短视频中的大规模多模态传播解读：TikTok 上视觉、音频与文本心理健康话语

Mingyue Zha[1],**Ho-Chun Herbert Chang**[1]

[1] 达特茅斯学院定量社会科学项目，美国新罕布什尔州汉诺威市 03755

arXiv:2601.15278v1 [cs.MM] 21 Jan 2026

摘要

短视频平台将文本、视觉和音频整合为复杂的传播行为，但现有研究将这些模态孤立分析，缺乏能够大规模解读它们共同作用的框架。本研究提出一套流程，将自动化多模态特征提取与基于夏普利值的可解释性相结合，以分析文本、视觉和音频如何共同影响互动。我们将此框架应用于 162,965 条 TikTok 视频和 814,825 张有关社交焦虑障碍（SAD）的图像，发现面部表情在预测观看量方面优于文本情感倾向；信息性内容比情感支持更能吸引关注；跨模态协同效应则呈现出依赖阈值的影响。这些发现表明，多模态分析能够揭示单模态方法无法观察到的交互模式。在方法论层面，我们贡献了一个可复现、可解释且适用于不同领域的多模态研究框架；在实质层面，我们深化了对算法中介环境中心理健康传播的理解。

引言

社交媒体的兴起从根本上改变了人际传播，使数字媒介支持的交流得以实现，并呈现出网络化、异步和公开的特征。近年来，多模态平台迅速普及，将文本、图像、视频、音频和短暂内容融入同步与异步形式之中。Instagram、TikTok 和 Snapchat 等平台更倾向于视觉和表演性传播，而非基于文本的交流（Violot 等，2024年；Huang、Lei 和 Chen，2025年；Xie、Lin 和 Bai，2025年），这要求人们具备“多模态素养”，并重塑了人们在线表达身份、情绪和关系联结的方式。Kress 和 van Leeuwen 将意义建构概念化为通过整合的符号系统实现，而非仅依赖口语语言（Kress 和 van Leeuwen，2001年）。如今，在线人际传播在视觉、文本和算法的交织环境中运作，因此不仅需要理论框架，也需要能够兼顾数字交互各种感官维度的方法论框架（Andalibi、Ozturk 和 Forte，2017年；Niu 等，2021年）。

尽管短视频日益重要，但大规模分析此类内容的方法工具仍不够成熟。现有的许多社交媒体研究聚焦于单一模态（最常见的是文本）（Islam et al. 2025），而多模态研究往往依赖小样本和人工标注（Wang 2021年；Wu et al. 2025年）。因此，对于如何从大型短视频数据集中系统地提取和解读多模态信号，目前尚未形成广泛共识（Ali and Molla 2025年）。研究人员需要一套有原则的工作流程，在统一框架内比较文本、视觉和音频特征的相对贡献。

机器学习领域的最新进展，包括预训练语言模型、图像识别和音频嵌入，使大规模整合这些技术在计算上成为可行之举（Wang 等，2023）。新兴研究表明，大语言模型辅助内容分析能够有效提取并标注在线视频中的视觉特征（Li 等，2024；Liu 等，2025）。然而，这些工具的部署往往较少关注可解释性、可复现性或跨模态可比性。

本文通过提出一套用于短视频社交媒体多模态分析的可扩展流程，弥补了这一不足。我们的方法整合了文本、图像和音频模态的自动特征提取，并结合基于沙普利值的可解释性，从而能够在统一的解释框架内直接比较异质特征。通过将多模态标注视为概率输入并应用夏普利值分析，我们的方法不仅能将注意力和参与度结果归因于单个模态特征，还能归因于它们之间的交互作用。此外，我们还在点级解释器的基础上构建了一个可解释框架，以助力社会科学解释。

我们通过一项关于社交媒体心理健康话语的案例研究，展示了这一方法论的实用价值。心理健康内容尤其适合进行多模态分析，因为情感表达分布在书面文字、口语和视觉呈现之中（Yazdavar 等，2020）。本文探讨了不同的多模态元素如何影响与社交焦虑障碍（SAD）相关的心理健康 TikTok 视频中的注意力动态。

文献综述

数字空间中的多模态传播

从以文本为主的社交媒体平台转向多模态社交媒体平台，标志着传播方式发生了根本性转变

。Kress 和 van Leeuwen 认为，当代传播依托整合的符号系统，而非仅仅依靠口头语言；其中，视觉、文本和听觉元素作为彼此独立但相互关联的意义建构资源发挥作用（Kress and Van Leeuwen 2001）。在数字环境中，这些模态相互作用，有时彼此强化，有时则相互矛盾。

TikTok、Instagram Reels 和 YouTube Shorts 等短视频平台充分体现了这种多模态复杂性。这些平台重视觉表现和算法策划，轻传统的文本话语。这些平台的可供性——竖屏视频格式、音乐整合、文本叠加功能以及快速内容消费——营造出独特的交际环境，在其中，成功取决于能否协调不同模态的元素。媒介丰富度理论认为，面对面交流传统上被视为“最丰富”的媒介，因为它整合了语言、声音和视觉线索（Daft 和 Lengel，1986年）。短视频平台则在异步、由算法介导的形式中近似实现了这种丰富度。

然而，我们对这些多模态元素在实际中如何发挥作用知之甚少；用于开展此类分析的综合框架虽日渐增多，但仍然稀缺。

数字空间中的心理健康披露

社交媒体已成为心理健康传播、同伴支持和健康信息搜寻的主要场所。关于网络心理健康披露的研究揭示了复杂的动机：个人寻求情感认可、实用的应对策略、通过常态化讨论减少污名，并与有相似经历的人建立联系（Andalibi、Ozturk 和 Forte，2017年）。TikTok 等平台已成为青年心理健康话语尤为重要的空间，用户创作的内容涵盖个人叙事、同伴支持、心理健康教育和倡导（Milton 等，2023）。

自我披露理论认为，透露个人信息具有建立关系的功能，但数字自我披露受到的限制不同于面对面互动。在线披露通常是公开的、持久留存的，并受到算法中介，其受众范围远超即时社交网络（Bazarova 和 Choi，2014 年）。数字披露的永久性和可见性既能放大其益处（找到社群、减少孤立感），也会加剧其风险（隐私顾虑、不受欢迎的注意力、污名化）。对于社交焦虑障碍患者而言，尤其如此——这种障碍的特征是强烈害怕受到社会评价以及处于表现情境中，而制作突出自我的视频内容则构成了一个显著的悖论。视觉自我呈现是社交焦虑障碍患者焦虑的核心诱因，却成了他们讨论自身社交焦虑经历的主要媒介。

对在线支持社群的研究指出，社会支持有不同形式：情感支持（表达共情与关怀）、信息支持（建议与实用指导）、工具性支持（有形的协助），以及评价支持（为自我评价提供反馈）（Cutrona 和 Suhr，1992年）。这些区别很重要，因为不同类型的内容可能满足不同受众的需求，也可能受到平台算法不同程度的重视。

近期针对 TikTok 心理健康内容的研究发现，算法推荐系统会影响哪些心理健康叙事能够获得曝光，这可能同时影响有益的同伴支持和错误信息的传播（Bickham 等，2024）。Milton 等（Milton 等，2023）记录了患有心理健康状况的 TikTok 用户如何通过算法策划的内容找到社群并获得认可，但也指出了人们对同伴生成的健康信息质量和准确性的担忧。围绕社交焦虑的社交媒体讨论因此成为一个有用的试验场，因为人际传播和自我呈现具有高度自我觉察的特性。然而，先前研究大多依赖定性方法或小规模定量分析，无法捕捉多模态表达的全部复杂性与多样性。

社交媒体的多模态分析：当前方法与局限

关于社交媒体观看量的研究，大多沿着单模态的路径展开。基于文本的研究考察语言特征——情感倾向、词汇多样性、情感强度和语言风格——如何预测受欢迎程度（Vilares、Alonso 和 G´omez-Rodr´ıguez，2015；Hong、Doumith 和 Davison，2013）。视觉分析则关注图像美学、色彩构成、人脸存在和物体识别（Trzcinski 和 Rokita，2017；Wu 等，2017）。音频研究虽然较少，但会考察音乐特征、语音韵律和言语模式（Li 等，2024）。这些单模态方法带来了宝贵的见解，但必然无法完整说明多模态内容的作用机制。

现有的少数多模态研究通常结合两种模态的特征——最常见的是文本和图像——并将拼接后的特征向量输入预测模型（Meghawat 等，2018）。这种方法虽然能提升预测准确率，优于单模态模型，却将多模态特征视为一个扁平的输入空间，因此难以理解不同模态各自的贡献及其相互作用。更重要的是，这些研究往往依赖小规模样本和人工标注，限制了分析规模和研究结果的泛化能力。正如 Chen 等（2025）所示，多模态标注需要投入大量人力；而平台规范快速演变，也意味着人工编码的数据集很快就会过时。

