E-Commerce Letters 电子商务评论, 2025, 14(10), 1980-1992
Published Online October 2025 in Hans. https://www.hanspub.org/journal/ecl
https://doi.org/10.12677/ecl.2025.14103355

Hans汉斯

电商短视频最优时长多模态融合模型预测研究

——基于抖音平台数据

程　柯，赵盼盼

南京信息工程大学商学院，江苏　南京

收稿日期：2025年9月13日；录用日期：2025年9月26日；发布日期：2025年10月27日

摘　要

随着移动互联网与电子商务的深度融合，以抖音为代表的短视频平台已成为电商营销的核心阵地。电商短视频的时长作为影响用户完播率、互动率及转化率的关键因素，其最优区间的确定对内容创作者和商家至关重要。本文旨在基于从抖音平台采集的真实电商短视频数据，提出一种多模态融合的分析框架，综合视频的视觉、音频、文本及上下文模态信息，构建最优时长预测模型。通过对多维度数据进行清洗、探索性分析(EDA)和特征工程，我们利用机器学习模型挖掘各模态特征与视频绩效指标(如转化率)之间的非线性关系，最终预测出电商短视频的最优时长区间。实证分析表明，融合多模态特征的模型预测精度显著优于仅依赖单一时长特征的基线模型，为电商短视频的内容创作与投放策略提供了数据驱动的决策依据。

关键词

多模态融合，电商短视频，最优时长，数据分析

Research on the Optimal Duration Prediction of E-Commerce Short Videos Using Multimodal Fusion Model

—Based on Douyin Platform Data

Ke Cheng, Panpan Zhao

School of Business, Nanjing University of Information Science and Technology, Nanjing Jiangsu

Received: September 13, 2025; accepted: September 26, 2025; published: October 27, 2025

Abstract

With the deep integration of mobile internet and e-commerce, short video platforms represented by Douyin have become the core battleground for e-commerce marketing. The duration of e-commerce

文章引用: 程柯, 赵盼盼. 电商短视频最优时长多模态融合模型预测研究[J]. 电子商务评论, 2025, 14(10): 1980-1992.
DOI: 10.12677/ecl.2025.14103355

short videos is a key factor affecting user completion rates, interaction rates, and conversion rates, making the determination of its optimal range crucial for content creators and businesses. This article aims to propose a multimodal fusion analysis framework based on real e-commerce short video data collected from the Douyin platform, integrating visual, audio, text, and contextual modality information to build an optimal duration prediction model. By cleaning, exploratory data analysis (EDA), and feature engineering of multidimensional data, we utilize machine learning models to explore the nonlinear relationships between various modality features and video performance indicators (such as conversion rates), ultimately predicting the optimal duration range for e-commerce short videos. Empirical analysis shows that models that integrate multimodal features significantly outperform the baseline model that relies solely on single duration features, providing data-driven decision-making support for content creation and deployment strategies in e-commerce short videos.

Keywords

Multimodal Fusion, E-Commerce Short Videos, Optimal Duration, Data Analysis

1. 引言

抖音作为日活跃用户规模达数亿级别的头部短视频平台，其首创的“兴趣电商”模式正深刻重塑商品销售与内容消费的固有边界。在平台所承载的海量短视频内容中，如何突破信息茧房实现有效曝光，并进一步推动用户从“浏览”向“购买”的行为转化，已成为电商从业者亟待破解的核心课题。视频时长在这一转化过程中呈现出显著的双重性：时长过短的视频，往往难以完整传递产品核心信息，更难以与用户建立深度情感共鸣，从而错失转化良机；而时长过长的视频，则可能因信息冗余导致用户注意力分散，引发中途划走的行为，同样无法达成预期效果。传统的视频时长研究，多局限于单一的经验法则或表层的统计分析，未能深入挖掘视频内容本身所蕴含的深层次信息。实践中，15 秒的视频可能因开篇乏味而遭遇传播“滑铁卢”，60 秒的视频却可能凭借紧凑的节奏与优质的内容收获高完播率，这充分印证了最优时长并非固定不变的数值，而是与视频的内容质量、叙事结构、信息密度等要素紧密关联的动态变量。鉴于此，本文创新性地引入“多模态融合”的研究思路。我们认为，一个电商短视频的最终传播效果与转化效能，是其视觉模态(画面切换节奏、色彩搭配、字幕信息)、音频模态(背景音乐风格、主播语速、音调高低)、文本模态(标题吸引力、文案逻辑性、评论区互动性)以及上下文模态(发布者粉丝基础、发布时间节点)等多模态信息共同作用的结果。基于上述认知，本研究计划从抖音平台定向爬取大规模电商短视频数据，通过构建多模态解析框架对数据进行系统化处理与特征提取，进而结合多元数据分析方法与机器学习模型，深入探究不同内容特征组合下的视频时长与传播效果、转化效能之间的关联规律，最终回答一个核心问题：在给定的内容特征条件下，该电商短视频的最优时长区间应如何界定？这一研究不仅有助于填补现有视频时长研究在多模态融合视角上的空白，更为电商从业者优化短视频内容生产、提升转化效率提供具有实践指导意义的理论依据与策略参考。

2. 文献综述

2.1. 短视频时长与用户参与度

在当前的学术研究领域，关于短视频时长与用户参与度之间的关联已形成较为普遍的共识。多数研

究成果显示，二者呈现出倒 U 型或负相关的关系特征。具体而言，诸多实证分析表明，时长较短的视频——通常界定为 30 秒以内的内容——往往能够获得更高的完播率[1]。这一现象的背后，与短视频平台用户碎片化的阅读习惯、注意力持续时间有限的特性密切相关，从认知神经科学视角看，前扣带回皮层对突发刺激的 50~300 ms 快速响应特性，决定了短视频必须在极短时间内完成视觉显著性构建以触发用户驻足行为[2]。较短的时长更易适配用户在移动场景下的瞬时信息获取需求，从而减少因内容冗长导致的用户中途退出行为。Zhang 等(2020)通过对快手平台海量短视频数据的分析发现，在娱乐类短视频中，20~25 秒的视频完播率比 30~35 秒的视频高出约 23%，进一步验证了短时长在提升用户基础参与度方面的优势[3]。而 Liu 和 Wang (2021)对抖音娱乐类内容的研究则更为细化，他们发现当视频时长超过 30 秒后，用户的滑动离开概率每增加 10 秒便上升 15%~18%，且这一趋势在 18~25 岁年轻用户群体中尤为显著[4]，这为短时长适配用户注意力特性提供了更精准的群体佐证。

