arXiv:2602.09637v1 [cs.CV] 10 Feb 2026

迈向基于大型语言模型的免训练多模态仇恨定位 *

Yueming Sun
混合智能实验室，杜伦大学 多模态智能实验室，
埃克塞特大学 yueming.sun@durham.ac.uk

Long Yang
混合智能实验室 杜伦大学
yang.long@durham.ac.uk

Jianbo Jiao The
MIx Group 伯明翰大学
j.jiao@bham.ac.uk

Zeyu Fu
多模态智能实验室 埃克塞特
大学 z.fu@exeter.ac.uk

摘要

在线视频中仇恨内容的泛滥对个人福祉和社会和谐构成了严重威胁。然而，现有的视频仇恨检测方案要么高度依赖大规模人工标注，要么缺乏细粒度的时间精度。在本文中，我们提出了**LELA**，这是首个基于免训练**大**语言模型（LLM）的仇**恨**视频**定**位框架。与依赖监督式流程的最先进模型不同，LELA 利用大语言模型和模态专属字幕生成，以免训练的方式检测并在时间上定位仇恨内容。我们将视频分解为五种模态，包括图像、语音、光学字符识别、音乐和视频上下文，并采用多阶段提示方案，为每个视频帧计算细粒度的仇恨分数。我们进一步引入组合匹配机制，以增强跨模态推理。在两个具有挑战性的基准测试 HateMM 和 MultiHateClip 上进行的实验表明，LELA 以较大优势超越了所有现有的免训练基线方法。我们还提供了大量消融实验和定性可视化结果，确立了 LELA 作为可扩展且具备可解释性的仇恨视频定位坚实基础的地位。

免责声明： 本文讨论仇恨或冒犯性内容，这些内容可能会令部分读者感到不适。请读者谨慎阅读。

1 引言

尽管人们已投入大量精力检测文本仇恨言论 [1, 2], ，仇恨视频检测仍相对缺乏研究。与文本不同，视频包含多种同步模态，包括视觉内容、语音、音乐和屏幕文字，这些内容共同作用，能够以隐含且依赖上下文的方式传达仇恨 [3]。对此类内容进行人工审核既耗时，又会造成心理伤害，且最终难以扩大规模，因此有必要开发自动化的多模态仇恨检测框架，以营造更安全的数字环境。

近期的基准测试（如 HateMM [4] 和 MultiHateClip [5]）通过提供涵盖多种模态的片段级标注，取得了重要进展。HateMM 包含具有视觉、听觉和文本元素的视频，而 MultiHateClip 则由短视频组成，这些视频均以英语和中文标注了仇恨性、冒犯性语言、受害者指向和模态归因。然而，在这些数据集上训练的现有方法 [5, 4, 6, 7] 仍主要依赖视频级或片段级标签，忽略了更精细的时间粒度。如图1所示，这类视频级分类系统本质上如同黑箱：模型只输出一个标签，却没有说明为何或在视频中的何处做出这一决策，因此难以确定模型的推理是否真正符合人类的理解。这促使我们探索视频仇恨定位，明确地

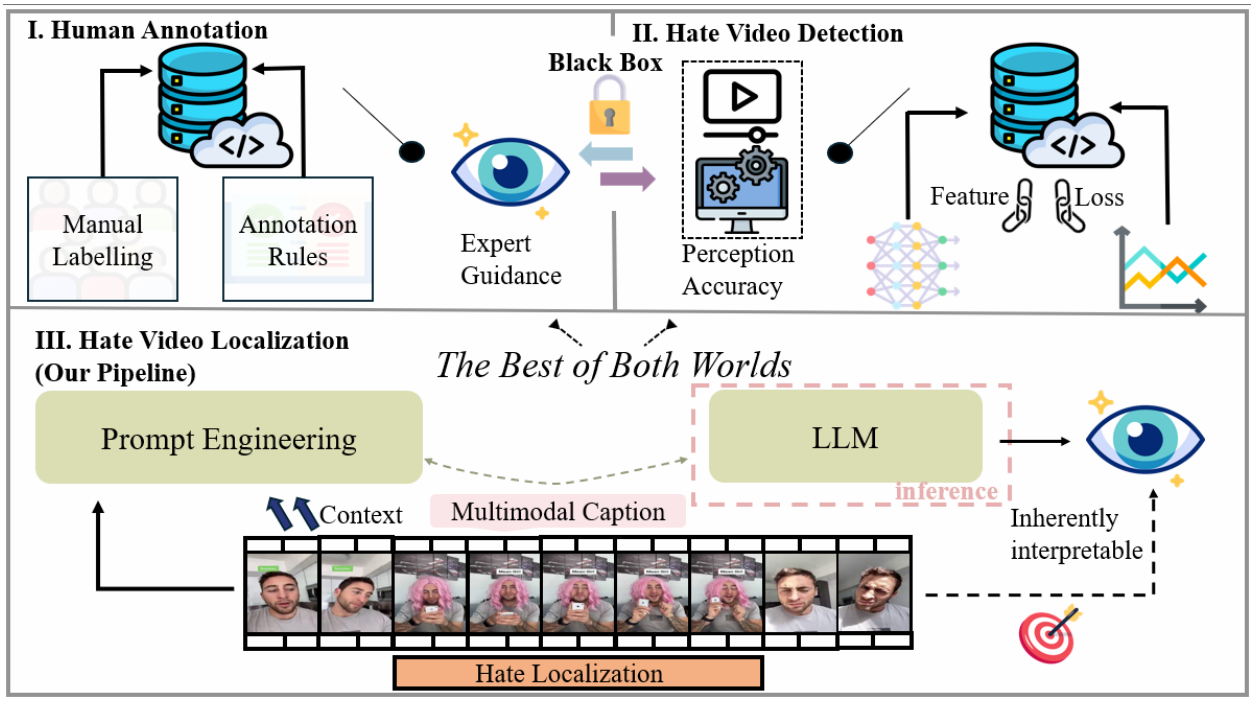

图1：我们提出了首个基于大语言模型的视频仇恨定位框架，该框架应对仇恨内容审核中的可解释性挑战，并从多模态视频输入中定位帧级仇恨内容。

针对仇恨内容进行帧级理解。我们的方法采用基于大语言模型的免训练范式，生成逐帧预测，从而提高模型决策的可解释性，并能超越粗粒度的视频级准确性，更全面地评估仇恨内容理解。