计算机科学领域的近期研究使用深度学习来分析视觉、音频和文本特征如何共同预测 TikTok 创作者影响力等级。然而，与深度学习中常见的情况一样，嵌入表示牺牲了可解释性——研究人员难以确定究竟是哪些具体特征驱动了预测，也难以了解来自不同模态的特征如何相互作用。对于希望理解内容特征塑造注意力之机制的社会科学家而言，仅凭预测准确性是不够的。

计算社会科学中的可解释性

对于社会科学家而言，可解释性不仅是一项技术要求——它是科学可推广性的基础，也是干预措施设计的规范性基础。沙普利值源自合作博弈论，为模型解释提供了一种有原则的方法（Winter 2002年）。其核心思想是将每个特征视为对预测“收益”作出贡献的“参与者”，并通过考察该特征在其他特征所有可能联盟中的平均边际效应，计算其贡献。SHAP（夏普利加性解释）将这一概念应用于机器学习模型，为模型提供了一个统一框架，将任何预测分解为各项特征贡献之和（Lundberg 和 Lee（2017））。

近期将 SHAP 应用于计算社会科学的研究，已证明其有助于理解表格数据场景中的特征重要性（Salih et al. 2025）。然而，SHAP 尚未被系统地应用于多模态社交媒体分析，以保留并利用模态结构。虽然已有标准的全局 SHAP 汇总（例如，平均绝对归因），但它们无法提供一种类似系数的方向性汇总，并根据特征存在/强度进行条件化；而这往往正是社会科学家所需要的。

现有文献揭示了几个关键空白。首先，鲜有研究致力于开发一套全面的自动化流程，以大规模提取短视频中的文本、视觉和音频模态所包含的理论驱动特征。其次，尽管已有多模态预测模型，但它们往往缺乏可解释性，难以理解模态内部及不同模态之间的具体特征如何影响结果。第三，心理健康传播研究尚未系统考察情感表达如何因模态而异。

本研究通过引入一套自动化多模态分析流程来弥补这些空白，该流程将用于语义特征的零样本分类与基于沙普利值的可解释性相结合。我们将这一框架应用于 TikTok 上有关社交焦虑障碍的心理健康话语，并探讨以下三个研究问题：

RQ1：在描述心理健康相关话语时，文本、音频和视觉情感倾向有何差异？

RQ2：文本、视觉和音频模态在多大程度上影响短视频中心理健康相关内容的注意力动态？

RQ3：文本、视觉和音频特征如何共同影响短视频中心理健康内容的注意力动态？

通过回答这些问题，我们在方法论层面（提供一个可复现的可解释多模态分析框架）和实质层面（深入揭示心理健康传播在算法媒介环境中的运作方式）均作出了贡献。

方法论

多模态流程概览

我们的分析流程将社交媒体视频中的文本、视觉和音频信息整合到一个统一的可解释多模态框架中。图 1 概述了该流程。每段视频都被分解为三种模态，并采用适合各自模态的特征提取方法独立处理。文本内容通过情感分析和零样本话语标注进行分析；视觉帧采用面部分析和零样本视觉标注进行编码；音轨则用于分析声学特征，并将口语内容转录出来，以供后续文本分析。对于各个模态，零样本标注会生成概率特征表示，从而能够对异质特征类型进行可加且可比较的归因。所有特征均被拼接到统一的多模态模型中，并通过夏普利值分析量化特征层面和模态层面对模型预测的贡献。

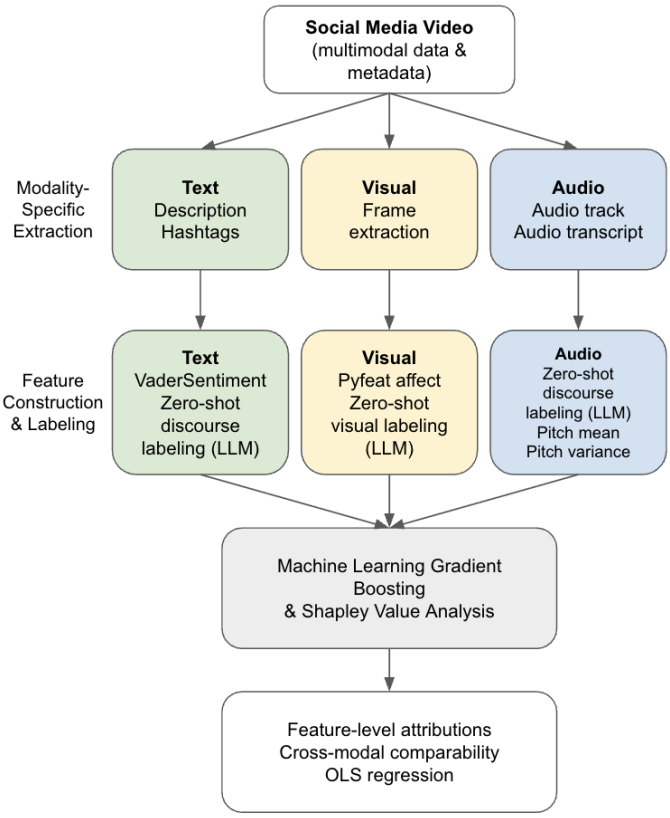


图 1：多模态分析流程概览

数据提取

我们通过 TikTok 研究 API，使用与社交焦虑障碍相关的关键词收集数据：“社交焦虑”“社交焦虑”“社交恐惧症”“社交恐惧症”“社交焦虑障碍”“社交焦虑障碍”。最终数据集包括 2020年1月1日至2025年1月1日期间在美国发布的 162,965 条帖子。对于每条帖子，API 返回了包括上传日期、视频 URL、描述文本、点赞数、评论数和观看数在内的元数据。所有相应的视频均使用 Python 模块 Pyktok 下载，以进行多模态分析。

文本模态

文本提取 我们分析了与每个视频相关的文本内容来源，包括视频描述、字幕和话题标签。所有文本内容均经过最低限度的预处理，包括转换为小写以及移除网址和非语义性伪元素；在保留非正式语言、表情符号和标点符号（如有）的同时，尽可能保留社交媒体话语所特有的情感和风格信号。

情感分析 为量化情感基调，我们采用了 VADER（效价感知词典与情感推理器），这是一种基于规则的情感分析工具，专为简短、非正式的社交媒体文本优化（Elbagir 和 Yang 2019年）。VADER 为每条文本输入生成四项评分：正向情感、负向情感、中性和复合情感。

零样本话语标注 除了表层情感倾向，我们还采用零样本分类来捕捉更高层次的交际意图。我们使用 GPT-4o-Mini，为预先定义的话语类别分配概率标签，包括自我披露、应对策略、幽默、信息支持、情感支持，以及其他与心理健康交流相关、具有理论依据的构念。我们通过结构化提示将视频描述和话题标签提交给模型，要求其为每个类别提供概率分数，而非分配离散类别。将话语标签视为连续的概率分布（0-100），使这些特征能够自然地融入下游预测模型。完整的分类提示见附录 A。

视觉模态

帧提取 对于每段视频，我们沿视频时长均匀提取五帧。选择这一采样策略，是为了在视觉内容的代表性覆盖、计算效率，以及时长略有差异的视频之间的一致性之间取得平衡。

均匀间隔采样可确保从每段视频的开头、中间和结尾捕捉视觉信息，减少对开场或高潮片段等时间上局部集中的时刻的偏向。与随机采样相比，这种方法可提供标准化的覆盖和可复现性（Brkic 等人 2025年）。与密集帧提取相比，它能降低计算成本，同时保留与视频格式、主体和环境相关的重要视觉线索；这些因素在短视频中往往相对稳定。多模态和基于视频的社交媒体分析领域的既有研究表明，少量具有代表性的视频帧足以捕捉与感知和参与度相关的高层次视觉特征（Kandhare 和 Gisselbrecht 2024年；Kennedy 等人 2025年）。每段视频提取五帧是在覆盖度和效率之间的良好折中，可支持大规模分析。

情感分析使用开源工具 PyFeat（Cheong 等人 2023年）对包含人脸的视频帧进行分析，提取动作单元（面部肌肉运动），并将情绪分为七类：快乐、悲伤、愤怒、恐惧、惊讶、厌恶和中性。每种情绪的强度值范围为 0 到 1。我们将这些情绪汇总为正性情感（快乐、惊讶、中性）和负性情感（愤怒、厌恶、恐惧、悲伤）指标。

零样本视觉标注 为系统地编码更高层次的视觉特征，我们使用结构化视觉提示将每一帧输入 GPT-4o-Mini，请求其为包括视频格式（自拍、专业、自制、模因）、镜头类型（特写镜头、宽、主观视角）、特效存在情况、可见人数以及环境背景在内的视觉属性提供概率分数。这种零样本视觉标注方法与文本标注流程相呼应，使不同模态之间能够保持一致的标注严谨性。完整的视觉分类提示见附录B。

音频模态

音频提取 我们使用 pydub 的 AudioSegment 类从视频文件中提取音轨，并将其转换为未压缩的 WAV格式。Pydub 是一种广泛使用的开源 Python 音频处理库（PyDub 开发者 2025年；FFmpeg开发者 2025年）。这些文件被用作音频特征提取和分析的输入。