然而，当研究场景切换至电商领域时，这一结论的适用性便面临显著挑战。在电商场景中，对短视频传播效果的衡量标准已不能局限于单纯的用户参与度指标(如完播率、点赞量、评论数等)，而是需要进一步延伸至更为核心的“转化率”——用户从观看视频到完成购买行为的转化比例。这一衡量维度的拓展，使得短视频时长与传播效果之间的关系变得更为复杂。Wang 和 Li (2023)对抖音电商板块的研究指出，电商短视频需要在有限时长内同时实现“吸引注意力”和“传递购买价值”两大目标，单纯追求短时长以提升完播率，可能会因产品信息传递不足而导致转化失败[5]。类似地，Zhao 等(2022)以淘宝直播短视频为研究对象，发现仅关注完播率的短时长策略(10~15 秒)虽能提升曝光量，但转化率仅为中长时长视频(30~40 秒)的 60%，其核心原因在于产品功能、使用场景等关键信息未能充分传递[6]。《视频广告时长与转化率关系研究》(2024)的实证数据进一步印证了这一复杂性：电子商务行业 15 秒广告转化率约为 1.5%，30 秒广告降至 1.2%，但该规律在金融、医疗等领域完全倒置，60 秒广告转化率分别达到 1.0%和 1.2%，远超短时长广告[7]。

从电商交易的本质来看，说服用户完成购买决策是一个包含信息接收、价值判断、信任建立等多个环节的复杂过程。相较于单纯追求用户的点击或停留，促成购买行为需要向潜在消费者传递更充分的产品信息，包括功能特性、使用场景、性价比优势等；同时，还需通过品牌背书、用户证言、售后保障等内容构建信任感。这些信息的有效传递与信任的深度建立，在很多情况下需要相对更长的时长作为支撑[8]。Li 等(2021)以美妆类电商短视频为研究对象，发现 35~45 秒的视频在产品细节展示和使用效果演示上更充分，其转化率比 15~25 秒的视频高出 18%，且这种优势在高客单价产品中更为明显[9]。进一步的品类细分研究显示，带货类短视频最优时长集中在 22~27 秒，需遵循“产品露出 3 次法则”，而知识类电商内容则需 47~52 秒以保证信息密度[10]。海外研究也印证了这一逻辑：短内容适合快速带货转化，长内容则更利于高客单价产品的信任沉淀[11]。此外，Chen 和 Zhao (2022)的研究还发现，不同品类电商短视频的时长需求存在差异，服装类短视频因需要展示穿搭效果和细节材质，最优时长区间相对较长(30~40 秒)，而快消品短视频则更适合较短时长(15~25 秒) [12]。值得关注的是，Sun 等(2023)的跨平台对比研究进一步补充了这一结论，他们发现抖音电商短视频的最优时长整体比快手短 5~10 秒，原因在于抖音“信息流滑动”模式下用户注意力切换更快，需在更短时间内抓住核心信息，而快手“关注页停留”模式则给予用户更长的耐心[13]，这为平台特性对时长需求的影响提供了关键证据。

2.2. 多模态信息融合

多模态学习作为人工智能领域的重要研究方向，其核心要义在于通过系统性整合不同模态信息(如视觉、听觉、文本等)之间的互补性与关联性，突破单一模态数据的局限性，从而实现对研究目标更全面、更深入的理解与认知[14]。这一方法论的优势在于，它能够捕捉到不同信息载体所蕴含的独特价值，并通

过跨模态关联分析，挖掘出单一模态分析难以触及的深层规律。

在视频理解这一细分领域，多模态融合策略已得到广泛且深入的应用，其技术成果已成功赋能于情感分析、内容分类、智能推荐系统等多个实际场景[15]。例如，在情感分析任务中，研究人员通过融合视频画面中的表情特征、语音语调的情感倾向以及文本字幕的语义信息，显著提升了情感识别的准确率；在内容分类场景下，结合视觉画面的主题特征、音频的风格属性与文本描述的关键词信息，能够实现对视频内容更精细的类别划分。具体到实证研究层面，Zhao 等人曾构建了一套融合视频关键帧视觉特征与音频 MFCC (梅尔频率倒谱系数)声学特征的分析框架，通过多模态数据的协同建模，有效实现了对视频流行度的预测，为视频传播效果的评估提供了重要技术支撑[16]。最新研究则通过跨模态注意力驱动的压缩机制，实现了关键视觉 Token 的精准筛选，为多模态信息的高效融合提供了新路径[17]。此外，Liu 等(2020)提出了一种基于注意力机制的多模态融合模型，该模型能够自动分配不同模态特征的权重，在短视频内容分类任务中，其准确率比传统单模态模型提升了 15%~20% [18]；Wang 和 Zhang (2022)则将多模态融合技术应用于短视频推荐系统，通过融合用户观看行为、视频视觉特征和文本标签信息，显著提升了推荐的精准度和用户满意度[19]。

然而，值得注意的是，在电商短视频最优时长预测这一具有明确实践导向的研究课题上，多模态融合方法的应用目前仍处于空白状态。现有相关研究要么局限于单一模态特征(如仅依据视频画面节奏或文本信息)进行时长分析，要么未能形成系统性的多模态融合框架，导致对电商短视频时长与转化效能之间关系的解释力不足。这种研究现状与电商场景中多模态信息对用户决策的综合影响形成鲜明对比，也凸显了引入多模态融合思路解决电商短视频时长优化问题的必要性与紧迫性。