与作为早期任务且已有许多强基线方法的视频异常检测（视频异常检测）不同， [8, 9, 10], 视频恶意内容检测本质上更具语义性，且对上下文更为敏感。传统视频异常检测主要关注视觉（偶尔也包括文本）模式，并分析某些视频帧是否偏离整体视频分布。相比之下，视频恶意内容检测依赖于协同的多模态信号，并需要依据有害或冒犯性行为的现实世界知识做出片段级判断。如图2所示，视频异常检测主要旨在识别原本正常视频中的短暂异常片段，而视频仇恨定位则在多种模态上运作，关注与仇恨内容相关的高层线索，并独立地对每个视频帧进行排序或赋予标签。

总之，在本文中，我们引入了仇恨视频定位任务，其目标不仅是在视频层面预测一段视频是否具有仇恨性，还要为各个视频帧或短片段赋予可解释的仇恨分数。给定由视频帧、音频和自动生成的字幕组成的多模态视频，模型必须生成时间分辨的仇恨程度概况，以突出导致最终决策的精确区域。这种帧级或片段级定位自然地支持人在回路审核、申诉程序中的证据标注以及模型审计，也可作为未来监督模型的细粒度监督信号。

基于这些洞见，本文的贡献主要有三点：

- 我们提出首个基于大语言模型的视频仇恨内容时间定位框架，无需依赖监督标注或进行任务专用模型调优，因此能够高度适应现实场景。
- 我们提出一种多阶段提示机制和组合匹配策略，利用特定模态的字幕，并发挥大语言模型的推理能力，以实现细粒度仇恨评分，同时提升可解释性。
- 我们在 HateMM 和 MHC-English 数据集上进行了大量实验，首次全面评估了大语言模型在多种模态、语言和文化语境下检测并定位仇恨内容的能力。

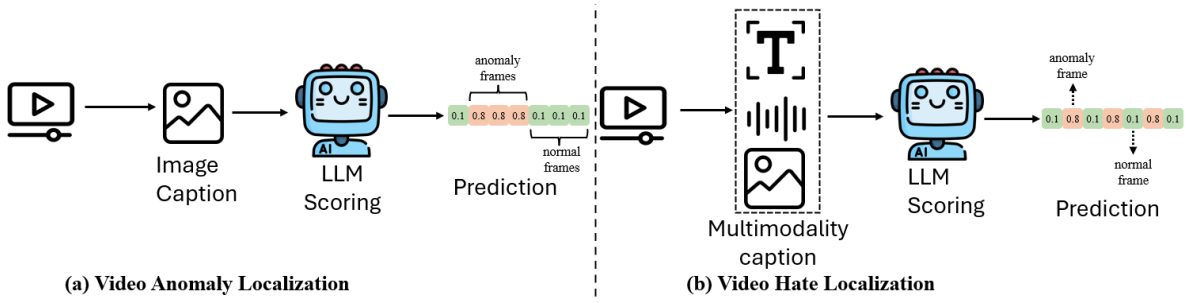

图 2：传统的视频异常检测（a）侧重于视觉信号，旨在从整体正常的视频中检测出短暂的异常片段，通常生成粗略的视频级异常分数。视频仇恨定位（b）则联合利用多种模态（图像、音频和文本），关注与仇恨或冒犯性内容相关的高层语义线索，并分配细粒度的帧级分数，从而能够精确定位视频中的有害区域。

2 相关工作

2.1 多模态仇恨视频检测

仇恨言论检测传统上以文本方法为主 [11], , 但随着模因和视频等有害视觉内容日益增多，研究重点逐渐转向多模态场景 [12, 13]。近期，研究者推出了多个综合性数据集，例如 Facebook 的仇恨模因（FHM） [12] 和多媒体自动厌女识别（MAMI）数据集 [14], , 推动了强大多模态模型的发展，使其能够联合分析图像和文本。这些基准测试表明，仇恨意图往往是通过视觉象征与语言线索之间微妙的相互作用来表达的，而不仅仅依靠直白的侮辱性言辞。

基于这些见解，后续研究开始将多模态仇恨检测从静态图像扩展到动态视频内容。Wang 等人 [15] 通过利用模因数据集扩充有限的仇恨视频数据，解决了数据稀缺问题；研究表明，跨模态迁移学习能够显著提升性能，并凸显了短视频模因与视频帧之间强烈的共享归纳偏置。MultiHateClip（MHC） [16],和 HateMM [17] 数据集通过引入多模态视频标注，为这类研究提供了宝贵的基准测试；不过，它们规模相对较小且来源领域特定，仍给泛化带来了重大挑战。多模态融合方法 [18, 19] 和大规模视觉语言预训练 [20, 21]的进展，则通过共享嵌入空间和交叉注意力机制更好地整合视觉与文本信号，进一步提升了多模态仇恨言论检测能力。

