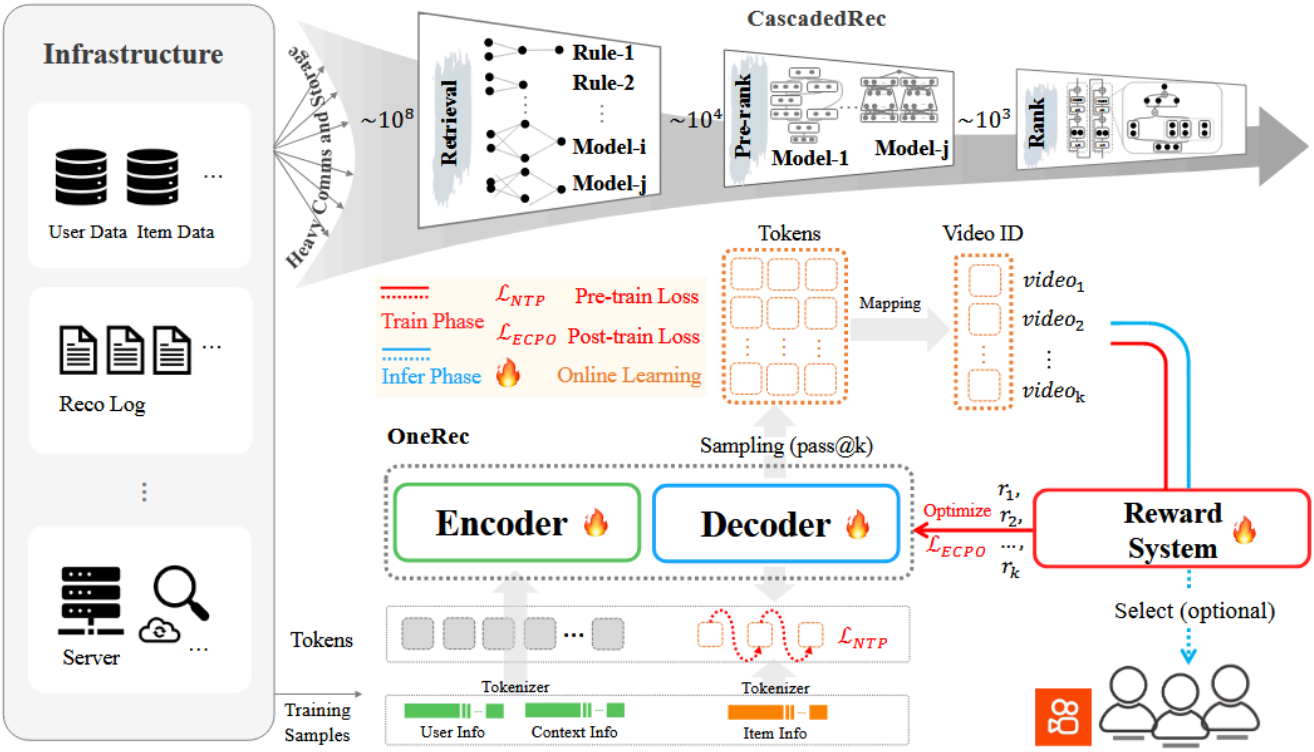
介绍

开篇提出，目前的大部分推荐系统还是按照多级级联结构（multi-stage cascaded architecture）比如召回-精排-重排）的方式而不是端到端的方法 -> 通信存储计算的碎片化,，每个阶段的优化目标不一致，传统推荐系统不能跟上ai演化的脚步
OneRec 的关键就在于：把检索和排序等所有阶段融合进同一个生成式模型，让系统能够一次性的学习并生成推荐结果。