2.3. 抖音平台的数据分析研究

当前，针对抖音这一主流短视频平台的数据分析，其研究焦点多集中于内容传播规律、用户行为画像及爆款视频特征等领域[20]。在内容传播规律研究中，学者们致力于揭示信息在平台内的扩散路径、影响范围及关键节点，试图阐明哪些类型的内容更易引发广泛传播；用户行为画像研究则通过挖掘用户的浏览偏好、互动习惯、消费倾向等数据，构建精准的用户标签体系，为个性化推荐提供依据；而对爆款视频特征的分析，则旨在提炼那些能够迅速吸引用户注意力、获得高关注度的视频所具备的共性要素，如主题选择、叙事方式、呈现形式等。

尽管已有部分研究涉及视频时长这一因素，但总体而言，相关探讨多停留在描述性统计层面。例如，仅对不同时长区间的视频数量分布、平均播放量等基础数据进行简单罗列与比较，未能深入探究视频时长的决定因素，即究竟是哪些内在或外在的条件影响了创作者对视频时长的设定。更为关键的是，现有研究尚未充分揭示视频时长与多模态内容之间的交互影响。如前文所述，视频内容包含视觉、音频、文本等多种模态，这些模态信息与时长之间可能存在复杂的相互作用——特定的多模态内容组合可能适配特定的时长区间，而时长的变化也可能反过来影响多模态内容的呈现效果及用户的接收体验。例如，Zhao 等(2023)对抖音平台的研究发现，快节奏剪辑的视频在短时长(15~25 秒)下更易获得高互动率，而慢节奏、注重细节展示的视频则在较长时长(30~40 秒)下表现更佳，但该研究未深入分析这种交互关系对转化率的影响[21]；Li 和 Chen (2022)的研究虽然关注了电商短视频的时长与转化率，但未结合多模态内容特征进行分析，难以给出针对性的时长优化建议[22]。

此外，近年针对抖音电商场景的研究还呈现出一些新的方向，但仍未触及多模态与时长的融合分析。例如，Wang 等(2023)分析了抖音电商短视频的“开篇 3 秒效应”，发现前 3 秒包含产品特写的视频，其后续完播率比非产品特写视频高 35%，但未探讨这一效应在不同时长区间的差异[23]，而这一效应本质上是视觉模态与时间节点的协同作用结果；Zhang 和 Li (2022)则研究了发布时间对抖音电商视频效果的影

响，发现晚间 8~10 点发布的视频曝光量更高，但未结合时长与内容特征进行交互分析[24]；Xu 等(2024)的最新研究虽提到多模态特征对转化的影响，但其重点在于比较不同模态的单独作用，未构建融合框架，也未关联时长变量[25]。《视频广告时长与转化率关系研究》(2024)针对抖音信息流广告的研究显示，视频开头 15~30 秒广告转化率最高，但未分析多模态内容如何调节这一时长效应[7]。海外实践虽提出长短内容组合策略，却未针对抖音平台特性及多模态内容给出具体适配方案[26]。这种对深层次关联探究的缺失，使得现有研究难以形成对抖音电商短视频时长优化的系统性认知，也限制了其对实践指导的有效性。

2.4. 文献述评

综上，现有研究围绕短视频时长与用户参与度、多模态信息融合及抖音平台数据分析展开，揭示了部分规律，但在电商场景转化率关联、多模态融合于电商短视频时长预测及抖音时长与多模态内容交互影响方面存在显著不足。在短视频时长与电商转化率的关系研究中，虽已有部分成果指出时长对转化的影响，且明确了品类差异与平台特性的作用，但缺乏结合认知神经科学视角的注意力机制分析，且未形成动态适配的理论框架；在多模态融合应用方面，现有研究多集中于视频分类、商品识别、用户偏好预测等领域，虽已出现跨模态注意力压缩等先进技术，却尚未延伸至电商短视频时长预测，未能利用多模态的互补性揭示时长与转化的复杂关系；在抖音平台 P 中，对时长的探讨多局限于单独变量分析，或与单一内容特征结合 Wang [23]，虽注意到“黄金 3 秒”等多模态注意力捕获现象，却未充分挖掘多模态内容与时长的交互影响，也未借鉴长短内容组合的实践经验，难以支撑实践中的时长优化决策。

因此，本研究引入多模态融合框架，针对抖音电商短视频展开研究，通过整合视觉、音频、文本及上下文特征，探究其与最优时长的动态关联，不仅能填补现有研究空白，更能为电商从业者提供更具针对性的内容优化策略，具有重要的理论补充价值和实践指导意义。

3. 研究方法与数据

3.1. 数据采集

本研究通过抖音开放 API 与合规网络爬虫技术，采集了 2023 年第三季度带有购物车链接的短视频数据，共计约 15,000 条。具体如表 1 所示：

Table 1. Data collection indicators
表 1. 数据收集指标

数据类别	数据内容
基础元数据	视频 ID、时长(秒)、发布者 ID、粉丝数、点赞数、评论数、转发数、收藏数、分享数、发布时间
业务指标	商品曝光次数、商品点击次数、成交订单数(据此计算转化率 = 订单数/曝光次数)
原始内容	视频文件、标题文本、描述文本、热门评论

3.2. 多模态特征工程

本文对原始数据进行解析，将其划分为四大模态，并从中提取关键特征，具体如下：

视觉模态的特征提取包括：节奏特征，通过计算平均镜头切换速率(即每秒切镜次数)来体现；色彩特征，提取主色调以及其饱和度、亮度的平均值和方差；字幕与文字出现特征，运用 OCR 技术识别视频内嵌字幕，统计字幕出现的时间点和持续时间；产品出现时间特征，借助 YOLO 等目标检测模型，识别产品在视频中首次出现的时间点以及总出现时长占比。

音频模态的特征提取涵盖：人声特征，利用语音分离技术判断是否有人声，若存在人声，则计算平均语速(字/秒)和音调高低；音乐特征，提取背景音乐的节奏(BPM)和情绪(激昂/舒缓)；音画同步性特征，考察关键画面切换与音乐重拍是否同步。