然而，大多数现有方法仍局限于视频级分类，未能捕捉仇恨表达的时间动态；这些表达往往仅在特定时刻出现，或体现在一系列视频帧与口语内容的相互作用中。此外，依赖冻结视觉语言模型的高层嵌入往往导致浅层的跨模态对齐，限制了模型理解细微、依赖上下文的仇恨信号的能力，也无法为其预测提供可解释的依据。相比之下，我们的工作明确针对仇恨内容的帧级定位，并利用大语言模型对丰富的模态特定字幕进行推理，从而超越粗粒度的视频级标签，迈向时间上更细粒度且更具可解释性的判断。

2.2 免训练视频异常检测

传统的视频异常检测（视频异常检测）方法通常需要大量数据标注和模型训练，其形式往往是单类分类器、基于重建的模型或预测网络，这些模型从大量正常视频中学习“正常”的概念。尽管在受限场景中有效，这些方法通常具有领域特异性，若不进行代价高昂的重新训练，便难以泛化到新环境或新的异常类型。近来，利用大规模预训练模型的免训练方法已成为一种有吸引力的替代方案，因为它们能够复用从多样化网络规模数据中获得的强大通用表示。

LAVAD [8] 提出了一种免训练方法，利用视觉语言模型（VLMs）为视频帧生成字幕，随后通过大语言模型（LLMs）估计异常分数。与传统的无监督方法和单类分类方法相比，该方法在标准基准测试中表现显著更优，展示了无需进行显式特定任务训练即可实现视频异常检测的可行性。其他研究也通过利用预训练视觉语言模型和大语言模型，探索了类似的免训练范式

以有效描述和分析视频内容 [22, 23, 24], , 例如将预测描述与正常性先验进行比较，或利用生成模型衡量重建难度，并将其作为异常的替代指标。
尽管取得了这些进展，免训练方法仍受限于生成描述的质量以及底层模型的跨模态对齐能力，因此可能会遗漏领域特定线索或细微的语义异常。此外，视频异常检测通常侧重于检测偏离预期模式的情况，往往依赖低层次的视觉或运动线索；而仇恨内容检测则要求跨多种模态进行高层次语义推理，并对社会、文化和法律规范保持敏感。仇恨本身具有主观性且依赖具体语境，这进一步增加了任务的难度，因此，直接应用免训练的视频异常检测框架不足以实现准确且可解释的仇恨检测。我们的工作可视为将免训练范式从通用异常检测拓展至更具挑战性的仇恨定位任务，并针对这一安全关键领域设计了专门的提示和聚合策略。

2.3 用于视频理解的大语言模型

大型语言模型（大语言模型）彻底改变了自然语言处理任务，展现出强大的推理、抽象和泛化能力 [25]。近期研究日益关注将大语言模型与视觉数据相结合，以推动视频理解任务的发展，例如异常检测、字幕生成和多模态内容审核 [8, 15]。Menapace 等人 [22] 利用大语言模型聚合时间信息，并生成内容丰富且契合语境的描述，用于异常检测；研究表明，对视觉内容进行文本摘要可作为下游推理的有效中间表示。同样，Lee 等人 [16] 利用大语言模型和视觉语言预训练，提升了视频中仇恨内容的识别能力，这表明在提供适当提示词时，大语言模型能够捕捉跨模态的细微语义关系。

除这些具体应用之外，还有一类更广泛的研究致力于探索多模态大语言模型和以视频为中心的架构 [19, 18, 26], : 在这类架构中，视频帧被编码为视觉词元，并输入大语言模型骨干网络，以实现关于动态场景的开放式问答、对话和推理。这些方法展现了大语言模型不仅能够进行内容分析，还能适应各种下游任务的潜力，而这得益于在大规模多模态语料库上的预训练和指令微调。然而，仍然存在一些挑战，尤其涉及时间信息的有效整合、长视频的处理，以及面向安全敏感任务的大语言模型输出校准 [15, 27]。在本文中，我们基于这些发展设计了一种多阶段提示框架，将仇恨定位分解为可解释子任务，并结合基于大语言模型的评分与模态感知组合匹配，以更好地利用时间线索和多模态线索。

3 方法论

3.1 问题形式化

给定一个输入视频 V，其包含多种模态，包括视频帧 $I = [I_1, I_2, \ldots, I_M] \in V$、文本信息 $T = [T_1, T_2, \ldots, T_M] \in V$和音频片段 $A = [A_1, A_2, \ldots, A_M] \in V$，其中 M表示视频帧的总数，现有的仇恨内容检测方法旨在学习一个分类函数 f，能够将每个模态特定输入帧归类为正常类别（非仇恨，得分为 0）或仇恨类别（得分为 1）。通常，检测模型 f 使用带标签数据集 $\mathcal{D}$ 进行训练，该数据集由形如（V_i、 y_i）的样本对组成，其中 V_i 表示视频序列， $y_i \in \{0, 1\}$ 表示其对应的二元仇恨标签。
形式上，给定一个数据集： $\mathcal{D} = \{(V_i, y_i)\}_{i=1}^{N}$，训练过程旨在通过最小化损失函数 $\mathcal{L}$，确定能够将视频输入准确映射到其对应标签的最优函数 f^* :

$$f^* = \arg\min_{f \in \mathcal{F}} \mathcal{L}(f(V), y) \tag{1}$$

其中 $\mathcal{F}$ 表示可能的分类函数的假设空间， $\mathcal{L}$ 为损失函数。现有数据集中标注标签 y 的粒度会因监督策略而显著不同。标签可以是表示帧级标注的稠密二进制向量，也可以是对应于整个视频片段的粗粒度单一二进制标签。然而，由于在整个视频序列上生成细粒度的帧级预测会带来相当大的计算开销，当前方法主要依赖视频级标注。
与传统方法不同，我们提出了一种用于仇恨视频定位的新框架，称为免训练仇恨视频定位（*HVL*）。该范式旨在仅在推理时利用预训练模型，为每个视频帧$I \in V$估算仇恨相关分数即，无需在带标签数据集 $\mathcal{D}$上进行任何训练或微调。

