COFFEE：广告排序系统特征增强嵌入协同设计框架

Sohini Roychowdhury、Doris Wang、Qian Ge、Joy Mu、Srihari Reddy
{sroychowdhury1,doriswm,qge2,joymu,sriharir}@meta.com Meta Platforms, Inc.

arXiv:2601.02807v1 [cs.IR] 6 Jan 2026

摘要

多样且丰富的数据源对于商业广告推荐模型准确评估用户在与内容互动前后的兴趣至关重要。尽管扩展用户互动历史可以改善用户兴趣预测，但遵循缩放定律原理，对来自多个来源的活动序列进行嵌入同样重要，这有助于确保用户和广告表示的新鲜度。本文提出一种新颖的三维框架，在不增加模型推理或服务复杂度的情况下增强用户广告表示。第一维度考察纳入多样事件来源的影响；第二维度关注更长用户历史的益处；第三维度则聚焦于通过额外的事件属性和多模态嵌入丰富数据。我们通过比较自然用户互动来源（如内容浏览）与广告曝光来源，评估来源增强框架的投资回报率（ROI）。与自然使用来源相比，即使在线序列长度仅为 10^2-10^4，所提出的方法也能使广告曝光来源的曲线下面积（AUC）和缩放曲线斜率提升1.56–2倍。此外，使用丰富的广告曝光事件来源时，点击率（CTR）预测相较于基线生产广告推荐系统的AUC提高了0.56%，从而改善了更长序列和离线用户广告表示的序列扩展分辨率。

关键词

缩放曲线、特征增强、基于事件的特征、序列、点击率

ACM 引用格式：
Sohini Roychowdhury、Doris Wang、Qian Ge、Joy Mu、Srihari Reddy。2026。COFFEE：用于特征增强嵌入的协同设计框架，应用于广告排序系统。载于*WSDM 2026* 论文集（*Web and Graph Workshop*）。ACM，纽约，纽约州，美国，5 页。

1 引言

表示学习是现代广告排序系统的核心组成部分，在从原始数据中提取有意义的特征并准确衡量用户意图方面发挥着关键作用 [1, 2]。广告排序模型的性能与高质量数据源的可用性和多样性高度相关，这些数据源涵盖自然使用、广告数据及其相互作用。关于大语言模型（LLMs）的早期缩放定律研究表明， [1] 短期内，基于Transformer的表示可使资源投入的收益线性提升；而改进数据编码并增强特征，则能带来长期的指数级收益。在本研究中，我们定义了一个实验框架，用于设计和评估基于用户参与活动的多模态数据源，并证明自然数据源和广告相关数据源对于提升推荐广告点击率（CTR）至关重要 [3]。

面向大规模广告推荐系统开发用户意图驱动的数据源序列面临两项主要挑战。首先，随着多模态内容源快速演进，如何获取具有显著全球覆盖范围的高质量用户与内容交互数据仍是一个尚未解决的难题 [4]。其次，多模态数据源的去噪复杂性，加之用户行为和交互偏好的快速变化，使实时推理成为必要 [2]。此外，对数据隐私和安全日益重视，促使相关法规和指南更加严格，进一步限制了某些数据源的可用性和覆盖范围。在本研究中，我们展示了一种模块化框架，用于从用户参与活动中学习个性化用户意图，并分别通过事件级、序列级和特征级增强来评估和提升学习效果。

本文做出了两项主要贡献。首先，我们提出了一种序列学习框架，将从用户参与活动或事件（例如内容浏览、点赞等）中收集的多模态带时间戳数据，融合为自动化事件特征（EBF）序列。这些 EBF 使基于序列的广告排序模型能够理解短在线序列中广告/内容交互前后的用户行为 [6]。其次，如图 1 所示，通过在多个维度扩展 EBF 序列，我们实现了投资回报率（ROI）提升 2 倍。

这些维度表示按以下方式扩展用户/广告序列：

- 更多事件：捕捉跨多个用户画像和多种表层参与的用户意图。在保持序列长度不变的同时，丰富用户意图档案。
- 更长的序列：通过增加序列长度，捕捉有关用户兴趣的更深层洞察和上下文。
- 语义增强：为每个用户参与事件捕捉补充性的细微语义信号，并利用多模态内容嵌入。