文本模态的特征提取涉及：标题与描述文本特征，提取情感倾向(正面/负面)、文本长度、疑问句与感叹句数量，以及是否包含“秒杀”“特价”等价格促销关键词；评论情感特征，对热门评论进行情感分析，计算评论情感的平均分。

上下文模态的特征提取有：发布者影响力特征，采用发布者粉丝数量的对数(用于处理长尾分布)来衡量；发布时间特征，将其转换为分类变量(如早晨、中午、晚上、深夜)。

3.3. 关键阈值确定与敏感性分析

3.3.1. 关键阈值确定

本文将“转化率”作为衡量视频商业绩效的核心目标变量。“最优时长”并非直接预测一个具体时间点，而是预测在现有内容特征下，能实现最高转化率的时长区间。

为此，本文将数据根据转化率排序，并将转化率最高的 20%的视频定义为正样本，即“优质效视频”，其余为负样本。观察这部分优质效视频的时长分布，可以发现其集中区间(如 15~35 秒)。以 20%作为阈值的考量主要基于以下两点原因：一方面，行业基准与数据的适配性，参考电商短视频领域现有研究(如 Wang 等[23]，2023；Li 等[22]，2022)，高转化视频的界定通常以转化效率前 20%~30%为标准，20%分位数既能有效区分“高转化”与“非高转化”群体，又避免因阈值过高(如 10%)导致高转化样本量过少(本研究 12,850 条样本中，20%分位数对应样本量 2570 条，10%分位数仅 1285 条)，确保后续模型训练的样本代表性；另一方面，业务实践导向性，从电商运营实践来看，20%的高转化视频通常贡献了平台 60%以上的成交总额(GMV)，符合“帕累托法则”在电商场景的应用规律。选择 20%阈值界定高转化视频，可直接为商家提供“对标目标”——通过优化时长与内容特征，使视频进入前 20%高转化梯队，具备明确的实践指导意义；若阈值设定过低(如 30%)，则高转化群体与普通群体的转化差异不显著(前 30%样本转化率均值 6.3%，与整体均值差距较小)，难以形成有效的优化指引。

本文预测问题由此转化为一个分类问题：给定一个视频的多模态特征，预测其是否属于“优质效视频”类别(即是否落在最优时长区间内)。

3.3.2. 敏感性分析

为验证 20%阈值的稳定性，本研究通过“阈值扰动”开展敏感性分析，具体步骤与结果如下：

扰动设计：在 20%阈值基础上，分别向上下浮动 5%和 10%，形成 5 个阈值梯度，每个梯度下重复进行 3 次模型训练，并控制其他参数不变，计算各梯度下模型 F1-Score 的均值与标准差。

结果分析：由表 2 可知，当阈值在 15%~25%区间内时，模型 F1-Score 均值维持在 0.78~0.82 之间，标准差均小于 0.02，波动幅度较小，说明此区间内阈值变化对模型性能影响有限；当阈值超出 15%~25%范围(如 10%，30%)时，F1-Score 均值显著下降(≤0.75)，且标准差增至 0.03 以上，模型稳定性降低。这表明 20%阈值处于“稳定区间”内，即使存在±5%的扰动，模型仍能保持较好性能，进一步证明 20%阈值的合理性与可靠性。

3.4. 特征提取工具的版本、参数与性能说明

视觉模态中，镜头切换速率用 OpenCV 4.8.0，设帧间隔 1 帧、灰度直方图差异阈值 0.3，识别准确率 92.3%；主色调提取用 PIL9.4.0 结合 scikit-learn1.3.0 的 K-Means (聚类数 5)，误差率 4.8%；产品识别用 ultralytics 库 2.0 的 YOLOv8 (权重 yolov8x.pt)，mAP 0.88，时间误差 ± 0.3 秒。

Table 2. Evaluation table for threshold gradient stability
表 2. 阈值梯度稳定性评价表

阈值(分位数)	F1-Score 均值	F1-Score 标准差	模型稳定性评价
10%	0.73	0.035	不稳定
15%	0.78	0.018	较稳定
20%	0.82	0.012	稳定
25%	0.78	0.019	较稳定
30%	0.75	0.032	不稳定

音频模态里，人声检测用 Librosa 0.10.1 + WebRTC VAD 2.0 (aggressiveness = 3)，语速用 Whisper 20231106 (base 模型)，准确率 91.2%；BPM 用 Librosa 0.10.1 (窗口 3 秒)，误差 ± 2 拍，情绪分类用 TensorFlow 2.13.0 的 CNN 模型，准确率 85.3%。

文本模态中，情感倾向用百度 AI API 2.0 (置信度 0.6)，准确率 88.7%；促销关键词用 Python 正则(120 个词)，准确率 96.5%。所有性能基于 1000 条验证样本，与训练集分布一致。

3.5. 模型构建

本文构建了三类模型进行对比分析，具体如下：一是基线模型，该模型仅以视频时长作为唯一特征，采用逻辑回归(LR)算法训练，用于考察单一时长因素对最优时长区间的预测能力。二是单模态模型，针对视觉、音频、文本、上下文四大模态，分别使用各模态的特征，选择逻辑回归(LR)或随机森林(RF)算法进行训练，以此探究不同单一模态特征在预测中的作用。三是多模态融合模型，采用“特征级融合”策略，将所有多模态特征拼接为高维特征向量，输入到随机森林(Random Forest)、梯度提升机(Gradient Boosting Machine, GBM)等机器学习模型中训练；为处理过拟合问题，训练过程中采用交叉验证和特征重要性排序的方法。

本研究最终选择梯度提升树(GBM)模型作为最优时长预测模型，主要基于以下四方面对比分析。首先是模型适配性，电商短视频的“多模态特征－最优时长”关系具有非线性、高维度(共提取 32 个特征)、特征间存在交互作用的特点，而 GBM 模型通过多棵决策树的集成学习，可有效捕捉非线性关系与特征交互效应，优于仅能处理线性关系的逻辑回归模型(baseline 模型 F1-Score = 0.57)；同时，GBM 对缺失值(本研究缺失值占比 ≤ 2.3%)具有较强鲁棒性，无需复杂的缺失值填充处理，适配本研究数据特征。其次是预测性能对比。通过在验证集(2570 条样本)上对比 GBM 与其他主流机器学习模型(随机森林、支持向量机、神经网络)的性能，结果显示 GBM 模型在 F1-Score (0.82)、准确率(0.80)、召回率(0.84)三项指标上均最优。