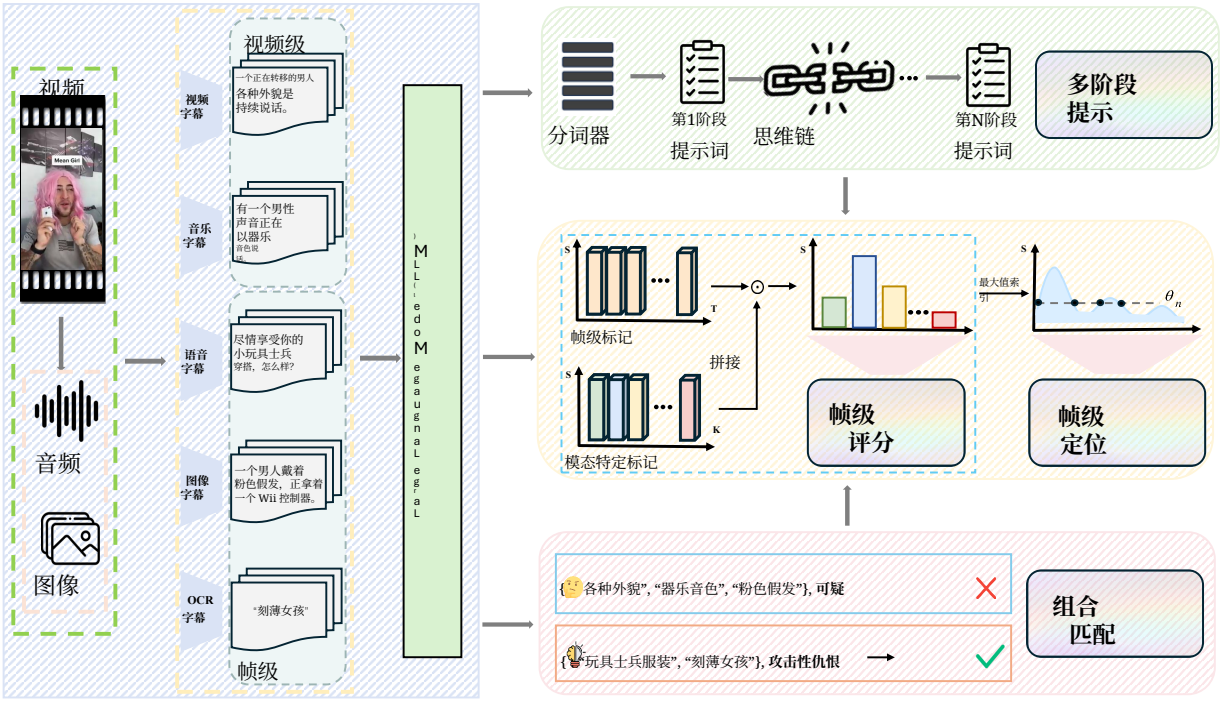

图 3：LELA 框架概览。给定输入视频，LELA 将其分解为五种模态：视频片段、静态图像、背景音乐、语音和 OCR 文本。每种模态都由专门的字幕生成模型处理，以提取与视频帧对齐的文本描述。多阶段提示策略（绿色）从互补视角（例如显性仇恨、隐含仇恨、目标群体）评估帧级模态描述；与此同时，组合匹配模块（粉色）则汇总跨时间和模态的显著信息。随后，将所得文本表示输入大语言模型，以生成各模态的评分；最终帧级仇恨评分取各模态评分中的最大值，从而实现对仇恨内容的细粒度定位，并为内容审核决策提供可解释的依据。

3.2 组合匹配

以往的实证研究 [28] 发现，多模态组件在仇恨检测中起着至关重要的作用。有些视频单独来看，其文本或视觉内容可能都不包含明显的冒犯性内容，但将两者结合起来却会产生新的含义。以往基于融合的研究 [29] [30] 主要关注通用的视觉和音频模态，但对于编码后的嵌入表示如何具体聚焦于细粒度文本内容，人们仍缺乏深入了解。为弥补这一不足，并凸显多媒体中的显式文本信息，我们提出将传统的视觉和音频模态分解为更细粒度的文本字幕。

具体而言，我们将视觉模态细分为图像字幕和 OCR（光学字符识别）字幕，并将音频模态细分为语音字幕和音乐字幕。对于视觉模态，我们使用 BLIP-2 [31] 生成全面的图像字幕，以描述视觉场景的整体内容；同时使用 EasyOCR [32]作为 OCR 字幕生成模型，准确提取视觉内容中出现的文本。对于音频模态，我们使用 Whisper [33] 模型转写口语内容（语音字幕），并利用 LP-Music-Caps [34] 生成背景音乐和环境音的描述性字幕。鉴于单个视频帧仅代表静态视觉快照，我们采用 PDVC [35]模型作为专门的视频字幕生成方法，用于描述视频帧之间的动态事件。通过对齐多种模态的字幕生成结果，我们提出的组合匹配方法增强了多模态嵌入空间中显式文本信号的表示能力，从而显著提升了仇恨内容检测的可解释性和准确性。

为增强跨模态的显式文本对齐，我们在每个视频帧处将模态特定字幕与语音字幕拼接 j:

$$\hat{C}_j^m = \text{Concat}(C_j^{\text{speech}}, C_j^m), \quad m \in \{\text{img}, \text{ocr}, \text{music}, \text{video}\}. \tag{2}$$

随后，我们使用大语言模型对拼接后的字幕进行总结（$\Phi_{\text{LLM}}^{\text{sum}}$）：

$$C_j^{m,\text{sum}} = \Phi_{\text{LLM}}^{\text{sum}}(\hat{C}_j^m) \tag{3}$$

生成 生成简洁而丰富的语义表示，以便通过多阶段提示进行后续评分 .