允许免费为个人或课堂使用而制作本作品全部或部分内容的数字或纸质副本，前提是这些副本不得用于营利或商业利益，也不得为此目的进行分发；且副本须附有本声明及首页完整引文。必须尊重本作品中由作者以外的其他方持有版权的组件。注明出处后允许摘编。若要以其他方式复制、转载、发布到服务器或分发至列表，须事先获得明确许可和/或支付费用。请向 permissions@acm.org 申请许可。*Web and Graph Workshop，2026* © 2026 版权所有归所有者/作者所有。出版权已授权给 ACM。ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM

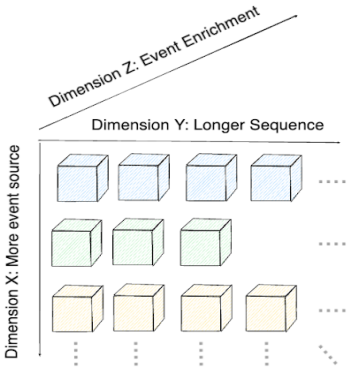

图 1：所提出的三维框架，可灵活适应用户意图，从而丰富在线 EBF 序列。

2 方法与材料

在本节中，我们定义来自多模态数据源的 EBF 序列，并分析在 [5] 中，序列增强对采用预训练嵌入的深度学习模型的影响。

2.1 广告推荐模型

本文使用的改进型广告推荐模型及其与标准稀疏神经网络模型系统的结构差异如图 2 所示。用于“序列学习推荐模型”的事件模块根据事件属性生成事件嵌入。因此，如果 EBF 流包含多个属性（图 2 中的主题、品牌、广告主等），则会进行线性压缩，将这些信息聚合为单个基于事件属性的嵌入。随后，事件模块将每项用户参与活动对应的编码时间戳与该单个基于事件属性的嵌入结合起来，为每位用户生成最终的事件级表示。因此，给定固定聚合时间窗口内（例如图 2 中的 30 天或 2 周）一组经时间戳编码的事件，以及每位用户的某一特定类型 EBF 事件源（例如图 2 中的自然展示、点赞、广告展示等），即可将其定义为跨用户活动的一系列基于事件属性的嵌入（每个事件含有相应属性），即公式 (1) 中的 $\{a_k^i\}$。

$$EBF^n(u) = [aggregate(\{a_k^i\})], \forall i \in [1, ...r], k \in [1, ...K]. \quad (1)$$

这里，i 表示经过属性级聚合的第 i 个带时间戳的用户参与活动。为确保实时推理，我们将事件属性的数量限制为最多 10 个。此外，如图 2 所示，单个事件 $\{a_k^i\}$ 可以是原始用户活动特征、语义 ID 或内容嵌入。

为进一步增强用户表征的动态性，可以增加 EBF 序列长度（属性长度 K 与事件长度 r 的组合）。在本研究中，我们通过分析 EBF 事件类型（n）、EBF 序列长度（r）以及 EBF 属性表征（K）对实时个性化和大规模推荐的不同影响，将这种动态用户表征方法从 [2] 扩展到更广范围。所提出的基于事件的框架能够捕捉长期用户行为，例如用户兴趣中的趋势和季节性模式，并支持长度更长的离线序列表征，其规模达到 10^6 及更高水平 [2]。

2.2 序列扩展模块

我们从以下 3 个维度评估增强 EBF 序列对提升广告/内容推荐效果的影响。

(1) 通过改变 (1) 中的 EBF^n，利用多样化信号进行扩展：通过筛选高价值 EBF，所提出的系统主动从多个垂直场景收集用户交互数据，例如社交媒体参与、内容消费和广告交互。这组多样化信号能够更全面地呈现用户意图，同时捕捉用户当前和长期的偏好，与先前研究 [2] 中的做法一致。

(2) 通过改变 (1) 中的 r，扩展序列长度：鉴于更长的用户交互 EBF 序列能够深入揭示不断变化的兴趣，所提出的系统专为处理更长的序列长度而设计。

(3) 通过在 (1) 中调整 K 以扩展语义：对多模态事件属性（文本、图像、时间上下文）进行编码、拼接和池化，形成稠密嵌入序列，从而得到语义更丰富的嵌入。例如，视频观看事件会利用视觉语言模型生成的内容嵌入和上下文元数据（如观看时长、设备类型）加以丰富。此外，还会应用定制的向量量化技术，高效编码每个属性嵌入中的属性，从而增强并更新事件级用户交互的表示 [7]。