我们的音频分析存在一项局限：短视频，尤其是 TikTok 上的短视频，有时会将音乐与语音叠加在一起。我们承认这是测量上的局限，因此将音频特征解读为复合信号，而不是分别孤立地分析语音或音乐。

声学特征提取 我们使用 Aubio（Brossier 2006年）提取声学特征，计算平均音高、音高变化程度以及背景音乐是否存在。这些特征捕捉了影响制作风格的副语言和音乐线索。

语音转录与标注 我们使用 OpenAI Whisper（Radford 等，2023年）生成每条音轨的转录文本。这些转录文本被纳入文本模态，并使用上文所述的相同情感和话语标注流程进行处理。

预测建模与夏普利值分析

建模框架 为了评估多模态特征如何影响注意力动态，我们采用 CatBoost（Prokhorenkova 等，2018年）构建梯度提升回归模型。梯度提升方法能够处理异质特征类型，捕捉非线性关系和高阶交互，并取得出色的预测效果（Chen 2016）。我们将观看数的对数值作为因变量，因为它反映的是初始注意力获取，而非观看后的行为（点赞、评论、分享）。所有输入特征均归一化至 0–100 的范围，以确保变量间标准化；其中，100 表示完全确信该特征存在。

我们将 CatBoost回归器设置为 1,000 次迭代、学习率为 0.1，最大树深度为 6。我们使用 80-20 的训练-测试划分训练模型，以评估预测性能。

SHAP 值分析 为了解释梯度提升模型的预测结果，我们使用 SHAP（SHapley Additive exPlanations，沙普利加性解释），根据合作博弈论中的夏普利值，将每个预测归因于输入特征（Lundberg 和 Lee（2017））。对于拟合模型 $f(\mathbf{x})$ 以及输入$\mathbf{X}$上的背景（参考）分布，SHAP 会构建一个加性解释模型

$$f(\mathbf{x}) \;=\; \mathbb{E}[f(\mathbf{X})] \;+\; \sum_{i=1}^{p} \phi_i(\mathbf{x}), \tag{1}$$

其中，$\mathbb{E}[f(\mathbf{X})]$ 是基线预测，而 $\phi_i(\mathbf{x})$是特征 i对观测 $\mathbf{x}$的贡献（以模型输出的单位表示）。夏普利值定义为某个特征在其余特征的所有子集（联盟）中的平均边际贡献：

$$\phi_i(\mathbf{x}) \;=\; \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N|-|S|-1)!}{|N|!}\,[\,v(S\cup\{i\}) - v(S)\,] \tag{2}$$

其中，$N=\{1,\dots,p\}$ 表示索引特征，$v(S)$表示仅包含子集 $S \subseteq N$ 中的特征时的模型值。

SHAP还可以进一步分解为主效应和成对交互效应。

对于每个观测$\mathbf{x}$，SHAP交互值将相对于基线预测的偏差分配为：（i）每个特征的主效应贡献；（ii）捕捉特征对如何超出各自单独效应进行联合贡献的额外项：

$$f(\mathbf{x}) \;=\; \mathbb{E}[f(\mathbf{X})] \;+\; \sum_{i=1}^{p} \phi_{ii}(\mathbf{x}) \;+\; \sum_{i<j} \phi_{ij}(\mathbf{x}). \tag{3}$$

其中，$\phi_{ii}(\mathbf{x})$表示特征i的主效应归因，$\phi_{ij}(\mathbf{x})$表示特征 $i \neq j$ 与 i和 j之间的交互归因。在不存在 i 与 j之间交互的加性模型中，$\phi_{ij}(\mathbf{x})$在各个观测中近似为零。

估计器曲面线性化

虽然 SHAP 蜂群图揭示了不同个体预测之间的异质性，但社会科学家通常需要类似于回归系数的特征重要性汇总指标。为此，我们计算特征加权 SHAP 贡献。首先，我们对底层特征进行标准化。中心化操作取决于具体领域（即使用均值或中位数）；由于我们的标注表示概率，因此选择 0.5。

因此，特征加权 SHAP 贡献可以表示为：

$$\beta_{i,SHAP} = \sum_{k=0}^{N} x_k \phi_k \tag{4}$$

这种聚合方法根据每条观测的特征幅度对其 SHAP 值加权，从而概括特征存在时对预测的影响强度。与原始 SHAP 均值不同，这种方法会考虑特征强度：只有在特征强烈存在（x_k 较高）时才持续具有较高 SHAP 值的特征，会获得相应的权重。

SHAP 分段普通最小二乘

为了刻画不同模态特征之间的阈值依赖交互，我们以中位数（概率特征取 50）为界划分特征空间，并在每个象限内分别估计线性关系。在连续域中，这对应于 SHAP 值 z 相对于变量 x 和 y 的关系：$z'' = \frac{\partial^2 z}{\partial x \partial y}$。

假设存在一个中位数阈值 x_0，使得当 x 超过该阈值时，y 的效应会出现分化：

$$\begin{cases} z'' > 0, & \text{if } x > x_0 \quad \& \quad y > y_0, \\ z'' < 0, & \text{if } x > x_0 \quad \& \quad y < y_0, \\ z'' = 0, & \text{if } x < x_0. \end{cases}$$

为了描述沿 y 轴旋转交互作用的类似情形，需反转关于 x 的条件（即，当 $x > x_0$ 时，$z'' = 0$）。第三种情形是，交互作用的方向取决于由两个变量的阈值所界定的象限，从而在不同象限之间发生符号变化：

$$\begin{cases} z'' > 0, & \text{if } xy > 0, \\ z'' < 0, & \text{if } xy < 0. \end{cases}$$

这种方法揭示了协同效应是仅在两个特征都超过阈值时才出现（非对称交互），还是会在所有象限中发挥作用（对称交互）。由于 SHAP值默认以零为中心，相关系数可直接估算各象限内交互作用的线性效应。

模态内与跨模态分析

我们使用 CatBoost 分别计算各模态内的 SHAP值，以量化各个文本、视觉和音频特征对预测观看次数的贡献。这种模态内分析方法使我们能够识别影响最大的特征，同时保留各模态的内部结构。为了进行跨模态比较并评估不同模态特征之间的交互作用，我们采用 XGBoost（Chen 2016），并基于相同的 80-20 训练-测试划分训练模型。模型最多训练 10,000 轮提升法，并根据测试集表现进行早停；每 1,000 轮评估一次，以监测收敛情况。随后，我们分析特征层面的 SHAP值，以考察不同模态特征之间的交互作用。

验证

为确保零样本分类的准确性，我们将随机抽取的 200 个视频帧（字幕、转录文本和图像）与人工标注进行对照，开展了人工验证。验证结果（附录 C）显示，字幕、转录文本和图像的总体准确率分别为 90%、86.5% 和 83.5%。具有明确行为指涉对象的类别（例如，医疗保健：100%，自我披露：100%）显示出近乎完美的一致性，而更抽象的构念则有效性较低（例如，情境压力源：65%，人工智能生成内容：60%）。为保持理论完整性，我们保留了所有特征，但在呈现结果时会谨慎解读有效性较低的特征。

结果

模态特定特征贡献

首先，我们使用 SHAP 值考察不同模态中的情感特征如何以不同方式预测曝光结果。图2a 展示了从文本（标题）、视觉（面部表情）和音频（语音转录稿）模态中提取的情感特征的 SHAP 值分布蜂群图。颜色渐变表示特征大小（高值为红色／粉色，低值为蓝色），而横向位置表示其对观看量的影响。快乐和中性情感的面部表情似乎能够预测更高的观看量，其正 SHAP 值（粉色）集中在零值右侧；相比之下，标题和转录文本的情感倾向指标则呈现出噪声较多、分布散乱的模式，在正负值范围内均有分布。