3.3 多阶段提示

我们任务的主要目标是通过分配一个从 0（非仇恨）到 1（仇恨）的连续分数，准确量化多模态上下文中的仇恨内容程度。基于初步研究（见附录），我们发现传统的单阶段提示在仇恨检测任务中的表现有限。为了解决这一问题，我们提出将推理过程分解为多个阶段，从而实现更细粒度且更易解释的评估。受思维链提示框架 [36] [37], 的启发，我们提出一种**多阶段提示策略**，引导大语言模型以结构化的步骤逐步理解仇恨内容并作出判断。

在第一阶段，我们通过明确的角色定义和针对任务的背景信息来引导大语言模型，这对于使其回应符合内容审核的规范和文化预期至关重要。如下所示，初始提示词界定了仇恨内容的范围，并设定了大语言模型的角色：

> 你是一名视频审核专家，负责评估仇恨内容。仇恨内容的定义是：煽动歧视、造成不适或痛苦，或基于种族、族裔、国籍、宗教、残障、年龄、退伍军人身份、性取向、性别认同等属性贬低个人或群体的视频。

在让模型理解这一界定后，我们提示它分析视频场景的多模态描述，鼓励它综合视觉、听觉和文本模态进行推理。该提示词旨在发挥大语言模型的文本推理能力，内容如下：

> 请结合{video/frame/music/OCR} 和{speech} 描述，分析视频中的视觉、文本和音频元素，以检测并标记任何仇恨内容。不要描述视频的字面内容；而应关注其隐含含义，以及是否传达了仇恨表达。

推理过程可抽象表述如下：

$$Rationale = \text{MLLM}(v_i^T, v_i^F, v_i^{\text{OCR}}, v_i^M, v_i^S) \tag{4}$$

其中，v_i^V、v_i^F、v_i^{OCR}、v_i^M和 v_i 分别表示视频级字幕、帧级字幕、从光学字符识别中提取的字幕、音乐字幕和语音字幕。输出记为推理依据，它呈现了模型基于多模态线索进行的中间推理。

最后，提示大语言模型根据生成的推理依据给出仇恨程度分数。该评分过程形式化如下：

$$s = \Phi_{\text{LLM}}(\text{Rationale}), \quad \in [0, 1], \tag{5}$$

其中，$\Phi^{(2)}$ 大语言模型(·)表示通过基于大语言模型的推理实现的评分函数。评分提示词如下：

> 给定{rationale}，如果你是一名内容审核专家，你会如何对该场景进行评分（范围为0到1）？ 其中，0表示普通（非仇恨）场景，0.5表示具有攻击性、冒犯性或敌意的场景，1表示明确具有仇恨性的场景。

这种多阶段提示　　该设计旨在提升仇恨性评分的粒度、可解释性和准确性
　　在多模态上下文中的有效性，确保使模型推理与特定任务定义及模态感知证据相一致。

3.4 仇恨视频的帧级定位

为了实现对仇恨内容的精确帧级定位，我们首先以语音模态生成的帧级字幕为分析基础。对于其余四种字幕生成模态（图像、光学字符识别、音乐和视频），我们分别将每个视频帧中各模态的字幕与语音字幕拼接，构建用于评分的全面多模态上下文。

如多阶段提示一节所述，在评分之前，我们会让大语言模型总结拼接后的字幕。构建好总结后的多模态字幕后，大语言模型会使用我们的多阶段提示方法，分别对每种模态进行评分。对于每个视频帧 j，这会得到模态特定仇恨度分数：

$$\begin{aligned} s_j^m &= \text{MultiStagePrompting}(C_j^{m,\text{sum}}), \\ & m \in \{\text{img}, \text{ocr}, \text{music}, \text{video}\}. \end{aligned} \tag{6}$$