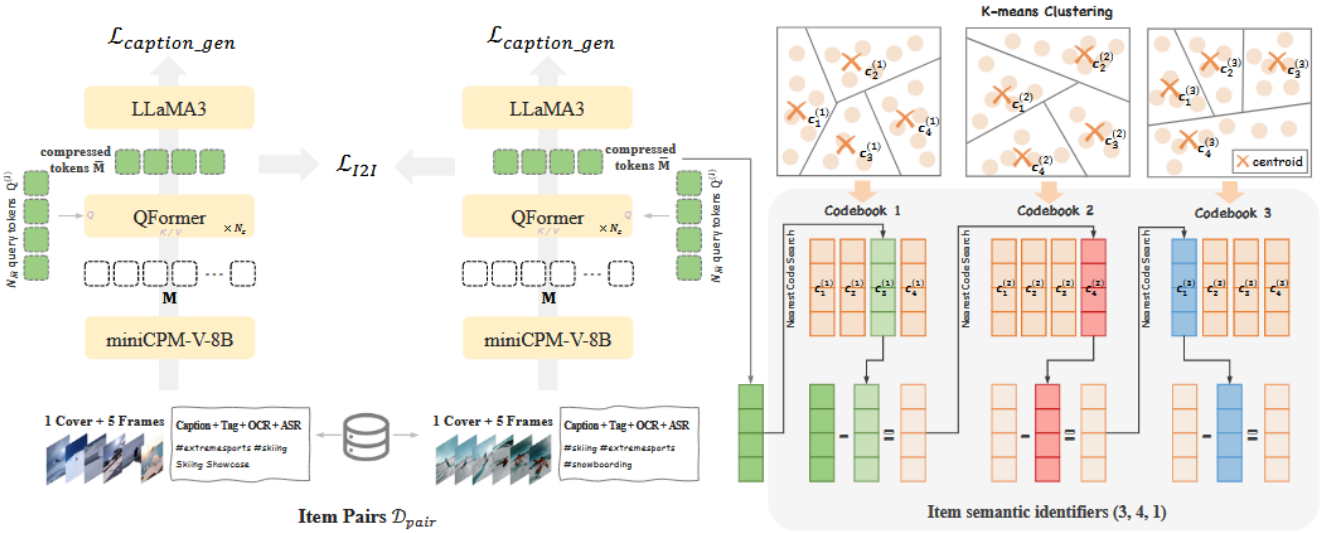
整体架构图（级联与OneRec的对比）

OneRec采用编码器-解码器架构，借助奖励模型，以端到端的方式生成用户喜欢的视频。



架构

分词器tokenizer

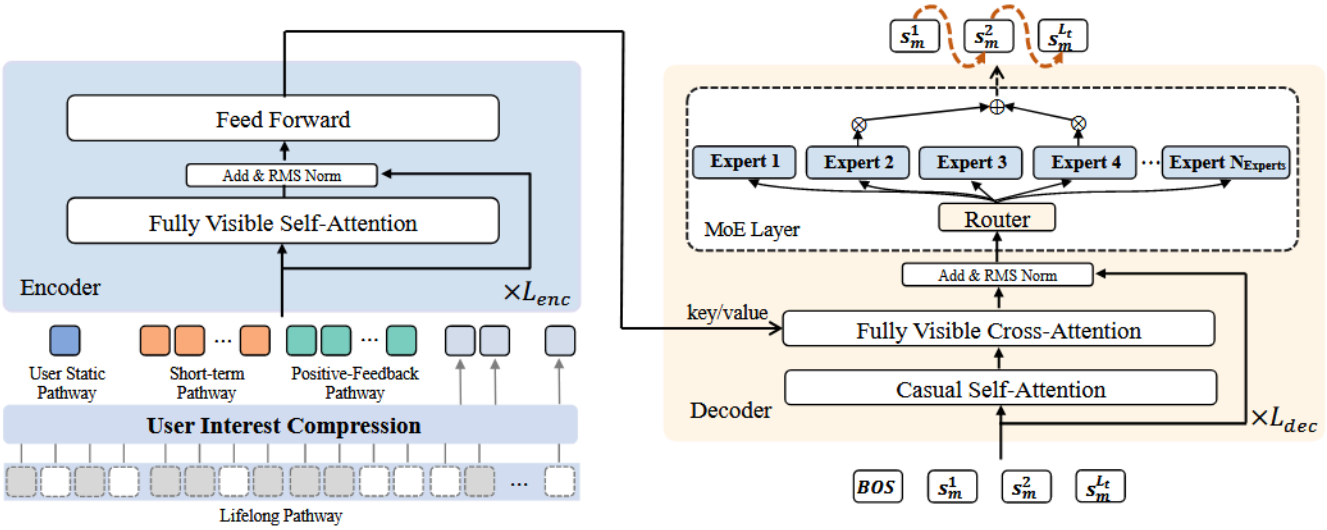


为什么不直接生成视频ID, 而是把视频token化?
ID训练参数量过大，token化最后模型生成的不是“视频具体 ID”，而是“视频语义编码”，再映射回具体视频。