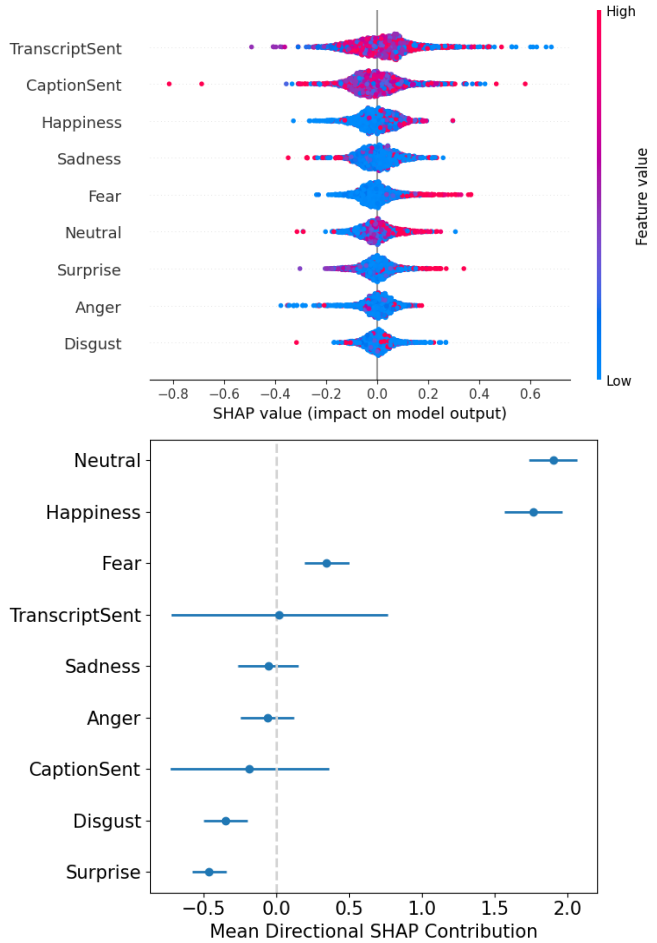


图2：a）视觉、文本和音频模态特征的 SHAP 蜂群图；b）点加权聚合 SHAP 值的森林图。

虽然蜂群图有助于探索分布模式，但其可解释性有限。例如，尽管文字转录稿和标题的情感倾向都具有较高的特征重要性，但这种重要性并不一定清晰明了。利用特征加权SHAP（公式4），便可以直接解读其方向和幅度。图2显示，与文本和语音情感倾向相比，快乐和中性情感的面部表情具有明确的正向SHAP贡献，而文本和语音情感倾向的SHAP值跨度较大。这表明，在多模态内容中，创作者的面部表情对病毒式传播性具有明确的影响（Slepian和Carr（2019）；Horstmann（2003））。

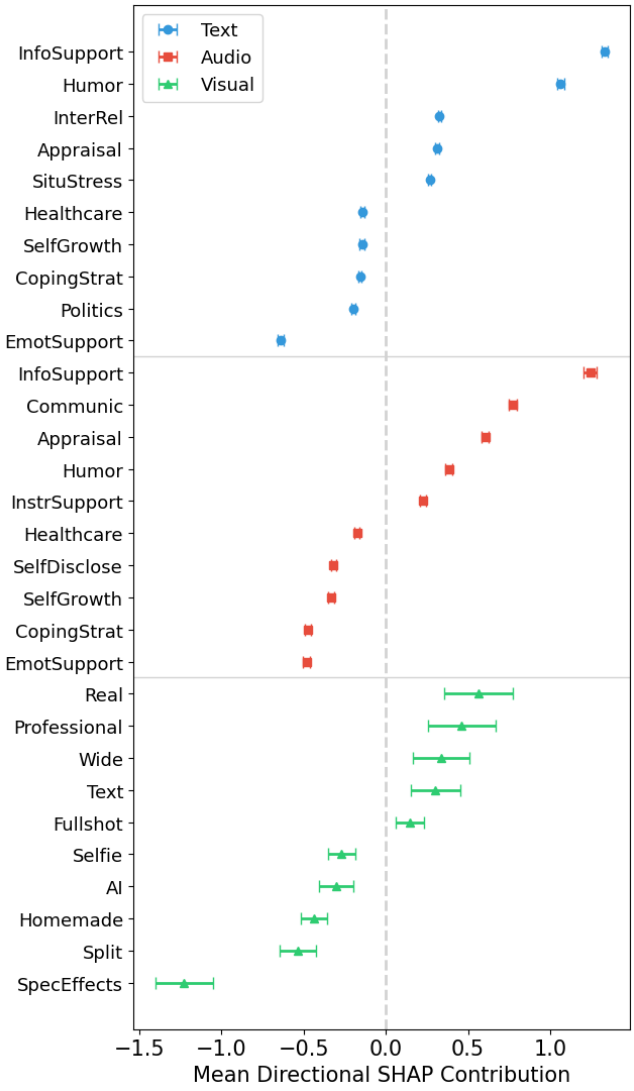


图3：文本（蓝色）、音频（红色）和视觉（绿色）特征的平均方向性 SHAP 贡献。

为确定哪些模态特定特征总体上最能预测观看量，我们计算了所有文本、视觉和音频特征的特征加权 SHAP 值。图3展示了各模态中影响力最大的十个特征（正向效应最明确的五个和负向效应最强的五个）。

文本和音频分析显示，被标记为与“信息支持”相关的帖子呈现出显著更高的正向情感 SHAP值，而被标记为

“情感支持”的帖子则呈现出较低的 SHAP 值。这表明，接触心理健康相关内容的受众更重视可操作的信息，而非情感表达。这一发现对于理解社交媒体平台如何作为非正式健康信息系统发挥作用具有重要意义。用户更偏好信息支持而非情感支持，这表明他们可能正通过这些平台主动寻求心理健康指导和实用的应对策略。这种行为与数字健康信息寻求领域的新兴研究相一致，并表明未来可以进一步研究如何审查或设计人工智能驱动的内容审核和推荐系统，在维护安全性和准确性标准的同时，支持心理健康信息的获取。视觉分析的结果印证了这些发现。

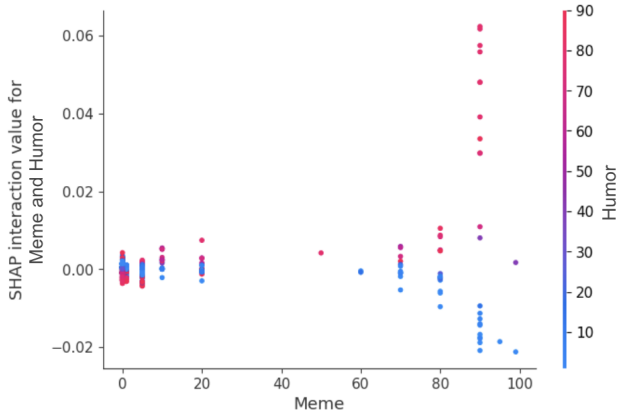
此外，呈现出“真实”（真实影像）的视频表现出最强的正向效应（平均 SHAP 值 ≈ 0.5），其次是看起来制作专业的视频、远景镜头和清晰可见的文字叠加（平均 SHAP 值 ≈ 0.3–0.4）。相反，特效带来的负向影响最为显著（平均 SHAP 值 ≈ -1.3）。分割画面、自制风格、人工智能生成内容和自拍构图均表现出轻微的负向效应。受众似乎不仅偏好信息丰富的内容，也偏好心理健康内容具有较高的感知真实性和精良的视觉制作。

跨模态交互

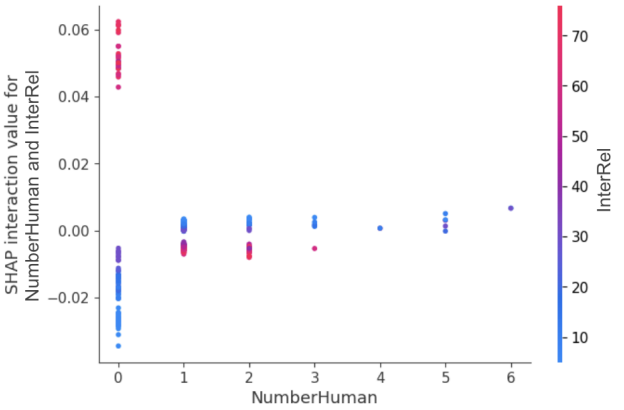
通过使用 XGBoost 建模，我们发现，视觉与文本模态之间的交互效应表明，不同程度的特征交互会协同提升受众参与度，而另一些则会降低受众参与度。

图 4a 展示了视觉模因格式与文本幽默如何相互作用并影响受众参与度。这种交互呈现出阈值效应：当模因格式的概率超过 50% 时，字幕中的幽默会显著提升受众参与度，这一点可由正向的 SHAP交互值看出。当两个特征的出现概率都很高时，这种协同效应最强。当模因概率阈值低于 50% 时，幽默对受众参与度没有明显影响，这表明模糊或不明显的模因格式无法激活受众对幽默内容的预期。这种单向模式表明，梗图形式会作为一种视觉线索，让受众倾向于对幽默内容给予积极反馈，但只有在模因格式足够容易辨认时才会如此。

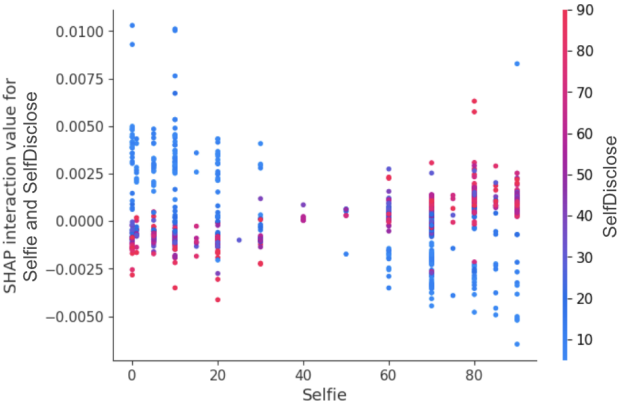
图4b展示了图像中人物数量与人际关系文本概率之间的交互作用，其效果与图4a相反。此处，差异出现在图表左侧。当视频中没有可见的人物时，讨论人际关系会增加观看量，这一点可由正向情感 SHAP交互值体现。然而，当视觉画面中出现一个或多个人物时，人际关系主题对互动度没有明显影响。这表明，对社会联结的非字面呈现可能会提升观看量。