2.3 EBF 数据源

- 自然曝光 EBF：当用户生成内容（UGC）在用户屏幕（网站或移动应用）上显示，且 >= 50% 的内容可见，并且用户观看至少 250 毫秒时，就会生成此事件。事件属性包括 {内容 ID（唯一标识符）、停留时长（可见时长）、媒体类型（例如图像、文本、视频）、位置（在用户个性化信息流中的排名）、时间戳}。
- 广告曝光 EBF：当广告内容的可见度超过 50%，且用户观看时长至少为 250 毫秒时，会生成此事件。事件属性包括 {semantic-ids（内容理解模型生成的广告丰富化元数据）、广告 ID、时间戳}
- 视频观看 EBF：当用户以 >= 30% 的可见度观看视频至少 1 帧后，会生成此事件。事件属性包括 {视频 ID（唯一标识符）、作者 ID、帖子 ID（适用时为父级内容）、停留时长（总观看时长）、页面 ID（上下文展示位置）、内容类型（广告或用户生成内容）、时间戳}

为了丰富 EBF，我们将内容理解嵌入的 k 近邻整理为额外的 EBF 特征 [7]。以下所有 EBF 实验均基于一个月内整理的超过 500 亿个用户样本进行训练。

3 实验与结果

本节通过两组实验展示使用 EBF 进行序列扩展的重要性。首先，我们量化

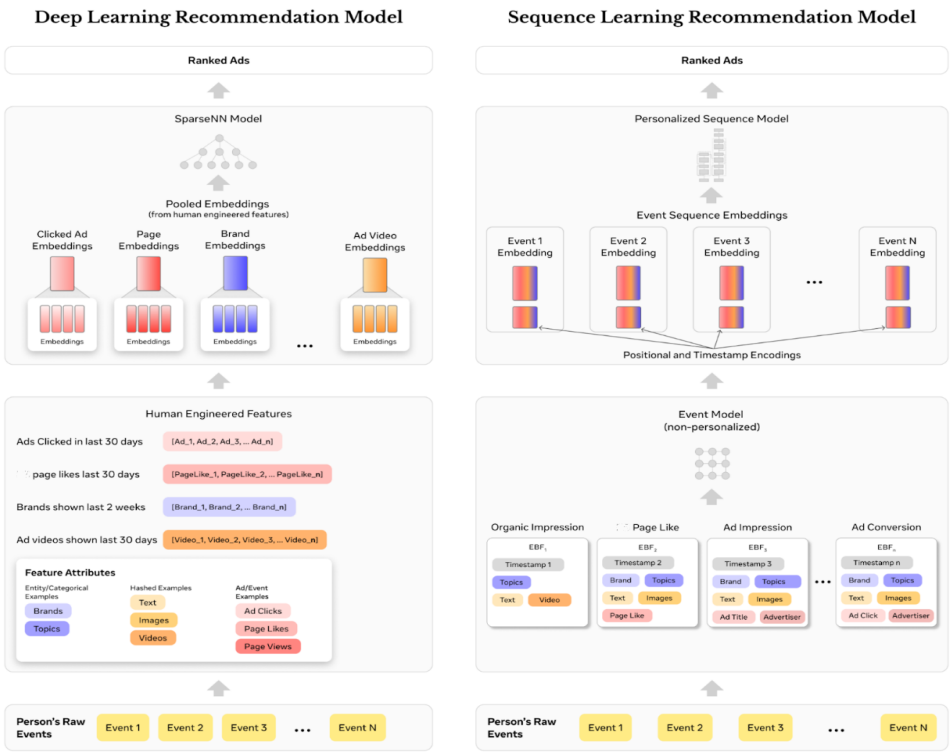

图 2：基于稀疏神经网络的模型（左）与序列学习推荐模型（右）的表示。

序列长度相对增加所带来的点击率提升，以及训练/服务成本的增加。其次，我们定性解释特定用户-广告对的预测意图改进。

3.1 定量 EBF 来源评估

图 2 中的“序列学习推荐模型”用于预测用户点击某条内容/广告的概率（点击率）。我们使用不同的 EBF，通过归一化熵（NE）评估点击率预测的改进情况，其定义如下：

$$\mathrm{NE} = \frac{\frac{1}{N}\sum\left(y_j \log p_j + (1-y_j)\log(1-p_j)\right)}{\hat{p}\log\hat{p} + (1-\hat{p})\log(1-\hat{p})}, \tag{2}$$

