亚马逊云科技

中国峰会

2026年6月23日 - 24日 · 上海 · 世博中心

亚马逊云科技
中国峰会

基于 SageMaker 的多模态广告推荐特征工程

杨 涛

Mobvista 广告算法技术负责人

冉晨伟

亚马逊云科技 应用科学家

客户业务总览 —— 汇量科技

客户

新加坡
10%

亚太地区（除新加坡）
46%

其他地区
44%

130+ 国家和地区

游戏
75%

非游戏
25%

效果收费[1]
CPI/CPL/CPE...

平台费[2]

汇量科技产品矩阵

第一阶段：广告联盟

- 聚合海外多品类非程序化及非标准化流量
- 日均触及 10 亿+活跃设备

第二阶段：程序化广告交易平台

- 日均处理 3000 亿+程序化广告请求，反馈延迟小于 50 毫秒
- All-in-Cloud 架构，自研系统实时调度100万核 CPU
- 日均触及 30 亿+活跃设备

↑ 持续优化效果

第三阶段： SaaS 工具包

统计与分析 | 素材管理 | 广告投放服务 | 算力优化

高效触及

全球用户

250+
国家和地区

35亿+
全球活跃设备

附注：汇量科技2025 年报数据

附注：
- 效果收费，为广告按每次安装/线索/行为等效果收费的定价模型
- 平台费，为平台按订阅时间/账户数量等收费的定价模型

*前述特定亚马逊云科技生成式人工智能相关的服务目前在亚马逊云科技海外区域可用。亚马逊云科技中国区域相关云服务由西云数据和光环新网运营，具体信息以中国区域官网为准。

AI 驱动的三方广告生态

典型业务模式 · AD ECOSYSTEM

DEMAND

Advertisers

以 CPI / CPA / ROAS / LTV 为目标，寻求高价值用户增长，持续投入广告预算。

广告主 · 增长目标

MATCH & OPTIMIZE

AI Advertising Platform

在每一次曝光中预测用户响应，实时匹配需求与供给，并用反馈持续优化。

连接器 + 优化引擎

SUPPLY

Publishers / Apps

提供广告库存与流量，通过广告实现流量变现。

媒体 / App · 流量供给

广告科技平台不是“卖广告位”
而是用 AI 在广告主增长目标、媒体流量供给与用户响应之间做实时匹配。

稀疏反馈下的实时决策系统

移动效果广告系统· INDUSTRY TRAITS

三大行业特征

广告推荐，
首先要读懂素材本身

multimodal · cold-start · sparse

01

多模态素材是核心供给资产

video / image / text / playable ads / dynamic catalog 持续迭代——广告对象不再只是一个 ID。

02

冷启动是高频问题

creative fatigue 不断刷新创意，新素材 / 新 App / 新 campaign 持续涌入，大量对象缺历史数据。

03

反馈信号极度稀疏

Display Ads CTR ≈ 0.27% · 电商转化 ≈ 2.69% · Netflix 评分密度 ≈ 1.18% ——不能只靠行为监督。

传统特征工程，读不懂广告素材

From Memorizing Ad IDs

To Understanding Ad Semantics

01

语义损失

图片 / 视频 / 文案被压成标签或浅层 embedding，丢失场景、情绪、卖点与用户意图。

02

跨模态无法对齐

图像里的“户外跑步场景”与文案里的“轻量透气跑鞋”，难以映射到同一个语义空间。

03

ID 特征泛化差

Ad / campaign / app ID 依赖历史点击转化，新对象冷启动时模型无从判断该投给谁。

仅有 ID 和 Embedding 还不够

Why Semantic ID

原始 ID 特征

ad_id / app_id

高效、稳定、易工程化——但无语义、冷启动差，只能记住“这是谁”。

记住 IS-WHO

连续 Embedding

text / image vec

能表达语义、可泛化——但连续向量不够结构化，存储 / 索引 / 序列建模都有工程挑战。

表达 LOOKS-LIKE

Semantic ID

[12, 58, 301]