(a) a) 迷因格式与标题幽默之间的 SHAP 交互作用



(b) b)画面中人物数量与标题中的人际关系叙事之间的 SHAP 交互作用



(c) c) 自拍视觉格式与标题中的自我披露叙事之间的 SHAP 交互作用

图4：SHAP交互图展示了跨模态特征对观看量的影响。y 轴显示 SHAP 值，x 轴显示某一特征的幅度。蓝色表示特征幅度较低，红色表示特征幅度较高。

最后，图4c展示了自拍构图与自我披露如何通过一种依赖阈值的关系相互作用，共同影响观看量。当自拍概率较低时（图表左侧），较高的自我披露程度（红色/粉色点）与负向情感的 SHAP 交互值相关，表明观看量下降。相反，当自拍概率超过约 60 时（图表右侧），较高的自我披露程度（红色/粉色点）对应的 SHAP 交互值转为正向情感，表明观看量上升。这一模式表明，自我披露叙事需要视觉自我呈现，才能有效提升观看量。个人故事讲述需要通过视觉自我暴露来传递真实性并建立虚拟联结。

与图2a中的蜂群图类似，SHAP交互图难以进行精确解读。为了在数学上建模跨模态交互如何随特征强度变化，我们以特征中位数为阈值，对四个象限中的SHAP交互值进行了分段回归分析。图5展示了三个交互对的回归系数（β值）。

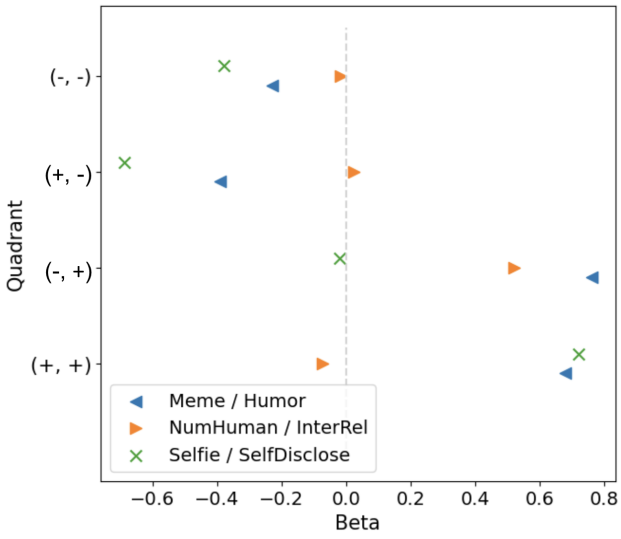


图5：SHAP交互值在各象限中的分段回归系数。每个交互对均以不同的标记表示。四个象限代表特征值的不同组合：(-, -)表示两个特征均低于中位数阈值；(+, +)表示两个特征均高于阈值；(-, +)和(+, -)表示混合情况，即一个特征高于阈值，另一个低于阈值。

模因/幽默交互（蓝色三角形）在（+, +）和（-, +）象限中呈现出明显的协同效应。这表明，文本中幽默出现的概率越高，视频的观看量也越高，而这一关系不受模因格式影响。（-, -）和（+, -）象限则显示观看量适度下降。人类数量/人际关系交互（橙色三角形）呈现出不同模式：如前所述，当没有人物但存在关系话语时（-, + 象限：$\beta \approx 0.3$），会产生正向影响；其他所有象限对观看量几乎没有影响。这种不对称性表明，受众可能更容易与抽象关系表征互动——例如文字截图、空白空间和隐喻性图像——从而将自身经历投射到含义模糊的视觉内容上。自拍/自我披露交互（绿色 ×）表现出最显著的依赖关系：当两种特征均不存在时，影响强烈为负（-, -象限：$\beta \approx -0.6$）；当出现自我披露但没有视觉自我呈现时，影响为负（-, + 象限：$\beta \approx -0.3$）；而当两者同时存在时，影响则显著为正（+, + 象限：$\beta \approx 0.7$）。

分段回归方法揭示了线性建模可能掩盖的交互结构，从而更细致地理解多模态特征如何塑造注意力动态。

单模态测量在多模态语境中并不充分

最后，我们构建了一个示例，说明单模态测量缺乏细致性。现有社会科学研究普遍假设，粗略特征——通常通过文本词汇测量——能够作为可靠且有效的代理指标，用于推断注意力动态。我们使用情感倾向这一社交媒体参与度的常见预测变量，对这一假设提出质疑，并表明同一内容中的情感测量会因模态不同而显著不同，而且文本情感无法捕捉不同领域中负向情感的细微差异。

图6展示了面部情绪（由 PyFeat 测量）与帖子说明文字和文字转录稿中的文本情感倾向之间的对比。由于面部情绪以概率表示，图中的点代表情绪概率与情感倾向相乘后的平均值（Wei、Noh 和 Chang，2025年）。

图6显示，文本情感倾向、语音情感与视觉情感之间的对应趋势并不一致。总体而言，字幕的情感表现明显更平淡、更含蓄，其加权平均分数在所有情绪类别中都集中于较窄的低分区间（≈0.05–0.10）。相比之下，文字转录稿呈现出显著更高且差异更明显的情感权重（≈ 0.17–0.29），表明口语内容中的情绪表达更强。因此，仅依据字幕进行情感倾向分析，可能会系统性地低估情绪强度和情绪多样性，而基于文字转录稿的分析则不会。字幕和文字转录稿情感倾向分析呈现的唯一一致模式是：在书面和口语模态中，悲伤对应的情感倾向分数都是最低的。

此外，将我们的发现与另一种话语领域进行比较，揭示了情境差异。有趣的是，近期一项政治传播研究发现，面部愤怒情绪与最负面的文本情感倾向相关（Wei、Noh和Chang（2025））。在心理健康领域，悲伤似乎与负面情感倾向的相关性最高。正如列夫·托尔斯泰在《安娜·卡列尼娜》（1877）中所写的那句名言：“幸福的家庭都是相似的；不幸的家庭各有各的不幸。”同样，.....的表现

负面情绪可能因领域而异，例如，政治领域表现为愤怒，心理健康领域则表现为悲伤。

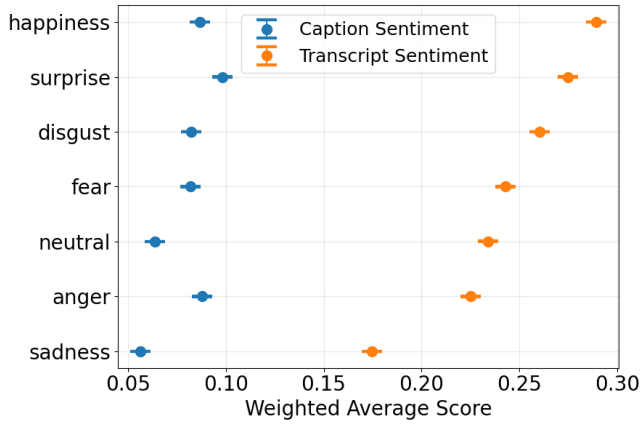


图6：面部情绪得分与标题及转录文本情感倾向的比较（含95%置信区间）

讨论

方法论贡献

短视频平台上的多模态传播给计算社会科学家带来了一个根本性挑战：如何在保留可解释性的同时，系统地大规模提取、整合并解读文本、视觉和音频模态中的异质特征。本研究通过提出一个方法框架来应对这一挑战，该框架将自动化概率特征提取与基于夏普利值的归因相结合，使研究人员能够在统一的解释框架中量化个人特征的贡献以及跨模态交互。

我们的方法带来了三项相互整合的方法论进展，共同将不透明的多模态内容转化为可解释、以理论为基础的变量。首先，零样本概率分类无需微调即可实现理论驱动的特征提取，并将话语类别和视觉属性表示为连续的概率分布（0-100），而非离散标签。这既保留了测量不确定性这一理论核心价值，也生成了可直接跨模态比较的特征。

其次，我们通过特征加权聚合（β_{SHAP}）扩展了SHAP 的可解释性，将点级归因汇总为类似回归系数的数值，社会科学家可据此清晰解读影响方向，同时兼顾梯度提升树的非线性复杂性。

第三，通过对SHAP交互值进行分段回归，在中位数阈值处划分特征空间，并估算各象限内的独立效应，从而揭示依赖阈值的跨模态协同效应。这使我们能够明确哪些情况下模态必须协同，哪些情况下则可独立运作。

将这一框架应用于162,965条关于社交焦虑障碍的TikTok视频，展现了其经验价值。我们发现，不同模态的情感倾向存在系统性差异：面部表情能够预测互动，而文本情感则产生嘈杂信号——这表明，单模态代理指标会引入系统性偏差。跨模态交互呈现出明显的阈值效应：只有当模因形式和幽默都强烈存在时，二者才会产生协同效应；自拍构图则会增强自我披露。这些依赖阈值的模式，既无法通过大规模人工编码检测，也无法通过缺乏可解释归因能力的学习嵌入发现。