其中，N 表示用户样本总数，y_j 表示实际标签，p_j 表示模型预测结果，$\hat{p} = \frac{1}{N}\sum y_j$ 表示先验概率。–

我们在第 2.3 节探讨了 EBF 数据来源中 EBF 变化的影响。为了评估每个 EBF 来源的投资回报率，我们采用图3所示训练容量扩展曲线下面积（曲线下面积）[1], 以及训练容量扩展曲线（表1）中的最佳拟合斜率作为衡量指标。我们通过将每个 EBF 的最大序列长度设为 200 至 2000，分别评估序列长度对三个事件来源的影响（图3中的橙色、红色和蓝色曲线）。此外，我们还评估了在现有高投资回报率 EBF 来源“广告曝光”之上添加额外增强特征所带来的事件增强效果（图3中的绿色曲线）。值得注意的是，类似的增强技术对视频观看和自然曝光 EBF 的影响要小得多，因此未在图3中列出。此外，我们还评估了 EBF 的影响通过 EBF 的三个维度——长度、事件类型（自然内容或广告）和额外丰富化——分别考察归一化熵增益和基础设施扩展成本的变化。

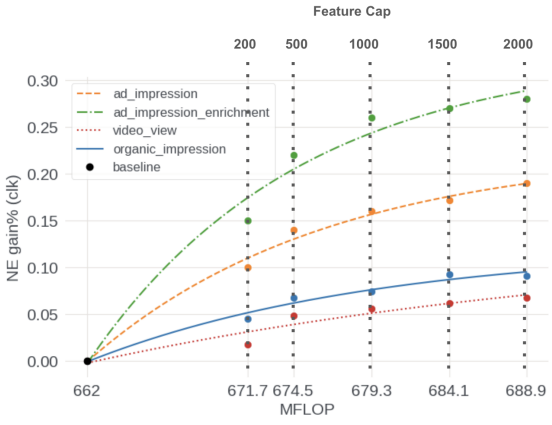

图 3：不同序列长度和丰富度级别下 EBF 来源的训练容量扩展曲线。广告曝光是最强来源（绿色和橙色），其次是自然曝光（蓝色）和视频观看（红色）。基线生产模型的性能由黑色原始数据点表示。

在各种 EBF 变体中，我们观察到以下三点。第一，归一化熵增益曲线在序列较长时往往会趋于饱和。第二，当序列长度超过 1000 后，归一化熵增益的增长速度会随着序列长度的增加而减缓。第三，对于广告展示，事件丰富化对归一化熵增益的提升幅度大于对自然内容曝光的提升幅度。此外，根据 [2]，中的先前工作，我们在表1中评估了图3所示归一化训练容量缩放曲线的曲线下面积和最佳拟合斜率。

表1：EBF 的投资回报率，以曲线下面积和训练容量缩放曲线的最佳拟合斜率衡量。

EBF 数据来源	AUC	斜率
视频观看	0.155	1.82
自然展示	0.235	2.37
广告展示	0.486	4.69
经增强的广告展示	**0.758**	**7.12**

在这里，我们观察到，EBF 增强操作使广告展示来源的 AUC 和斜率分别提升了 1.56 倍和 1.52 倍。相反，对于视频观看和自然展示等自然来源，随着序列长度增加，NE 增益呈现饱和趋势。此外，广告展示来源的投资回报率参数几乎是自然来源的 2 倍。这一结果符合预期，因为与自然来源相比，广告展示更能体现用户对某个广告的未来兴趣，而自然来源则更能体现未来的广告互动。