离散、语义化、可组合、可生成——把有语义的 embedding 变成可工程化的特征。

可工程化语义

Semantic ID 的价值：
把“有语义的 embedding”压缩为“可工程化、可组合、可生成的离散特征”。

Semantic ID 如何产生

From Creative to Discrete ID

Ad Creative

image / text / video

Continuous Embedding

连续向量

Semantic ID

[12, 58, 301]

STEP 01

多模态Encoder + 业务对齐

co-click / co-conversion

视觉模型懂猫狗、文本模型懂文案，
但广告推荐需要懂“用户为什么会点”。

STEP 02

Codebook 量化

压缩 / 去噪 / 结构化

Codebook 是一组语义原型，把连续向量映射到有限 token。

Semantic ID 对广告推荐的价值

Why Semantic ID Fits AD-REC

COLD-START

冷启动更友好

只要有图片、视频、文案就能生成语义特征，行为数据不足时为推荐提供先验信号。

01

MULTIMODAL

多模态融合更自然

来自统一的 business-aligned embedding，融合视觉、文案、视频与用户响应相似性。

02

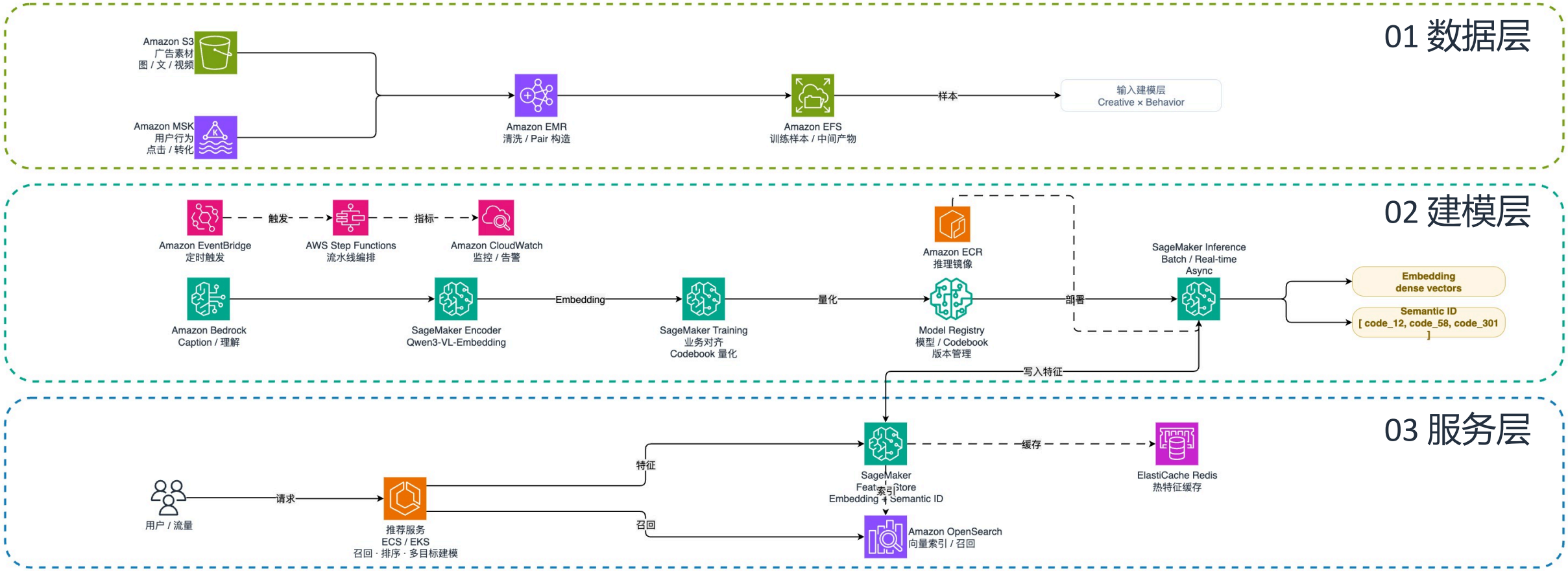
PLUG-IN

更易进入推荐系统

离散 token 直接进召回 / 排序 / CTR-CVR / 序列建模，并为生成式推荐预留接口。

03

基于 SageMaker 的端到端架构