该框架凭借其模块化架构，可推广至心理健康传播以外的领域。研究人员可以替换为适用于特定领域的分类方案（如政治框架、虚假信息指标），同时保持相同的分析流程。

心理健康传播

除方法论贡献外，我们的研究结果还揭示了心理健康内容在算法中介环境中的运作方式所呈现的若干实质性模式，并对健康传播研究和平台设计均有所启示。

在情感分析中，面部表情比文本情感倾向更为重要，这表明视觉自我呈现比文本更能有效传递真实性。正向情感的面部表情，尤其是快乐和中性情感，与更高的.....相关；这或许是因为它们向接触敏感内容的受众传递了平易近人和情绪安全的信号。这与有关非言语传播的研究一致，该研究指出，面部是传递情感真实性的优先渠道（Slepian和Carr（2019））。这也对同伴支持干预和心理健康倡导活动有所启示，表明在多模态语境中，视觉上的亲和力或许比文本更能吸引受众。

用户更偏好信息支持而非情感支持，这表明TikTok 的作用更像是非正式的健康信息系统，而非传统的支持小组。标注为提供信息支持的内容，其观看量高于提供情感肯定或共情的内容。这一模式表明，用户主动寻求可付诸行动的指导，而非情感联结，并将短视频平台视为提升心理健康素养的教育资源。这种行为对于理解数字健康信息寻求具有重要意义，也引发了有关信息质量、准确性和潜在危害的问题。如果受众优先选择来自同伴创作者的信息内容，而非情感支持，平台就应考虑推荐算法是否会在无意间放大健康错误信息或不合格建议，同时压低经专业审核的资源的传播。

局限性与未来工作

我们的研究存在三项局限，提示了方法上的改进方向。首先，TikTok 将音乐与语音叠加的做法会干扰音频归因，因为我们的处理流程

无法分离这些声源。因此，声学特征捕捉到的是复合信号，而非分别提取语音或音乐；采用音频源分离技术可以实现更精确的归因。其次，我们聚焦于 TikTok，这限制了研究的泛化能力，因为 Instagram Reels 和 YouTube Shorts 等平台的算法推荐系统、用户人口统计特征和界面可供性各不相同。跨平台比较研究有助于区分普遍适用的多模态原则与平台特有的影响。第三，将视频视为静态快照，忽略了内容中的多模态构成如何随时间演变。未来研究若能纳入时间特征提取，便可揭示视频播放过程中叙事动态和注意力保持模式的变化。

该流程的模块化架构有助于实现这些扩展，同时支持向其他领域进行适配。研究人员可以替换分类方案（政治框架、虚假信息指标）、结果变量（分享、评论、转化）和情境特征（创作者人口统计特征、发帖时间），以探讨政治传播、公共卫生宣传活动或商业内容中的相关问题。正如负面情感与面部情感的呈现存在差异一样，跨模态交互的动态也可能产生富有成效且具有领域针对性的结果。

结论

随着计算社会科学日益应对来自意义建构以多模态方式运作的平台的丰富非结构化数据，此处开发的方法工具将成为必不可少的基础设施。本研究提出了一个框架，将自动化概率特征提取与基于夏普利值的归因相结合，使研究人员能够量化个人特征的贡献和跨模态交互，同时保留可解释性。我们的贡献与希望在于，这一框架有助于构建计算社会科学所需的必要条件：大规模分析、坚实的理论基础，以及足以支持社会科学推断的可解释性。

参考文献

Ali, A.; 和 Molla, D. 2025年。《多模态文本摘要系统性文献综述》。*ACM ComputingSurveys*，58(3): 1–38。

Andalibi, N.; Ozturk, P.; 和 Forte, A. 2017年。《Instagram 上的敏感自我披露、回应与社会支持：以#抑郁症为例》。载于*2017*年美国计算机协会计算机支持的协同工作与社会计算会议论文集，1485–1500。Bazarova, N. N.; 和 Choi, Y. H. 2014年。《社交媒体中的自我披露：将功能方法拓展至社交网站上的披露动机与特征》。传播学杂志，64(4):635–657。

Bickham, C.; Kazemi-Nia, K.; Luceri, L.; Lerman, K.; 和 Ferrara, E. 2024年。《隐于众目之下：探索心理健康、饮食失调与 TikTok 内容审核的交集》。载于*ICWSM* 最脆弱群体福祉数据研讨会论文集。

Brkic, M.; Razzouki, A. F.; Tevissen, Y.; Guetari, K.; and Yacoubi, M. A. E. 2025年. 帧采样策略至关重要：小型视觉语言模型的基准测试。*arXiv*预印本 *arXiv:2509.14769*. Brossier, P. M. 2006年. 面向交互式应用的音乐音频自动标注. 博士论文. Chen, H.; Bisbee, J.; Tucker, J. A.; and Nagler, J. 2025年. 为社交媒体帖子添加标签：向编码员展示多模态内容，能否带来更好的人工标注和更优的机器分类器？《政治科学研究与方法》, 1–13. Chen, T. 2016年. XGBoost：可扩展的树提升系统。康奈尔大学. Cheong, J. H.; Jolly, E.; Xie, T.; Byrne, S.; Kenney, M.; and Chang, L. J. 2023年. Py-feat：Python面部表情分析工具箱。《情感科学》, 4(4): 781–796. Cutrona, C. E.; and Suhr, J. A. 1992年. 压力事件的可控性与对配偶支持行为的满意度。《传播研究》, 19(2): 154–174. Daft, R. L.; and Lengel, R. H. 1986年. 组织的信息需求、媒介丰富度与结构设计。《管理科学》, 32(5): 554–571. Elbagir, S.; and Yang, J. 2019年. 使用自然语言工具包和VADER情感分析进行Twitter情感分析。载于国际工程师与计算机科学家多学科会议论文集, 第122卷, 13–15. 国际工程师协会. FFmpeg开发者. 2025年. FFmpeg：用于录制、转换和串流音频与视频的完整跨平台解决方案。https://www.ffmpeg.org/. FFmpeg开源多媒体框架的官方网站。Hong, L.; Doumith, A. S.; and Davison, B. D. 2013年. 协同因子分解机：对用户兴趣建模并预测Twitter上的个人决策。载于第六届*ACM*网络搜索与数据挖掘国际会议论文集, 557–566. Horstmann, G. 2003年. 面部表情传达的是什么：情绪状态、行为意图，还是行动请求？《情绪》, 3(2): 150. Huang, Q.; Lei, S.; and Chen, Z. 2025年. 短视频应用中的拟社会互动与问题性使用：揭示其中介机制。《心理学前沿》, 16: 1584685. Islam, M. M.; Kakouros, S.; Heikkil¨a, J.; and Oussalah, M. 2025年. 迈向视频摘要的自动化多模态方法：构建文本、音频与基于面部线索的摘要之间的桥梁。*arXiv*预印本 *arXiv:2506.23714*. Kandhare, M.; and Gisselbrecht, T. 2024年. 多模态RAG检索的视频帧采样方法实证比较。*arXiv*预印本 *arXiv:2408.03340*. Kennedy, C. J.; Gajjar, B.; Chang, H.-C. H.; Unger, J. B.; and Vassey, J. 2025年. 电子烟社交媒体营销的人口统计趋势：通过计算机视觉感知性别呈现与面部年龄。《尼古丁与烟草研究》, ntaf057.