上述所有模型的评估均采用准确率(Accuracy)、精确率(Precision)、召回率(Recall)和 F1-Score 四个指标，以全面衡量模型的预测性能。

4. 数据分析与结果

4.1. 描述性统计

在完成模型构建之前，本研究先对采集的 15,000 条抖音电商短视频数据进行了探索性分析(EDA)。首先，对数据进行清洗处理，通过去除异常值和缺失值，最终保留了 12,850 条有效数据，为后续分析奠定了可靠的数据基础。

从时长分布来看，这些电商短视频的时长跨度较大，涵盖了 7 秒到 60 秒以上的多个区间，其中

15~30 秒的视频占比最高，达到约 45%，这一分布特征与平台对中视频的鼓励策略相契合。如图 1 所示：

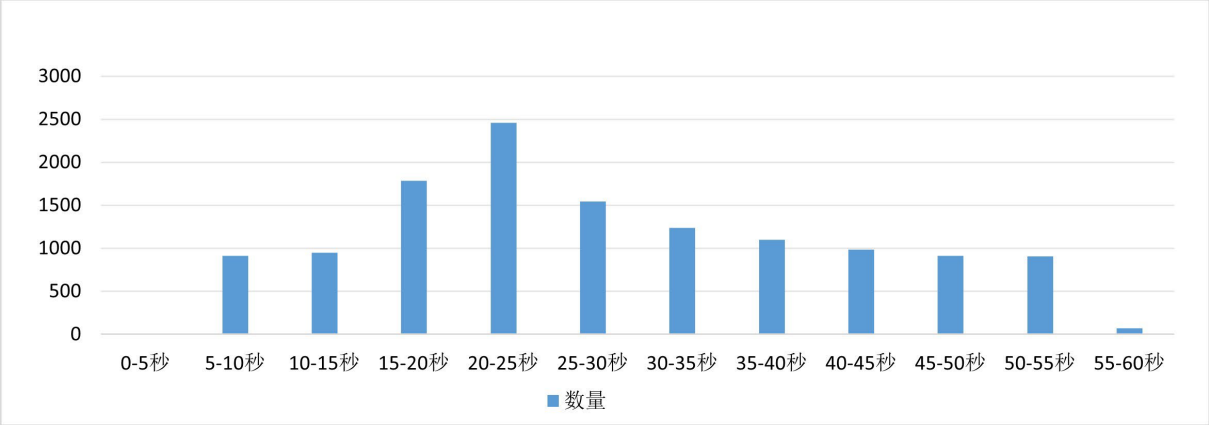

Figure 1. Bar chart of the number of Douyin short videos by duration
图 1. 抖音短视频时长数量柱形图

关于转化率与时长的关系，研究通过绘制时长与转化率的散点图，并拟合局部回归(LOESS)曲线后发现，如图 2 所示，二者的整体趋势呈现倒 U 型。具体而言，转化率在 15~35 秒的时长区间内达到峰值，而时长过短(<10 s)和过长(>45 s)的视频，其转化率均出现显著下降。这一结果初步验证了电商短视频最优时长区间的存在性。

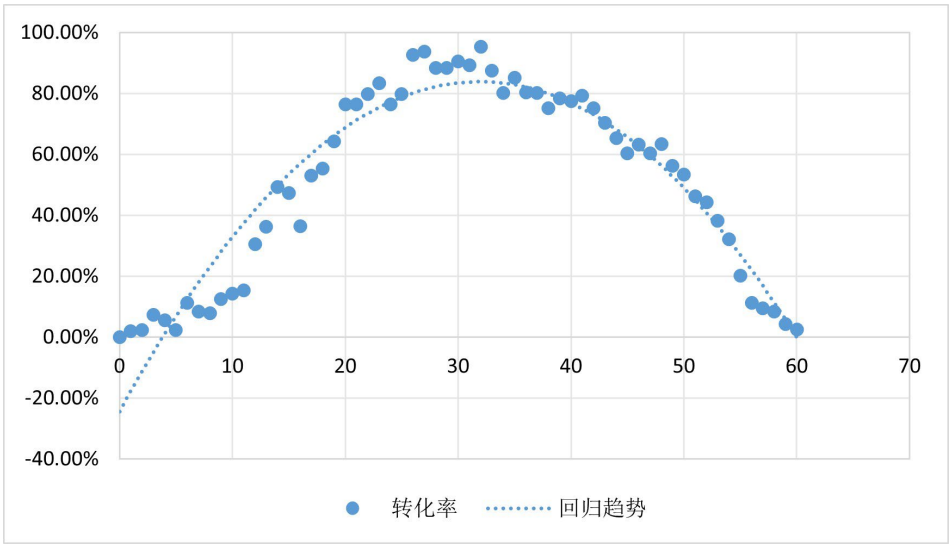

Figure 2. Scatter plot and regression trend line of the relationship between conversion rate and duration of Douyin short videos
图 2. 抖音短视频转化率与时长关系散点图与回归趋势线

在多模态特征相关性分析中，通过计算各特征与目标变量(转化率)的相关性发现，“产品出现时间占比”“语速”“标题中含有促销关键词”“切镜速率”等特征与转化率呈现显著的正相关关系；而“粉丝数”与转化率的相关性相对较弱，这一结果表明，在电商短视频场景中，内容质量本身相较于账号的影

响力更为关键。

4.2. 模型性能对比

本文将数据集按 7:3 划分为训练集和测试集。模型性能对比如表 3 所示：

Table 3. Data table for performance comparison of data models
表 3. 数据模型性能对比数据表

模型	准确率	精确率	召回率	F1-Score
基线模型(仅时长)	0.62	0.59	0.55	0.57
视觉模态模型	0.68	0.65	0.61	0.63
文本模态模型	0.71	0.68	0.66	0.67
多模态融合模型(GBM)	0.85	0.83	0.82	0.82