分层

3 层

数据 / 建模 / 服务

服务

10+ 亚马逊云科技

编排 / 训练 / 推理 / 存储

特征

2 类

Embedding + Semantic ID

SageMaker：特征生产流水线

PRODUCE SEMANTIC ID FEATURES

Training Jobs 01

标准化训练：自定义脚本 / 镜像，按需启动 GPU

HyperPod 02

大规模分布式训练，弹性、韧性基础设施，支持 EKS

Real-time Endpoint 03

新素材上线即时生成特征，低延迟同步

Async Inference 04

大输入 / 视频 embedding，近实时批量刷新

Model Registry 05

encoder / codebook 版本编目、审批、可回滚

Pipelines / MLOps 06

串联训练-评估-注册-部署，可自动化、可复现

特征存储与服务

— 01 / MANAGE

Feature Store

统一特征管理，训练 / 推理一致性，是离线与在线的衔接点。

— 02 / RETRIEVE

OpenSearch

稠密向量检索与语义召回，k-NN 找相似素材 / 创意。

— 03 / SERVE

ElastiCache

低延迟在线特征缓存，支撑高并发、毫秒级在线读取。

— 04 / TRACE

Offline / S3

历史快照、训练回溯、A/B 与版本审计，可重算可归因。

Semantic ID 调优是高维实验搜索问题

A High-Dimensional Search Problem

01 设计空间 · DESIGN SPACE

维度耦合
难以人工穷举