Kress, G. R.; and Van Leeuwen, T. 2001年. 多模态话语：当代传播的模式与媒介。（无标题）. Li, S.; Wu, S.; Liu, T.; Zhang, H.; Guo, Q.; and Peng, Z. 2024年. 理解图文帖子及其在在线哀伤支持社区中获得的社会支持。载于国际*AAAI*网络与社交媒体会议论文集, 第18卷, 917–929. Liu, J.; Su, Y.; Seth, P.; et al. 2025年. 大型语言模型能否掌握视觉内容中的概念？一项关于抑郁症主题YouTube Shorts的案例研究。*arXiv*预印本 *arXiv:2503.05109*. Lundberg, S. M.; and Lee, S.-I. 2017年. 一种解释模型预测的统一方法。神经信息处理系统进展, 30. Meghawat, M.; Yadav, S.; Mahata, D.; Yin, Y.; Shah, R. R.; and Zimmermann, R. 2018年. 一种预测社交媒体人气的多模态方法。载于*2018*年*IEEE*多媒体信息处理与检索会议（*MIPR*）, 190–195. IEEE. Milton, A.; Ajmani, L.; DeVito, M. A.; and Chancellor, S. 2023年. “我在这里看到了自己”：TikTok上的心理健康内容、社群与算法策划。载于*2023*年*CHI*计算机系统中的人为因素会议论文集, 1–19. Niu, S.; Bartolome, A.; Mai, C.; and Ha, N. B. 2021年. #宅家#与我同行：YouTube创作者如何帮助缓解COVID-19带来的孤独感？载于*2021*年*CHI*计算机系统中的人为因素会议论文集, 1–15. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A. V.; and Gulin, A. 2018年. CatBoost：基于类别特征的无偏提升法。神经信息处理系统进展, 31. PyDub开发者. 2025年. PyDub：通过简单易用的高级接口处理音频。https://www.pydub.com/. PyDub Python音频处理库的官方网站和文档。Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023年. 通过大规模弱监督实现稳健的语音识别。载于国际机器学习会议, 28492–28518. PMLR. Salih, A. M.; Raisi-Estabragh, Z.; Galazzo, I. B.; Radeva, P.; Petersen, S. E.; Lekadir, K.; and Menegaz, G. 2025年. 关于可解释人工智能方法的观点：SHAP与LIME。《先进智能系统》, 7(1): 2400304. Slepian, M. L.; and Carr, E. W. 2019年. 真实感的面部表情：情绪变化性会增强对可信度和领导力的判断。认知, 183: 82–98. Trzcinski, T.; and Rokita, P. 2017年. 使用支持向量回归预测在线视频的人气。《*IEEE*多媒体汇刊》, 19(11): 2561–2570. Vilares, D.; Alonso, M. A.; and G´omez-Rodr´ıguez, C. 2015年. 词汇与句法处理在Twitter消息极性分类中的效用。《信息科学与技术协会期刊》, 66(9): 1799–1816. Violot, C.; Elmas, T.; Bilogrevic, I.; and Humbert, M. 2024年。YouTube 短视频与常规视频：用户参与度和内容创作趋势的比较分析。载于第*16*届 *ACM Web Science Conference* 论文集，213–223。Wang, X.; Chen, G.; Qian, G.; Gao, P.; Wei, X.-Y.; Wang, Y.; Tian, Y.; and Gao, W. 2023年。大规模多模态预训练模型：全面综述。*Machine Intelligence Research*，20(4)：447–482。Wang, Y. 2021年。多模态分析：研究短视频与戏剧实践。第*104*届新闻与大众传播教育协会年会论文集，1：1–19。Wei, C.; Noh, S.; and Chang, H.-C. H. 2025年。面孔胜于言辞：2024年美国总统选举中的情绪与文本情感。载于 *ACM Web Conference 2025* 伴随论文集，1390–1393。Winter, E. 2002年。沙普利值。博弈论及其经济学应用手册，3：2025–2054。Wu, B.; Cheng, W.-H.; Zhang, Y.; Huang, Q.; Li, J.; and Mei, T. 2017年。利用深度时间上下文网络对社交媒体热度进行序列预测。*arXiv* 预印本 *arXiv:1712.04443*。Wu, P.; Zou, S.; Chen, C.; and Song, Y. 2025年。污名化的温床还是支持的来源：抖音心理健康相关视频的多模态分析。*Computersin Human Behavior*，108716。Xie, X.; Lin, Y.; and Bai, Q. 2025年。拟社会关系对短视频平台用户绝望感的恢复功能：一项调节中介研究。*BMC PublicHealth*，25(1)：3668。Yazdavar, A. H.; Mahdavinejad, M. S.; Bajaj, G.; Romine, W.; Sheth, A.; Monadjemi, A. H.; Thirunarayan, K.; Med- dar, J. M.; Myers, A.; Pathak, J.; 等。2020年。社交媒体中的多模态心理健康分析。*PLOS ONE*，15(4)：e0226248。

附录

文本/转录分类提示

系统提示

你是一名公共卫生研究人员，正在分析社交媒体帖子文本中的心理健康话语。你的任务是评估以下文本：以下格式说明极其重要，请务必严格遵守：请以百分比可能性表示概率（即，如果极不可能则为1%，如果极有可能则为99%）。请根据以下问题输出一个 JSON 对象。仅返回结构化 JSON，不要添加任何其他文字、描述或解释。心理健康主题请给出文本内容与应对策略相关的概率。请给出文本内容与社交互动中的沟通相关的概率。请给出文本内容与人际关系相关的概率。请给出文本内容与自我成长相关的概率。请给出文本内容与情境压力源相关的概率。请给出文本内容包含幽默用法的概率。社会支持类型请给出文本内容与情感支持（例如，表达同理心、爱、信任和关怀）相关的概率。请给出文本内容与工具性支持（例如，提供有形援助和服务）相关的概率。请给出文本内容与信息支持（例如，建议、提议和信息）相关的概率。请给出文本内容与评价支持（例如，有助于自我评价的信息）相关的概率。根据内容判断文本是在寻求信息、提供信息、两者皆是、两者皆非还是不明确。请将结果作为字符串返回：‘”seeking”‘、‘”providing”‘、‘”both”‘、‘”neither”‘或‘”unclear”‘。内容主题请给出文本内容与政治议题相关的概率。还请给出文本内容与医疗保健相关的概率。还请给出文本内容与外交事务（例如外交政策、战争、全球领导人）相关的概率。还请给出文本内容与经济议题相关的概率。还请给出文本内容与气候议题相关的概率。还请给出文本内容与 COVID-19 相关的概率。还请给出概

文本内容与移民相关的概率。请同时提供文本内容与犯罪或执法相关的概率。请同时提供文本内容与技术相关的概率。请同时提供文本内容与大语言模型或人工智能聊天机器人相关的概率。如果提及大语言模型或人工智能聊天机器人，还请将其立场描述为“正向情感”、“中性”、“负向”或“不明确”。请同时提供文本内容涉及倡导的概率，例如宣传某项事业，或呼吁社会或政治变革。请同时提供文本内容旨在提高意识的概率，例如向他人普及某个具体议题或状况。请同时提供文本内容包含行动呼吁的概率，例如鼓励受众采取某个具体步骤（如捐款、主动联系或寻求帮助）。请同时提供文本内容通过个人叙事进行自我披露的概率，例如分享个人经历或心理健康历程。如果文本提及任何心理健康状况（如抑郁症、焦虑症、创伤后应激障碍），请将所有提及的状况列为字符串。如果文本提及其他社交媒体个人资料（如 @username），请列出所有提及的用户名。如果文本提及 TikTok 以外的社交媒体平台（如 Instagram、Twitter），请列出所有提及的平台。

视觉分类提示

系统提示

您是一名公共卫生研究人员，正在分析社交媒体帖子视觉内容中的心理健康话语。您的任务是评估以下图像。以下格式说明极其重要，请务必严格遵守：如果图像中包含任何文本，请将其全部提取为一个字符串。请以百分比可能性表示概率（即，非常不可能时为 1%，极有可能时为 99%）。您需要针对以下问题输出一个 JSON 对象。只返回结构化 JSON，不要添加任何其他文字、描述或解释。基本识别 用不超过 200 个字描述该图像。以字符串形式返回。说明该画面最有可能是真实照片还是人工智能生成。请提供该画面包含文本的概率（标记为 Text:）。请提供该画面使用特

殊效果的概率（标记为 SpecialEffects:）。

文本内容：
如果画面中包含任何文本，请将其全部提取为一个字符串（标记为 TextContent:）。
如果包含文本，请指出文本使用的语言（标记为 Language:）。如果包含多种语言，请将它们分别列为多个字段。如果画面中包含任何话题标签，请将其全部提取为一个字符串（标记为 Hashtags:）。
如果画面中包含任何地点，请将其全部提取为一个字符串（标记为 Location:）。
如果画面中包含任何日期，请将其全部提取为一个字符串（标记为 Date:）。
如果画面中包含任何链接或二维码，请将链接全部提取为一个字符串（标记为 Link:）。
如果画面中包含任何歌曲名称或音乐，请将其全部提取为一个字符串（标记为Music:）。视觉内容

如果图像中包含个人，该人物直接与哪些物体互动？将所有物体以字符串形式列出。如果有多个个人，请全部列出（标记为RelevantObjects:）。
你能从画面中识别出任何个人或公众人物吗？如果可以，他们是谁？（标记为KnownIndivid- uals:）

场景所在的地点是哪里（例如卧室、汽车、公园、办公室等）（标记为Location:）？
请提供视频镜头风格为特写镜头的概率（标记为 CloseUp:）。请提供视频镜头风格为全景镜头的概率（标记为FullShot:）。请提供视频镜头风格为双人镜头的概率（标记为Two Shot:）。请提供视频镜头风格为主观视角的概率（标记为POV:）。请提供视频镜头风格为宽的概率（标记为Wide:）。请提供画面被分割成多个分割区域的概率（标记为Split:）。画面被分割成多少个分割区域？（标记为Segments:）。

分割区域是如何划分的？垂直、水平还是其他方式？（标注为：方向）。
请提供该录制内容属于自拍的概率，即画面非常近，且明显为手持拍摄（标注为：自拍）

请提供该录制内容属于自制的概率，即非专业拍摄，且明显不是手持拍摄，例如使用三脚架拍摄的素材（标注为：自制）。
请提供该录制内容属于专业拍摄的概率（标注为：专业）。