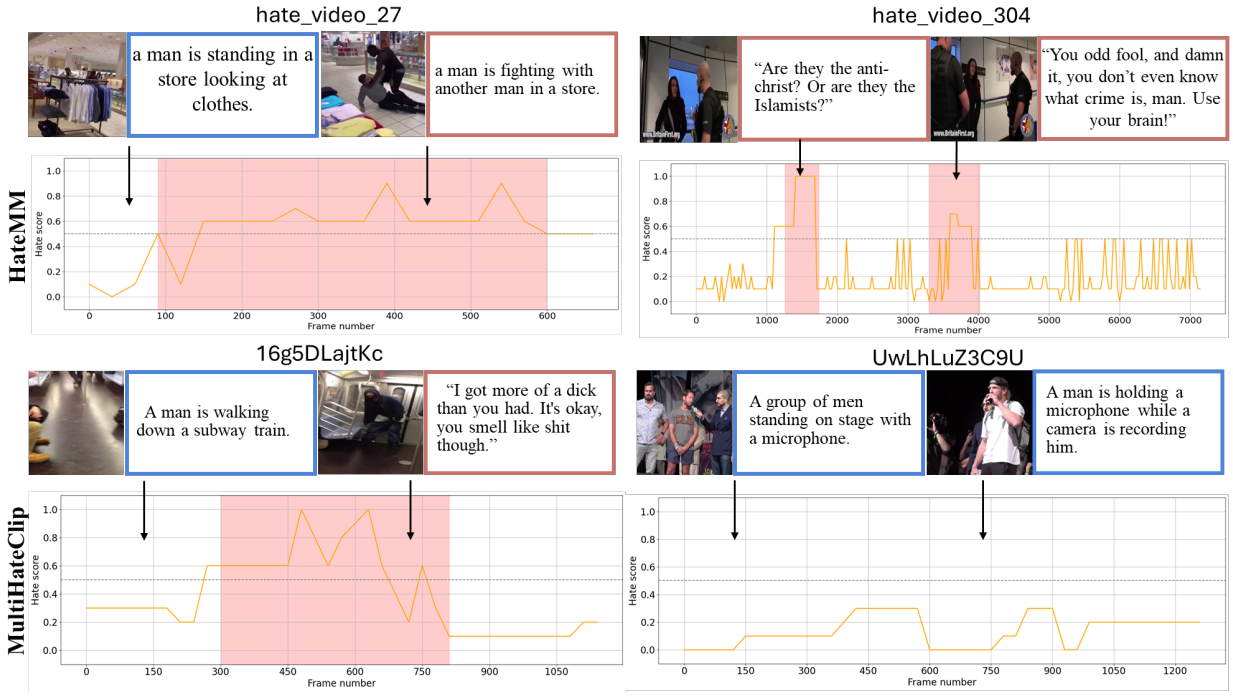

图 4：我们展示了本框架在四段测试视频上获得的定性结果，其中包括来自 HateMM 数据集的两个示例（上行）以及来自 MultiHateClip 数据集的两个示例（下行）。对于每段视频，我们绘制了由本方法计算得到的各视频帧预测仇恨评分。我们展示了选定的关键帧及其最相关的模态描述。被预测为非仇恨的视频帧以蓝色边界框标出，而被预测为仇恨的视频帧以红色边界框标出。真实仇恨片段亦以粉色高亮显示，为评估我们定位结果的准确性与可解释性提供参照。

获得模态特定仇恨度分数后，我们选取所有模态中的最高分，作为当前视频帧的仇恨强度，其公式如下：

$$s_j^{\text{final}} = \max\{s_j^{\text{img}}, s_j^{\text{ocr}}, s_j^{\text{music}}, s_j^{\text{video}}\}. \tag{7}$$

该策略体现了一项内容审核原则：若任一模态强烈表明仇恨性，则该视频帧应被标记以供审核。通过对整段视频的每一视频帧重复这一评分过程，我们得到完整的时序仇恨性概况。最后，我们应用决策阈值 τ 对帧级预测进行二值化：

$$H_j = \mathbb{I}(s_j^{\text{final}} > \tau), \tag{8}$$

其中 $\mathbb{I}(\cdot)$ 为指示函数。 τ 的取值可根据运行约束或验证表现进行调整。由此得到仇恨标注的时序序列，可用于视频级聚合、可视化或下游政策执行。

4 实验

我们在 HateMM [4] 和 MultiHateClip(MHC) [5] 数据集上验证了免训练方法**LELA**，并将其性能与最先进的免训练基线方法进行比较。为验证核心设计选择，尤其是所提出的多阶段提示和组合匹配策略，我们进行了全面的消融研究。此外，我们还进行了模态特定的消融实验，并尝试使用不同的大语言模型进行评分，以评估大语言模型在多模态仇恨视频理解中的有效性。

4.1 与最先进方法的比较

据我们所知，目前没有任何现有方法明确针对帧级视频仇恨定位任务。为与我们的免训练框架进行公平且有意义的比较，我们提出了一组免训练基线

方法	PR-AUC (%)	ROC-AUC (%)	Acc	M-F1	F1分数（H）	P(H)	R(H)
HTMM [4]	-	-	0.7481	0.7353	0.6728	0.6954	0.6546
MHCL [5]	-	-	0.7503	0.7407	0.6238	0.6642	0.6831
HVGuard [38]	-	-	0.8563	0.8597	0.8479	0.8228	0.8009
Yue 等人。[39]	-	-	0.821-	-	0.771-	0.798-	0.754-
CMFusion[40]	-	-	0.823-	-	0.860-	0.817-	0.908-
Wang 等人 [41]	-	-	0.82–	0.82–	0.80–	0.80–	0.79–
MM-HSD [42]	-	-	0.878-	0.874-	0.853-	0.849-	0.857-
ZS-CLIP [43]	0.5216	0.5367	0.5019	0.6455	0.5699	0.5425	0.5215
零样本 ImageBind [44]	0.5237	0.5683	0.5317	0.4813	0.6419	0.5042	0.4981
LLAVA-1.5 [27]	0.5327	0.5529	0.5304	0.5149	0.4971	0.5653	0.4742
LAVAD [8]	0.5781	0.6163	0.5716	0.5862	0.5319	0.5784	**0.6827**
Lin 等人 [10]	0.6239	0.5674	0.6738	0.6127	0.5573	0.7132	0.4568
LELA（我们的方法）	**0.7264**	**0.6756**	**0.7148**	**0.7043**	**0.6484**	**0.8152**	0.5481