— Input Modality — text / image / video

— Business Alignment — co-click / co-conversion

— Training — contrastive loss / 负采样

— Codebook — 容量 / 层数 / 碰撞率 / collapse

02 AUTO-RESEARCH 工作流

人定义赛道
AI 在赛道内实验

— Human defines the arena — 目标 / 数据 / 约束 / 评测

— AI searches the space — 假设 / 配置 / 运行评估

— System keeps the winners — 沉淀 / 归纳 / 下一轮

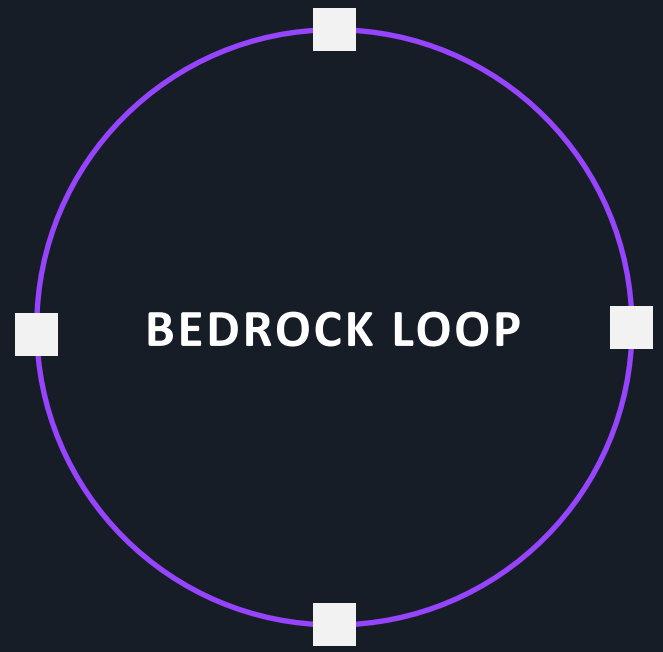
Synapse：基于 Bedrock 构建的 Auto-Research 工具

01 **Project Space**
把研究目标、数据、约束、评测脚本变成机器可读上下文。

02 **Hypothesis Canvas**
把“可以试试”变成可验证、可比较的结构化研究假设。

03 **Experiment Board**
配置 / 指标 / 日志 / 版本 / 产物路径，全程可追踪。

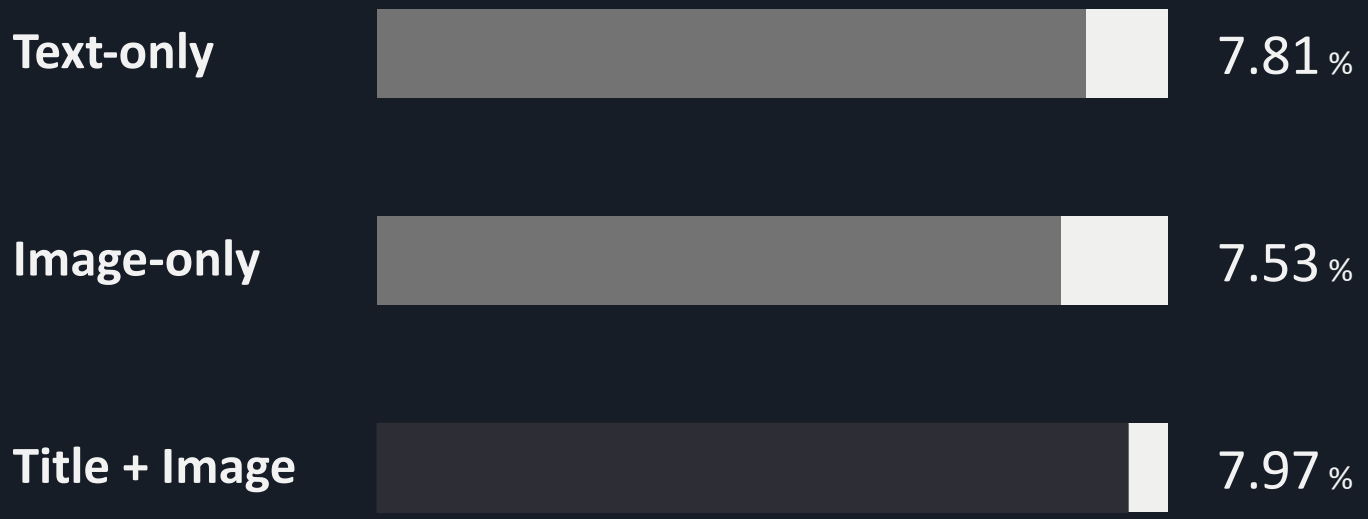
04 **Insight Memory**
自动沉淀最佳实践，进入下一轮实验上下文。



*前述特定亚马逊云科技生成式人工智能相关的服务目前在亚马逊云科技海外区域可用。亚马逊云科技中国区域相关云服务由西云数据和光环新网运营，具体信息以中国区域官网为准。