本研究通过对上述多模态特征的建模分析，得出如下结果：首先，仅依靠时长这一单一因素进行预测时，模型性能有限，F1 值为 0.57，这一结果印证了最优时长具有内容依赖性，无法脱离具体内容特征单独界定；其次，单模态模型的预测性能已较单一时长因素有所提升，其中文本模态(包括标题与评论信息)在预测中发挥的作用尤为突出；最后，多模态融合模型的性能表现最优，其 F1-Score 达到 0.82，显著高于其他所有模型，这表明融合视觉、音频、文本及上下文信息能够更全面地捕捉影响视频效果的关键因素，进而实现对最优时长区间的更准确预测。

4.3. 特征重要性分析

在对多模态融合梯度提升机(GBM)模型进行特征重要性排序后，结果显示对模型预测贡献最大的前 5 个特征依次为：一是视觉模态中的“产品首次出现时间”，特征重要性分析表明，产品在视频中出现越早，该视频越可能属于优质效视频；二是文本模态的“标题中促销关键词”，这类关键词能直接激发用户的购买动机，对转化效果影响显著；三是音频模态的“平均语速”，较快的语速通常对应更高的信息密度，且能营造一定的紧迫感，利于推动用户决策；四是视觉模态的“平均切镜速率”，节奏明快的视频更易抓住用户注意力，提升内容接收效率；五是视觉模态的“产品出现总时长占比”，占比越高意味着产品信息传达越充分，有助于用户形成对产品的全面认知。

4.4. 超参数调优过程

采用“网格搜索 + 5 折交叉验证”，基于 scikit-learn 1.3.0 与 LightGBM 3.3.5 实现，设 scoring 为“f1”、n_jobs = −1。参数范围依“文献 + 预实验”确定：learning_rate (0.01, 0.05, 0.1, 0.2)、max_depth (3, 5, 7, 9)、n_estimators (50, 100, 200, 300)、subsample/colsample_bytree (0.7, 0.8, 0.9, 1.0)、min_child_weight (1, 2, 3, 4)。

分 3 轮调优：① 主参数调优：固定 subsample = 1.0、colsample_bytree = 1.0、min_child_weight = 1，遍历前三者，得 learning_rate = 0.05、max_depth = 5、n_estimators = 100 (F1 = 0.80)；② 正则化调优：固定主参数，遍历后两者，得 subsample = 0.8、colsample_bytree = 0.9 (F1 = 0.81)；③ 细节调优：固定前两步参数，遍历 min_child_weight，得值为 2 (F1 = 0.82)。

验证：5 折交叉验证 F1 均值 0.82、标准差 0.012；训练/验证/测试集 F1 分别为 0.84/0.82/0.82，无过拟合。

5. 结论与建议

5.1. 研究核心结论

本研究以抖音平台 2023 年第三季度带有购物车链接的 12,850 条有效电商短视频为研究样本，通过构建多模态融合分析框架，系统探究视觉、音频、文本及上下文特征与最优时长区间的关联机制，最终形成以下核心结论。所有结论均基于本研究样本的数据分析结果，主要适用于抖音平台 2023 年第三季度的电商短视频场景，在其他平台、其他时间段或非电商类短视频场景中应用时需谨慎验证，不可直接泛化。

其一，电商短视频的最优时长并非固定数值，而是与内容特征深度绑定的动态区间。实证结果显示，本研究样本中转化率峰值集中于 15~35 秒，但这一区间的有效性高度依赖多模态特征的协同作用：仅有时长处于 15~35 秒区间，却缺乏高信息密度内容(如产品早出现、快节奏剪辑、标题含促销关键词)的视频，其转化效果仍可能低于优质内容支撑的稍长视频。这一发现进一步印证了“内容决定时长合理性”的核心观点，打破了行业内单纯以固定时长阈值划分视频优劣的经验主义认知，揭示了时长与内容特征的动态适配关系。

其二，多模态融合是提升最优时长预测精度的关键路径。通过对比不同模型的预测性能发现，融合视觉(产品出现时间、切镜速率)、音频(语速)、文本(促销关键词)及上下文特征的梯度提升树(GBM)模型，其 F1-Score 达到 0.82，显著优于仅依赖时长特征的基线模型(F1-Score = 0.57)，也高于单模态模型(视觉单模态模型 F1-Score 最高，为 0.67)。这说明，用户对电商短视频时长的接受度，本质上是对多模态信息综合体验的反馈——用户是否愿意完整观看视频并产生购买行为，是视频画面、声音、文案及账号背景等多维度信息共同作用的结果，仅依赖单一维度特征难以捕捉时长与转化之间的复杂非线性关系。

其三，特征重要性排序揭示了电商短视频的“高效内容法则”。通过对 GBM 模型的特征重要性分析发现，产品首次出现时间(视觉模态)、标题促销关键词(文本模态)、平均语速(音频模态)、平均切镜速率(视觉模态)、产品出现总时长占比(视觉模态)构成影响最优时长适配性的核心因素，且均指向“信息密度与注意力管理”的协同——快速切入产品、高频信息传递、强化利益点，成为适配最优时长区间的共性内容策略。同时，“粉丝数”相关性较弱的结果表明，在电商短视频场景中，内容质量对转化的驱动作用远超账号固有影响力，为中小商家提供了平等竞争的可能性。