最后，我们分析了采用所提出的基于 EBF 序列的模型相较于基线 [5] 方法在点击率方面的提升（图3中的原始数据点）。我们将序列长度为 200 的已部署基线模型，与序列长度上限为 200 的效果最佳的增强型广告展示 EBF 进行比较。基线的点击率预测 AUC 为 0.6721，而增强效果最佳的 AUC 为 0.6759，因此在序列长度上限为 200 时，点击率预测 AUC 提升了 0.56%。

3.1.1 限制条件分析。定量分析表明，曲线下面积和斜率都是衡量每个 EBF 来源投资回报率的指标。我们的研究扩展包括将每个来源扩展到长度依次为 10^6 [6] 的离线序列，以确定每个 EBF 来源的最优运行条件。此外，与本文所展示的较短在线序列相比，训练更长的离线用户广告序列的成本低 8 倍。

3.2 定性 EBF 来源可解释性

为了评估 EBF 来源如何提升用户参与度预测，我们利用广告与用户生成序列之间的交叉注意力权重。对于特定广告位，我们提取注意力权重最高的历史用户生成内容事件，以评估用户历史记录与投放广告之间的相关性。对于自然展示事件来源，我们分析新自然展示 EBF 来源带来的 NE 增益，其来源与使用历史用户互动活动数据源的情况有何不同。我们得出以下发现。首先，如图 4 所示，较高的交叉注意力权重能够准确捕捉用户广告意图。虽然注意力权重较高的自然帖子（中图）与用户点击的投放广告（左图）相关，但用户广告

使用历史互动权重进行预测（右图）可能与用户意图的动态相关性较低。这表明，有必要针对用户个性化 EBF，而不是依赖历史广告互动。其次，自然序列能够识别



图 4：使用自然展示 EBF 来源相较于历史基线信号改进广告预测。f ying i内容 f或新的广告 ds in the ab存在 o f related 用户参与度 .

在图 5 中，我们观察到，与历史互动转化事件（右图）相比，高注意力分数的自然内容（中图）更有可能与被点击的投放广告（左图）相关。



图 5：相较于基线，利用自然序列识别新广告的用户意图。

4 结论与讨论

在这项工作中，我们展示了 EBF 来源（包括广告展示、自然内容曝光和视频观看）可以在序列长度、属性和事件多样性方面进行扩展，从而恰当地建模随时间大幅变化的用户与广告互动水平。我们观察到，具有高层级表示属性的广告曝光来源，在在线序列长度仅为 200 时，能够将点击率预测AUC 提高 0.56 %；同时，随着离线序列长度增加至 10^6 [6] 量级，其扩展能力也很强。此外，我们发现，通过内容嵌入为 EBF 序列补充语义信息，可以进一步提升广告曝光 EBF 序列和自然序列带来的收益。未来的工作将致力于集成压缩后的基于大语言模型的嵌入，以便从更长、更丰富的用户历史记录中学习，用于实时预测。

参考文献

[1] Newsha Ardalani 和 Carole-Jean 等人。Wu。2022。理解推荐模型的缩放定律。*arXiv* 预印本 *arXiv:2208.08489*。[2] Alaa Awad 和 Denisa 等人。Roberts。2023。adSformers：Etsy Ads 中基于短期序列和表征多样性的个性化。*arXiv* 预印本 *arXiv:2302.01255* (2023)。[3] Qiwei Chen、Yue Xu、Changhua Pei、Shanshan Lv、Tao Zhuang 和 Junfeng Ge。2022。面向点击率预测的高效长序列用户数据建模。*arXiv* 预印本 *arXiv:2209.12212* (2022)。[4] Paul Covington、Jay Adams 和 Emre Sargin。2016。用于YouTube推荐的深度神经网络。载于第 *10* 届 *ACM* 推荐系统会议论文集。191–198。[5] Maxim Naumov 和 Dheevatsa 等人。Mudigere。2020。DLRM：一种先进的开源深度学习推荐模型。*Meta AI* 2 (2020)。

[6] Dinesh Ramasamy 和 Shakti 等人。Kumar。2025。SIDE：用于从序列中有效学习的语义 ID 嵌入。*AdsKDD* (2025)，KDD Workshop。

[7] Anima Singh、Trung Vu 和 Nikhil et.al. Mehta。2024。借助语义 ID 实现更好的泛化：推荐排序案例研究。载于第十八届 *ACM* 推荐系统会议论文集。1039–1044。