图文融合优于单模态

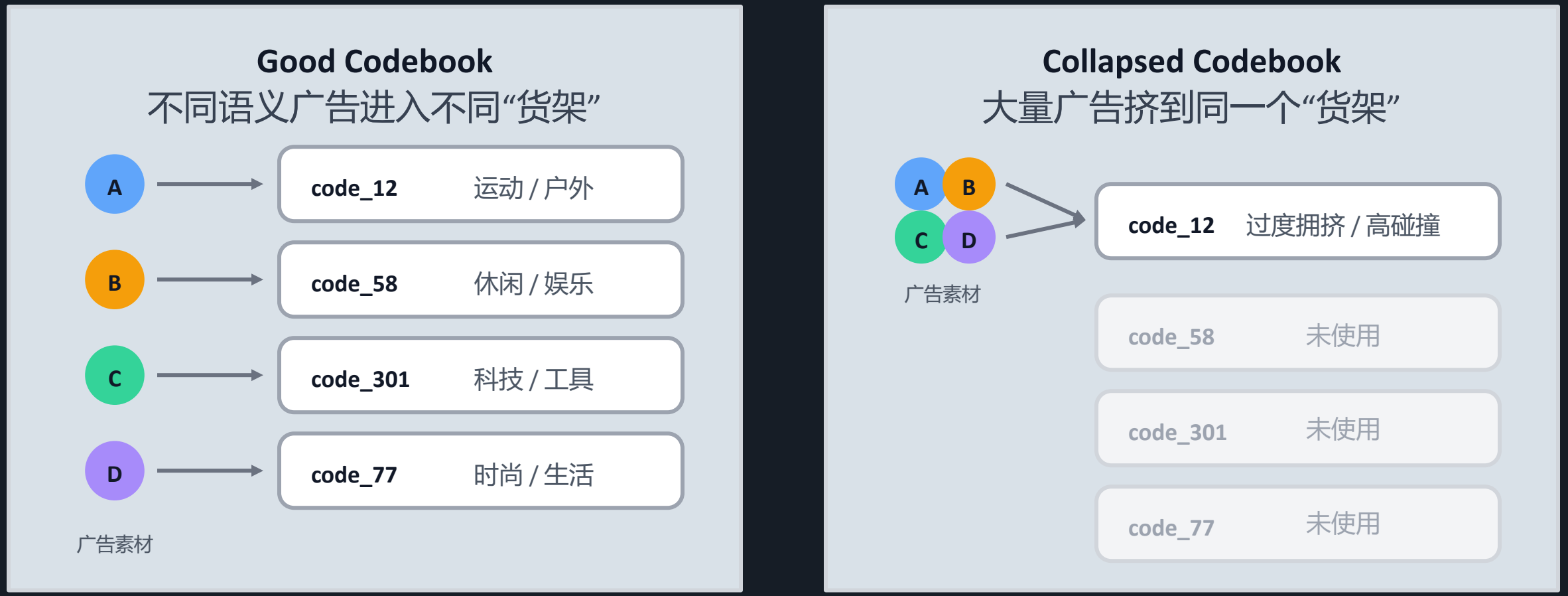
Multimodal > Single-modal



R@10（越高越好）· 文本回答“它是什么”，图片回答“它看起来如何”

——图文融合让 Semantic ID 同时捕捉文本语义与视觉语义。

Semantic ID 的风险：Codebook Collapse



理想状态：语义分散到不同 token；坍缩状态：大量广告挤到少数 token

最佳实践：Codebook 设计与初始化

01 容量与层数 · CAPACITY

10× 广告数量

— 数据 25K items · 最优 512×512（2 层，SID 长度 2）→ R@10 7.97%

— 对比 64×64×64（3 层）→ R@10 7.86%：少层优于多层

— 经验：容量取广告数 10×，优先 2–3 层，碰撞率高时再扩容

02 K-MEANS INIT

73.6% → 9.8%

— 无 K-means init：碰撞率高达 73.6%

— 启用 K-means init：碰撞率降至 9.8%

— 先在 embedding 空间找合理中心，再训练 codebook，显著降低 collapse 风险

最佳实践：建模方法调优

R3-VAE > RQ-VAE

同样生成离散 Semantic ID，不同模型路线显著影响最终推荐效果。

从“记住 ID”到“理解并生成语义 ID”

END-TO-END VALIDATION

Semantic ID + 生成式推荐，端到端超过传统序列推荐基线。

传统序列推荐 · UNIQUE ID + S3-REC

8.03 % R@10

Item 以 Unique ID 表示，主要依赖历史行为学习。

生成式推荐 · SEMANTIC ID + T5

8.29 % R@10

T5-Encoder-Decoder 基于行为序列生成下一个 Semantic ID。

Takeaways

读懂素材，
才能读懂用户。

多模态对比学习特征 + Semantic ID，
在 SageMaker 上端到端工程化落地。

01 **多模态语义特征价值**

业务对齐的多模态 Semantic ID，让冷启动与稀疏反馈下的广告推荐更稳。

02 **SageMaker 端到端工程化**

训练、推理、版本管理与特征服务，把实验性表征做成生产级流水线。

03 **特征建模调优**

基于 Bedrock 的 Auto-Research，把高维 codebook 调优变成可复现的研究闭环。

反馈二维码

您的反馈信息对我们非常重要

请您扫描“调查问卷”二维码，填写问卷

*前述特定亚马逊云科技生成式人工智能相关的服务目前在亚马逊云科技海外区域可用。亚马逊云科技中国区域相关云服务由西云数据和光环新网运营，具体信息以中国区域官网为准。

亚马逊云科技
中国峰会

Thank you