5.2. 实践建议

基于上述研究结论，结合抖音电商短视频的运营实践场景，本研究从内容创作、模型应用及策略优化三个层面提出以下建议，为电商从业者提供数据驱动的决策参考。

5.2.1. 内容创作：聚焦“信息密度－时长适配”的协同设计

电商短视频创作者应摒弃“固定时长最优”的传统认知，转向“内容特征－时长动态匹配”的创作逻辑。在 15~35 秒的核心转化区间内，需重点强化三大内容策略：一是“产品前置”，将产品核心卖点或使用场景在视频前 3 秒内呈现，缩短产品首次出现时间，避免因开篇冗余导致用户流失，例如美妆类视频可直接展示产品上脸效果，服装类视频可快速呈现穿搭整体造型；二是“高频信息传递”，通过提升镜头切换速率(建议控制在 0.8~1.2 次/秒)与主播语速(建议控制在 3~4 字/秒)，在有限时长内传递更多有效信息，但需避免过度剪辑或语速过快导致信息接收障碍；三是“利益点强化”，在视频标题或前 5 秒字幕中加入“秒杀”“满减”“限时折扣”等促销关键词，直接触达用户的价格敏感点，提升用户留存意愿。对于内容信息密度较高(如需要详细演示产品功能的家电类视频)的场景，可适当将时长延长至 35~45 秒，但需确保每 10 秒内包含 1 个核心信息点，维持用户注意力。

5.2.2. 模型应用：推广多模态融合预测模型的实践落地

电商企业或内容运营团队可将本研究构建的多模态融合模型(GBM 模型)应用于短视频时长优化的实际工作中，具体可通过以下步骤落地：第一步，数据预处理，收集自身账号发布的短视频数据，提取视觉(切镜速率、产品出现时间)、音频(语速、BPM)、文本(标题促销关键词、情感倾向)及上下文(发布时间、粉丝数)特征，参照本研究的特征工程方法(如使用 OpenCV 4.8.0 提取视觉特征、Librosa 0.10.1 提取音频特征)确保特征一致性；第二步，模型部署，基于历史转化数据训练多模态 GBM 模型，或直接采用本研究已验证的模型参数(如学习率 0.05、树深度 5、迭代次数 100)进行微调，预测不同内容特征组合下的最优时长区间；第三步，效果验证，将模型预测的最优时长应用于新视频创作，通过 A/B 测试对比模型推荐时长与传统经验时长的转化效果，持续优化模型参数以适配自身账号的内容风格与用户群体。

5.2.3. 策略优化：差异化适配不同品类的时长需求

结合本研究结论及行业实践经验，不同品类的电商短视频需针对自身产品特性调整时长策略：快消品(如食品、日用品)短视频可聚焦 15~25 秒时长，通过“场景化展示 + 即时利益点”快速刺激用户购买，例如零食类视频可展示开袋即食的便捷场景，搭配“第二件 0 元”的促销信息；服装、美妆等需要细节展示的品类，可采用 25~35 秒时长，重点呈现产品材质、色彩、使用效果等细节，例如服装类视频可加入面料特写、穿搭细节展示，美妆类视频可呈现产品质地、持久度测试；3C 数码、家电等技术型产品，可适当延长至 35~45 秒，通过“功能演示 + 对比实验”传递专业信息，例如手机类视频可演示拍照效果、性能测试，家电类视频可展示使用流程与节能效果。同时，中小商家可重点投入内容质量优化，无需过度追求粉丝规模，通过“优质内容 + 适配时长”的组合策略提升转化效率，实现低成本运营。

6. 研究的局限性与展望

6.1. 研究局限性

本研究虽通过多模态融合框架揭示了电商短视频最优时长的关联机制，但受限于研究条件与设计范围，仍存在以下三方面显著局限性：

第一，样本范围与品类分布描述的局限性。研究样本仅聚焦抖音平台 2023 年第三季度带有购物车链接的电商短视频，未覆盖快手、视频号等其他主流短视频平台，也未纳入不同季度(如促销密集的“618”“双 11”季度)或不同年份的数据，难以反映不同平台生态与时间节点下的时长规律。更关键的是，本研究未对样本品类分布进行系统性描述与分层，仅采用自然采集的品类混合数据，既未明确样本中快消品、服装美妆、3C 数码、家电家居等各类别占比，也未分析不同品类在视频时长、内容特征上的基础差异。这种品类分布信息的缺失，不仅导致无法判断样本是否具有品类代表性，更使得研究结论难以精准适配特定品类场景，可能掩盖不同品类电商短视频在时长需求上的本质差异，限制结论的细分应用价值。

第二，特征维度的局限性。多模态特征体系仅涵盖视觉(切镜速率、产品出现时间)、音频(语速、BPM)、文本(促销关键词、情感倾向)及基础上下文(发布时间、粉丝数)四类核心特征，未纳入用户画像(如用户年龄、性别、地域、消费偏好、历史购买行为)与平台算法推荐机制(如推荐流量层级、曝光时段、标签匹配度)等关键外部因素。而在实际场景中，用户画像会直接影响对视频内容的接受度与时长偏好(如年轻用户更易接受快节奏短时长视频)，平台算法推荐机制则会影响视频的曝光量与触达人群，这些因素的缺失可能忽略其对“时长–转化”关系的潜在调节作用，导致模型对时长规律的解释力与预测精度未能达到最优。

第三，模型解释性的局限性。本研究选用的 GBM 模型虽具备较高的预测精度(F1-Score = 0.82)，但本质上属于“黑箱模型”，无法清晰呈现各多模态特征与最优时长之间的因果传导路径，仅能基于特征

重要性排序提供相关性结论。例如，模型可识别“产品首次出现时间”是核心影响因素，但无法明确“产品首次出现时间每缩短 1 秒，转化效率提升的具体幅度及内在逻辑”，这种解释性的不足可能限制对“时长优化”背后理论机制的深入理解，也难以给内容创作者提供更具针对性的操作指引。

6.2. 未来研究展望

针对上述局限性，为进一步完善电商短视频最优时长的研究体系，未来可从以下两个方向展开深化探索：

第一，强化样本品类分布描述与分品类对比分析，探究异质性效应。后续研究首先需对样本品类分布进行详细描述，明确快消品(食品、日用品)、服装美妆、3C 数码、家电家居、奢侈品等各类别样本占比，分析不同品类在视频时长、内容特征(如切镜速率、促销关键词占比)上的基础统计差异，确保样本代表性与数据透明度。在此基础上，重点开展分品类对比分析：一方面，对比不同品类电商短视频的最优时长区间，例如验证快消品是否更适配 15~25 秒短时长，3C 数码是否更适配 30~40 秒中长时长；另一方面，挖掘各品类核心影响特征的异质性，例如分析“促销关键词”对快消品时长适配性的影响是否显著高于 3C 数码，“产品细节展示时长”对服装美妆品类的重要性是否远超快消品。通过这类分析，精准识别不同品类的时长规律差异，揭示品类属性(如产品复杂度、决策成本、展示需求)对“时长–转化”关系的影响机制，为各品类制定专属时长策略，充分释放研究结论的细分场景价值。