表 1：HateMM 数据集上监督方法与零样本方法的比较。

方法	PR-AUC (%)	ROC-AUC (%)	Acc	M-F1	F1分数（H）	P(H)	R(H)
HTMM [4]	-	-	0.6861	0.7456	0.6817	0.6875	0.6597
MHCL [5]	-	-	0.7498	0.7317	0.6153	0.6750	0.6783
HVGuard [38]	-	-	0.8090	0.6646	0.4556	0.4722	0.5000
Yue 等人。[39]	-	-	0.78–	-	0.77–	0.80–	0.77–
Wang 等人 [41]	-	-	0.82–	0.81–	0.76–	0.87–	0.68–
ZS-CLIP [43]	0.5181	0.5449	0.5021	0.6395	0.5794	0.5371	0.5289
零样本 ImageBind [44]	0.5135	0.5753	0.5391	0.4756	0.6392	0.5074	0.5014
LLAVA-1.5[27]	0.5319	0.5438	0.5342	0.5062	0.4974	0.5766	0.4694
LAVAD [8]	0.5865	0.6302	0.5833	0.5799	0.5344	0.5671	**0.6926**
Lin 等人 [10]	0.6147	0.5626	0.6174	0.5849	0.5637	0.7233	0.4813
LELA（我们的方法）	**0.7227**	**0.6733**	**0.7124**	**0.6923**	**0.6568**	**0.8217**	0.5387

表 2：MHC 数据集上有监督方法与零样本方法的比较。

这些基线源自视觉语言模型（VLMs），并结合视频异常检测（VAD）领域的先进方法。此外，为了进行全面评估，我们还纳入了仇恨视频检测（HVD）领域的监督分类方法。

性能指标。 我们使用两项标准指标评估帧级仇恨定位的性能：ROC-AUC 和平均精度（AP）。ROC-AUC 衡量在视频帧级计算的受试者工作特征曲线下面积。这是一项与阈值无关的指标，用于评估模型在所有可能的决策阈值下区分仇恨与非仇恨视频帧的能力，因此尤其适用于定位任务。我们还报告**平均精度（AP）**，其定义为帧级精确率-召回率曲线下面积。在仇恨内容稀少的类别不平衡场景中，这项指标尤为重要。我们的评估协议遵循 [8]中提出的既定做法。

定性结果。 图 4 展示了我们的方法框架在 HateMM 和 MultiHateClip 数据集中的示例视频上的定性结果。我们展示关键帧及其多模态摘要，以说明模型的可解释性。在仇恨视频示例中（例如第 1 行和第 2 行第 1 列），帧级摘要有效捕捉了敌意、冒犯性或攻击性内容的上下文线索。这些丰富的文本描述有助于大语言模型准确定位仇恨片段。相比之下，在一个非仇恨示例中，摘要保持中立，且我们的模型在所有视频帧上始终给出较低的仇恨分数。这证实了模型区分有害内容与无害内容的能力。

定量结果。 我们在 HateMM 和 MHC 数据集上评估了监督方法和零样本方法的性能，结果见表 1 和表 2。HTMM 和 MHCL 等监督基线方法表现出较高的准确性和 F1 分数，但 PR-AUC 表现不佳。相比之下，我们提出的 **LELA**框架

表3：大语言模型在 HateMM 数据集上的性能比较。

方法	大语言模型规模	ROC曲线下面积_	PR曲线下面积_	Acc	宏平均 F1	F1分数 (H)	P(H)	R(H)
开源大语言模型								
DeepSeek-R1-1.5B [45]	1.5B	55.69	49.60	56.99	54.90	45.21	55.03	38.37
DeepSeek-R1-7B [45]	7B	64.73	60.58	64.67	63.37	56.67	71.33	48.89
Qwen2.5-3B [46]	3B	57.49	51.42	58.60	56.70	47.62	58.33	40.23
Qwen2.5-7B [46]	7B	62.14	57.29	63.98	62.49	55.03	68.33	46.07
LLaMA-2 7B [47]	7B	60.97	54.23	61.29	59.60	51.35	63.33	43.18
专有大语言模型								
Gemini-2.0 Flash [48]	-	70.28	**68.98**	70.97	70.08	64.94	**83.33**	53.19
GPT-4o Mini [49]	-	**72.64**	67.56	**71.51**	**70.46**	**64.90**	81.67	**54.85**

表4：大语言模型在 MHC 数据集上的性能比较。

方法	大语言模型规模	ROC曲线下面积_	PR曲线下面积_	Acc	宏平均 F1	F1分数 (H)	P(H)	R(H)
开源大语言模型								
DeepSeek-R1-1.5B [45]	1.5B	55.41	49.18	56.62	54.47	45.20	55.07	38.12
DeepSeek-R1-7B [45]	7B	63.28	59.33	63.70	62.41	55.92	70.81	48.05
Qwen2.5-3B [46]	3B	57.03	51.37	58.21	56.35	47.43	58.26	40.02
Qwen2.5-7B [46]	7B	61.82	56.51	62.41	61.21	53.77	67.45	45.12
LLaMA-2 7B [47]	7B	61.42	53.97	60.92	59.27	50.89	63.02	42.87
专有大语言模型								
Gemini 1.5 Pro [48]	-	70.57	67.12	70.84	70.01	**65.82**	**83.01**	53.05
GPT-4o Mini [49]	**-**	**72.27**	**67.33**	**71.48**	**70.43**	64.84	81.52	**54.81**