传统方法只用内容（标题、标签等）生成语义 Token，忽略协同信号（用户行为）。
OneRec 改进为：多模态内容（文本+图像+语音） + 协同信号（用户点击关系） → 融合后再分词 Tokenize
协同信号 = 看过 A 的人也看 B

分词使用RQ-KMeans（残差量化聚类）来生成分层语义 ID
关于RQ-kMeans:
第 1 层：粗聚类：把视频大致归类（如“运动”、“美食”、“游戏”）
第 2 层：在对应类别内部再聚类（如“滑雪”、“篮球”、“瑜伽”）
第 3 层：再细化（如“滑雪教学”、“滑雪比赛”、“搞笑滑雪事故”）
于是一个视频可能变成这样的 Token：`(3, 4, 1)` 分别对应 3 层的语义编码。

编码-解码器架构



编码器：刻画用户特征

编码器使用四个角度融合用户特征：
第一个是user static pathway, “用户静态特征”部分，就像档案卡，记录年龄、性别、地区等基本信息。
第二个是 short-term pathway, “短期行为”部分，抓取最近 20 条观看记录，不仅看视频编号，还关注作者、标签、时间、停留时长和点赞、评论等互动情况。
第三个是positive-feedback pathway, “正反馈”部分，会集中处理那些用户反应特别强烈的 256 条，比如反复观看、点赞、转发这些高参与度的内容。
第四个lifelong pathway “终身历史”路径，将用户的长期兴趣进行压缩，但直接使用注意力机制的计算量太大，因此先用一种分层的 K-means 聚类，把相似的视频归到同一簇里，再从每个簇里选出最能代表那簇的item，把整个历史压缩成大约 2,000 条“精华”信息。接着再用轻量版的 QFormer（带 128 条可学查询向量）把这 2,000 条内容进一步浓缩，既抓住了长期兴趣的核心，又保证了计算可行。