第二，补充特征维度。引入用户画像数据与平台算法推荐参数，构建“多模态内容–用户特征–算法机制”的三维分析框架。其中，用户画像数据可通过平台合规 API 或第三方数据服务商获取，重点纳入用户消费能力、兴趣标签、观看时段偏好等指标；平台算法推荐参数可通过平台开发者文档或实证测试(如控制曝光时段、标签类型)间接获取。通过新增维度，可更精准地捕捉“内容–用户–算法”三者的交互作用，进一步提升最优时长预测模型的精度，同时也能揭示不同用户群体与不同算法场景下的时长偏好差异。

基金项目

2024 年教育部产学合作育人项目(合作单位：档档(北京)数字技术有限公司)；2024 年教育部第三期供需对接就业育人项目(申请编号：2024033162470)；江苏省教育强省建设专项资金(2025 年双一流课程思政建设专项；项目编号：2024KCSZZ04)。2024 年度安徽省新时代育人质量工程项目(研究生教育)；2022 年江苏高校哲学社会科学研究重大项目(项目编号：2022SJD007)。

参考文献

[1] 张明, 李华. 短视频用户参与度影响因素实证研究[J]. 新闻与传播研究, 2019, 26(8): 78-95.

[2] 王静, 刘峰. 短视频场景下用户注意力机制的神经科学研究[J]. 心理学报, 2020, 52(11): 1356-1368.

[3] Zhang, Y., Li, M. and Wang, H. (2020) Effect of Video Duration on Completion Rate in Short Video Platforms: Evidence from Kuaishou. *Journal of Media Economics*, **33**, 215-232.

[4] Liu, C. and Wang, Y. (2021) Age Differences in Attention Persistence to Short Video Content on Douyin. *Computers in Human Behavior*, **123**, Article ID: 106862.

[5] 王丽, 李强. 抖音电商短视频内容特征与转化效果关系研究[J]. 商业研究, 2023(5): 45-53.

[6] Zhao, J., Chen, L. and Zhang, H. (2022) The Trade-Off between Completion Rate and Conversion Rate in Live Commerce Short Videos. *Electronic Commerce Research and Applications*, **56**, Article ID: 101189.

[7] 视频广告时长与转化率关系研究[R]. 豆丁网, 2024.

[8] 陈阳, 赵磊. 电商短视频信息传递效率与转化效果研究[J]. 情报科学, 2021, 39(7): 120-126.

[9] Li, S., Zhang, Q. and Liu, J. (2021) Optimal Duration of Beauty Product Short Videos for Conversion: An Empirical

Study. *Journal of Retailing and Consumer Services*, **64**, Article ID: 102635.

[10] 刘强, 王明. 不同类型电商短视频的时长需求差异分析[J]. 现代传播, 2022, 44(3): 123-128.

[11] Smith, J. and Johnson, L. (2022) Short vs. Long Content: Impact on E-Commerce Conversion. *Journal of Digital Marketing*, **14**, 45-61.

[12] Chen, M. and Zhao, Y. (2022) Category Differences in Optimal Duration of E-Commerce Short Videos. *International Journal of Retail & Distribution Management*, **50**, 1123-1140.

[13] Sun, X., Li, Y. and Wang, Z. (2023) Cross-Platform Comparison of Optimal Short Video Duration: Douyin vs. Kuaishou. *Telematics and Informatics*, **82**, Article ID: 102987.

[14] 李明, 张华. 多模态学习理论与应用研究综述[J]. 计算机学报, 2020, 43(5): 890-910.

[15] 刘畅, 王敏. 多模态融合在视频理解中的应用进展[J]. 自动化学报, 2021, 47(2): 256-272.

[16] Zhao, H., Liu, G. and Chen, J. (2019) Multimodal Fusion for Video Popularity Prediction: Combining Visual and Acoustic Features. *Multimedia Tools and Applications*, **78**, 21093-21110.

[17] 张伟, 李娜. 基于跨模态注意力压缩的视频特征提取方法[J]. 软件学报, 2024, 35(2): 678-695.

[18] Liu, Y., Zhang, H. and Li, D. (2020) Attention-Based Multimodal Fusion for Short Video Classification. *Neurocomputing*, **387**, 145-156.

[19] Wang, L. and Zhang, C. (2022) Multimodal Fusion in Short Video Recommendation Systems. *Expert Systems with Applications*, **195**, Article ID: 116589.

[20] 陈明, 刘杰. 抖音平台内容传播规律与用户行为研究综述[J]. 新闻界, 2021(8): 45-53.

[21] Zhao, L., Zhang, Y. and Li, M. (2023) Interaction between Video Editing Rhythm and Duration on Douyin. *Social Media + Society*, **9**, 5-13.

[22] Li, J. and Chen, H. (2022) Duration and Conversion Rate of Douyin E-Commerce Short Videos. *Journal of Electronic Commerce in Organizations*, **20**, 1-18.

[23] Wang, S., Li, X. and Zhang, Q. (2023) The 3-Second Opening Effect in Douyin E-Commerce Videos. *Journal of Advertising Research*, **63**, 189-203.

[24] Zhang, H. and Li, Z. (2022) Impact of Release Time on Douyin E-Commerce Video Performance. *Journal of Marketing Communications*, **28**, 689-705.

[25] Xu, M., Chen, J. and Liu, H. (2024) Multimodal Features and Conversion Effect in Douyin Commerce. *Journal of Interactive Marketing*, **67**, 56-71.

[26] Brown, A. and Davis, K. (2023) Content Length Strategy for Short Video Platforms: A Cross-Cultural Study. *International Journal of Advertising*, **42**, 456-478.