请提供该画面描绘模因的概率（标注为：模因）？

这则模因来自哪部卡通或动漫作品（《恶搞之家》、《辛普森一家》等）？（标记为 MemeSource:）请提供该画面是绘画作品/插画的概率。

（标记为 Illustration:）请提供该画面中有人跳舞的概率（标记为 Dance:）。请提供该画面中有人唱歌的概率（标记为 Sing:）。

零样本标注验证

表 1：按类别划分的字幕标注有效率

类别	有效性
评价支持	75%
应对策略	90%
情感支持	90%
医疗保健	100%
幽默	95%
信息支持	90%
人际关系	100%
政治议题	100%
自我成长	95%
情境压力源	65%
总体	**90%**

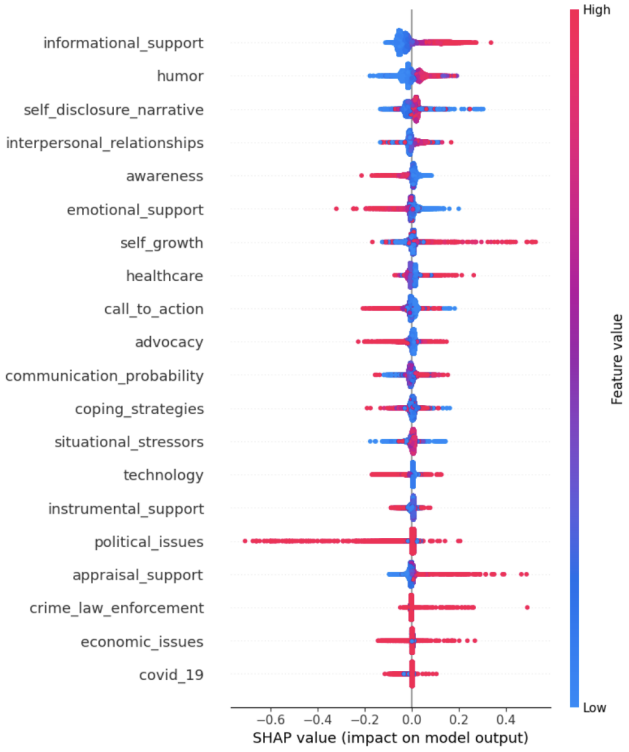
表 2：按类别划分的文字转录稿标注有效率

类别	有效性
评价支持	85%
沟通概率	75%
应对策略	90%
情感支持	70%
医疗保健	90%
幽默	85%
信息支持	95%
工具性支持	85%
自我披露叙事	100%
自我成长	90%
总体	**86.5%**

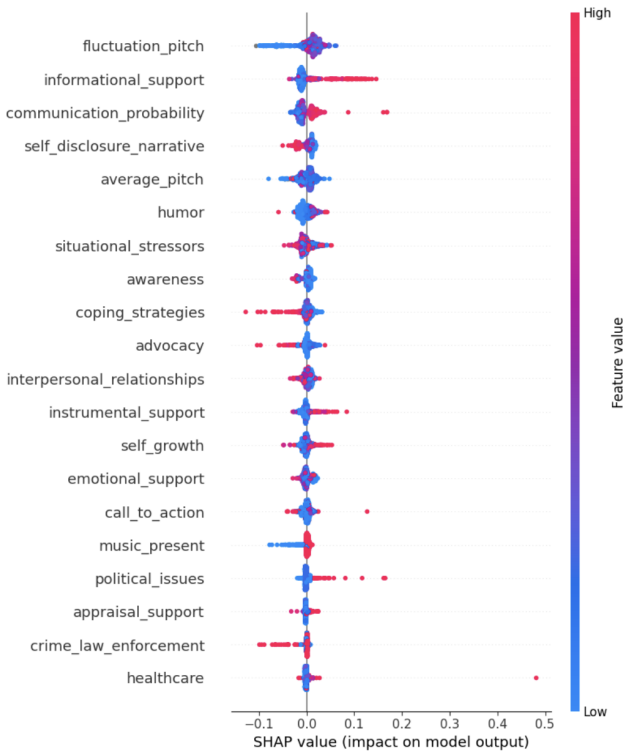
表3：按类别划分的图像标注有效性百分比

类别	有效性
AI	60%
全景镜头	75%
自制	100%
专业	70%
Real	100%
自拍	75%
特效	75%
分割画面	100%
Text	85%
Wide	95%
总体	**83.5%**

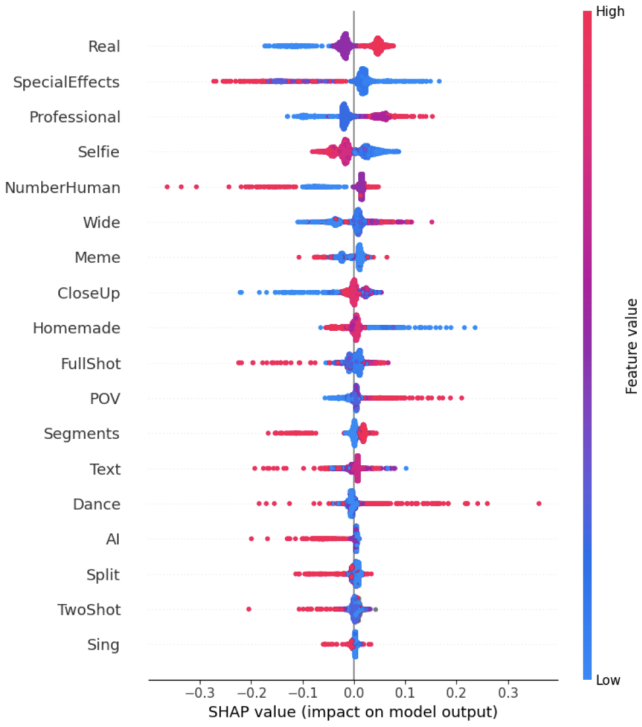
完整 SHAP 蜂群图



(a) 概率文本特征的 SHAP 值分布。



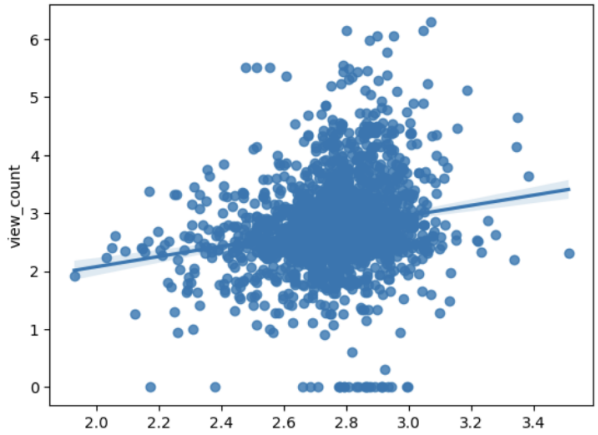
(b) 概率听觉特征的 SHAP 值分布。



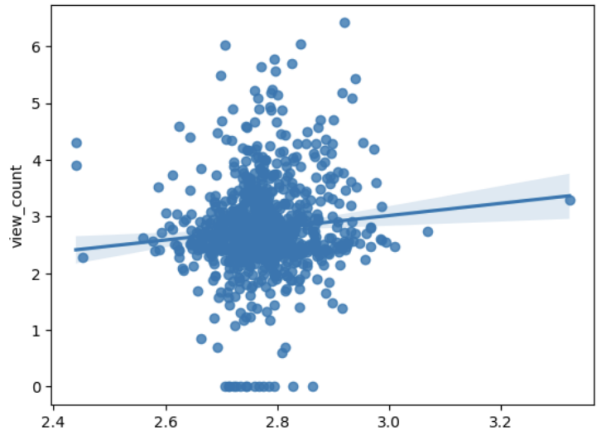
(c) 概率视觉特征的 SHAP 值分布。

图 7：文本、视觉和听觉特征的 SHAP 总结图。颜色表示特征幅度，其中蓝色表示幅度较低，红色表示幅度较高。

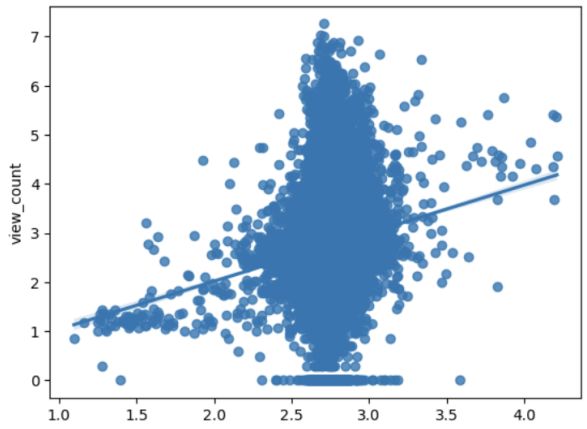
回归图



(a) 散点图及拟合回归线，展示使用文本特征预测的观看量（x 轴）与实际观看次数（y 轴）之间的关系。



(b) 散点图及拟合回归线，展示使用视觉特征预测的观看量（x 轴）与实际观看次数（y 轴）之间的关系。



(c) 散点图及拟合回归线，展示使用听觉特征预测的观看量（x 轴）与实际观看次数（y 轴）之间的关系。

图 8：文本、视觉和听觉特征的回归图。