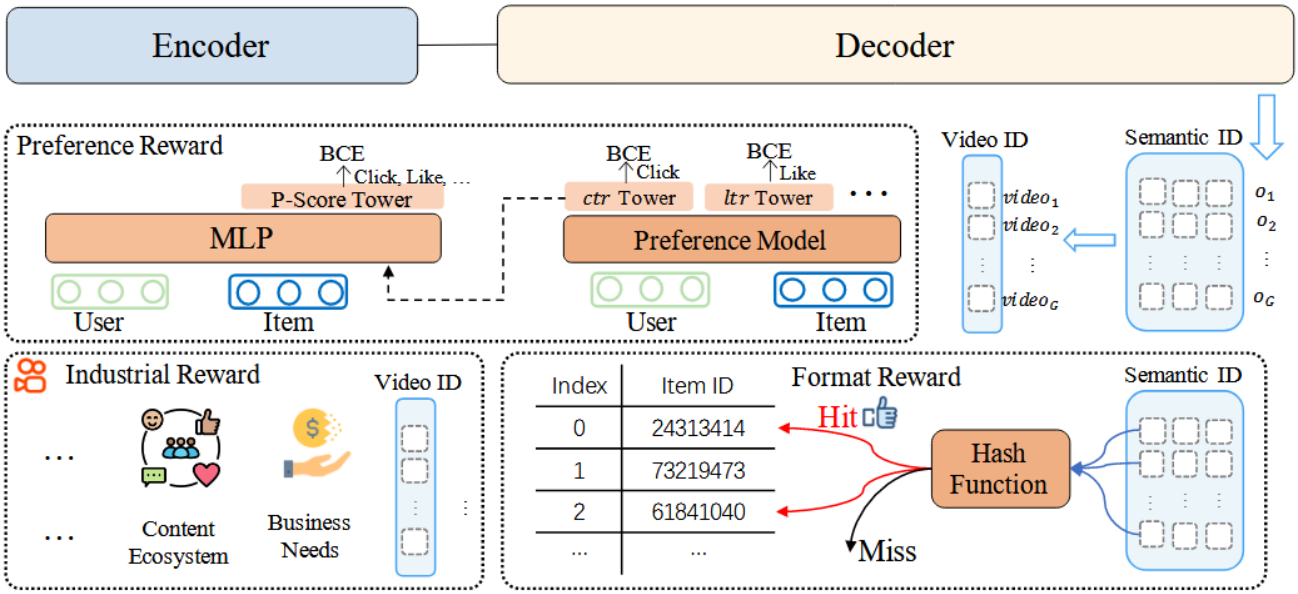
解码器：生成推荐

解码器从一个可学习的起始符开始，逐点写出下一位视频编号（point-wise generation）
预测过程解码器的每一层都做三件事：第一，用“因果注意力”让已经生成的编号彼此参考，好比先前写过的词会影响接下来要写的词；第二，用“交叉注意力”把用户兴趣的整体画像也拉进来，比方说把用户的历史偏好当作上下文，帮助模型判断下一步最合适的编号；第三，再把这条信息送进一个“混合专家”网络（MoE），这里面藏着很多小专家，模型会根据当前输入挑出排名前几的专家来帮忙，既能扩充模型容量，又不会让某个专家过载，保证所有专家都能被合理利用。

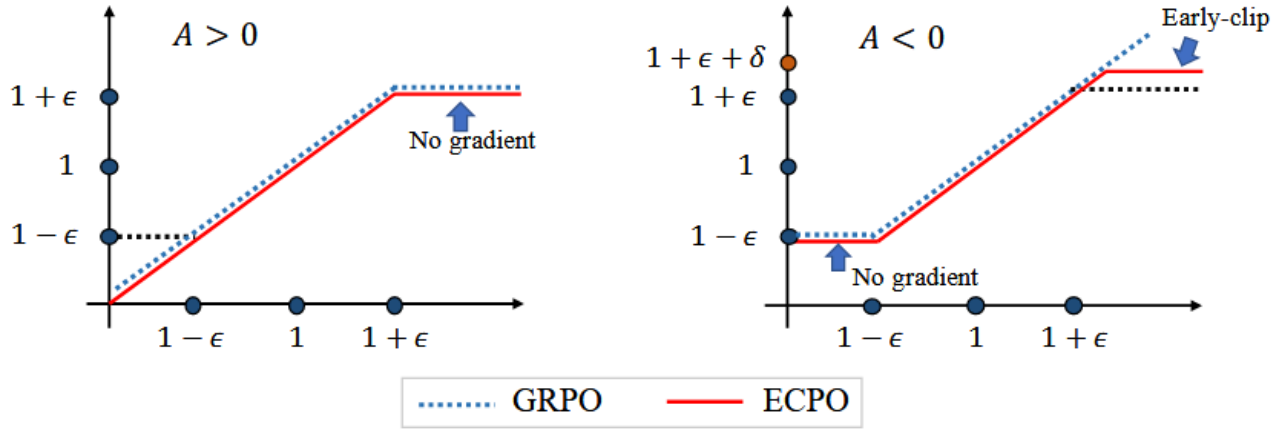
强化学习

为了让模型更加懂得客户，OneRec使用 on-policy（基于模型当前生成结果）强化学习在生成的物品空间中对模型进行训练。通过奖励信号，模型能够感知到更加细粒度的用户偏好信息。
OneRec使用的三类奖励：

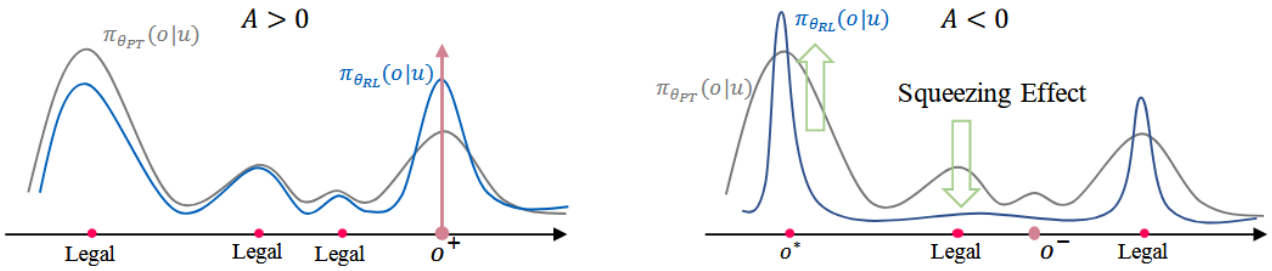
- **偏好奖励（Preference Reward）**：用于对齐用户偏好，把点击、点赞、观看时长等多种反馈融合成一个“P-Score”



然后用一种叫 ECPO（Early Clipped GRPO，早期裁剪GRPO）的算法，沿着这个分数不断优化模型，让推荐更“对胃口”。

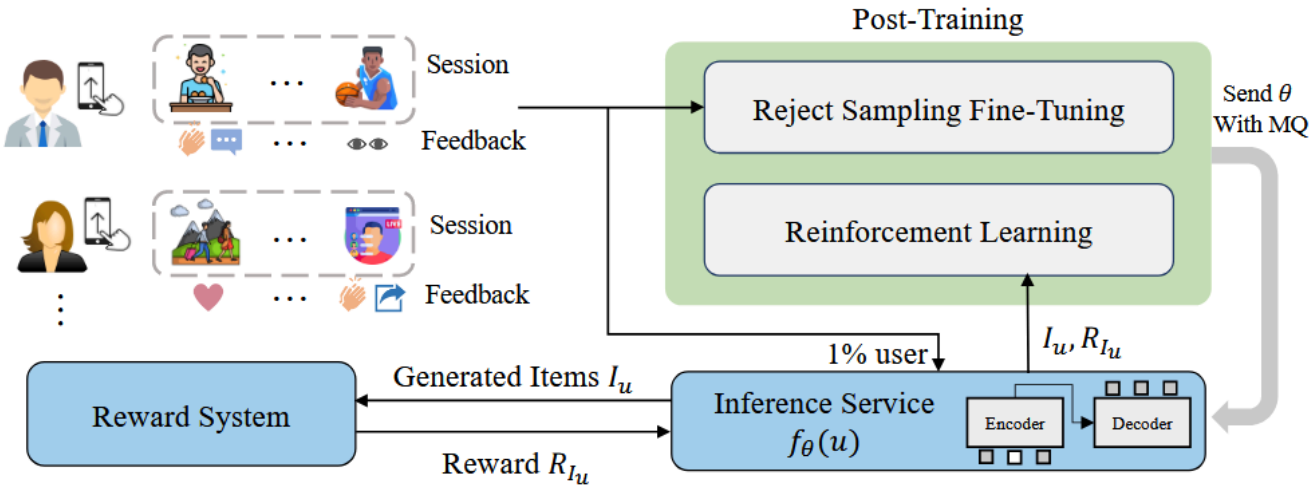


- **格式奖励（Format Reward）**：从一批生成结果中随机抽取若干条，把能对应到实际视频的标记当作奖励，非法或找不到视频的直接丢掉。这样，模型就学会只“写”那些能映射到真实内容的序列，不会再出现“空洞编号”。



- **行业特定奖励（Industrial-specific Reward）**：适配特定业务场景下的特殊需求。举个例子，如果平台想适当压缩低质内容或提高新作者曝光，就可以在奖励函数里给这部分内容打上不同的加减分，让模型在统一的学习过程中自然兼顾这些商业或生态指标。

训练框架



预训练

关于设备：OneRec 背后用了 90 台服务器，每台配备 8 块旗舰 GPU
平台每天处理大约 180 亿条训练样本，产生 540 亿个语义 ID 标记。OneRec 的几种模型从几千万到 26 亿参数不等，一般跑到 1,000 亿标记左右就能收敛。

后训练

使用拒绝采样微调（RSFT）和强化学习（RL）
对于 RSFT，我们根据播放时长过滤掉曝光会话中排名后 50% 的样本。
对于 RL，我们从 RSFT 数据中随机选取 1% 的用户，用于生成 RL 训练样本。（RL 样本的生成环节被拆到外部推理服务去跑，训练程序则定期把最新参数推送给它，形成紧密的闭环优化。）

其他知识

MFU
Model FLOPs Utilization ，模型算力利用率
表示模型在理论最大算力的基础上，实际利用了多少 GPU 运算能力。
一般推理阶段的 MFU 比 训练阶段高一些

FLOPs
每秒浮点运算次数（Floating-Point Operations Per Second）

QPS
Queries Per Second，每秒请求数 / 每秒查询数
在推荐系统中，QPS表示系统每秒能够处理的推荐请求数量，是衡量服务 吞吐能力与并发性能的关键指标。

OPEX
Operational Expenditure ，运营成本 / 运维支出