在所有指标上始终优于其他零样本模型，甚至取得了与监督方法相当的结果。这些发现凸显了 LELA 有效捕捉并整合不同模态多模态信号的能力。

4.2 不同大语言模型的评估

我们在两个基准测试 HateMM 和 MHC 英语（表 3 和表 4）上评估大语言模型的仇恨理解能力。在 HateMM 上，**DeepSeek-R1-7B** 在开源模型中表现最佳，ROC_曲线下面积为 64.73%，PR_曲线下面积为 60.58%，准确性为 64.67%，宏平均 F1 为 63.37%；其精确率较高（71.33%），但召回率中等（48.89%）。**Qwen2.5-7B** 表现相近，而 **DeepSeek-R1-1.5B** 和 **Qwen2.5-3B** 等较小规模的变体表现明显较差。MHC 英语数据集上也呈现类似趋势：**DeepSeek-R1-7B** 再次优于其他开源模型（ROC_曲线下面积为 63.28%，PR_曲线下面积为 59.33%），其次是 **Qwen2.5-7B**，而较小规模的模型仍然表现不佳。

在两个数据集上，**专有大语言模型**都展现出明显优势。**GPT-4o Mini** 的整体表现最佳（例如，在 Hate MM 上的 ROC_曲线下面积为 72.64%，在 MHC 上为 72.27%），且精确率较高（81.67%、81.52%），F1(H) 分数也较为均衡（64.90%、64.84%）。**Gemini-2.0 Flash** 的表现同样出色，尤其是精确率较高，但召回率和稳定性略逊一筹。根据分析，我们在实验设置中选择 GPT-4o Mini 作为骨干大语言模型。

4.3 消融研究

多阶段提示。 为验证我们所提出的多阶段提示机制的有效性，我们通过将整个过程分解为三个功能组件，开展了系统的消融研究：语境化、理由生成和最终决策。表6展示了通过选择性启用或禁用各个组件所得到的四种提示策略的性能。当仅执行最终决策而没有先进行上下文处理或推理时，模型的曲线下面积最低，仅为56.48%，这表明直接评分不足以可靠地评估仇恨内容。将语境化与最终决策结合后，性能显著提升，曲线下面积增加至65.82%。此外，单独使用理由生成可将性能提升至59.70%，表明中间推理具有积极作用。包含全部三个组件的完整模型取得了最高的曲线下面积68.28%，证明了语境基础与中间推理之间的互补性。这些

决策阈值	准确性
0.3	0.4287
0.4	0.5153
0.5	0.7148
0.6	0.6789
0.7	0.5482

表5：不同决策阈值下的准确性。

语境化	理由生成	最终决策	曲线下面积(%)
✗	✗	✓	56.48
✓	✗	✓	65.82
✗	✓	✓	59.70
✓	✓	✓	**68.28**

表6：消融研究　　多阶段提示策略。第4行表示完整模型。曲线下面积指　　o ROC_AUC.

发现共同凸显了：上下文指导和结构化推理对于在多模态场景中实现准确且稳健的仇恨性评分都至关重要。

组合匹配。 我们通过组合匹配实验进一步考察所提出多模态框架中各模态组件的有效性和必要性。具体而言，我们以多阶段提示机制分析的基线语音字幕为起点，逐步加入其他模态：视觉（图像帧）、光学字符识别提取文本、背景音乐，最后加入完整的视频级上下文。所有实验均使用 GPT-4o Mini 作为骨干大语言模型，以隔离模态组合对最终性能的影响。表7总结了这些结果。

4.4 敏感性分析

对于大语言模型的评分决策，我们在定性和定量实验中均将 0.5 设为阈值。为评估模型在不同决策边界下的稳健性，我们通过在一系列阈值范围内测量分类准确率来开展敏感性分析。如表5所示，准确性在标准阈值 0.5 处达到峰值；当阈值向更低或更高的方向偏移时，准确性均出现明显下降。这表明，模型在 0.5 阈值下表现最为可靠，同时对阈值变化具有中等程度的敏感性，说明大语言模型已有效建立了一套连贯的标量评分系统。

表7：使用 GPT-4o Mini 作为骨干大语言模型的模态组合消融研究。

模态组合	曲线下面积（%）
语音（基线）	68.28
+ 图像	68.89
+ 光学字符识别	71.47
+ 音乐	71.75
+ 视频	**72.27**

5 结论

本文提出了 **LELA**，一种用于视频中帧级仇恨内容定位的新型免训练框架。LELA 利用多模态字幕分解和多阶段提示，引导大语言模型进行结构化推理。在 HateMM 和 MultiHateClip 数据集上的广泛评估表明，LELA 表现更优。该工作凸显了基于语言模型的推理在复杂内容审核任务中的潜力，并为未来开展可扩展、免训练的多模态理解研究铺平了道路。

参考文献

[1] Zeerak Waseem。你是种族主义者，还是我看错了？标注者对 Twitter 仇恨言论检测的影响。载于自然语言处理与计算社会科学第一届研讨会论文集，第 138–142 页，2016 年。[2] Shivam Sharma、Firoj Alam、Md Shad Akhtar、Dimitar Dimitrov、Giovanni Da San Martino、Hamed Firooz、Alon Halevy、Fabrizio Silvestri、Preslav Nakov 和 Tanmoy Chakraborty。检测和理解有害模因：综述。第 5597–5606 页，2022 年。[3] S. Singh、S. Dewangan、G. S. Krishna、V. Tyagi、S. Reddy 和 P. R. Medi。用于暴力检测的视频视觉变换器。*arXiv* 预印本 *arXiv:2209.03561*，2022 年。[4] Mithun Das、Rohit Raj、Punyajoy Saha、Binny Mathew、Manish Gupta 和 Animesh Mukherjee。HateMM：用于仇恨视频分类的多模态数据集。载于国际*AAAI*网络与社交媒体会议论文集，第 17 卷，第 1014–1023 页，2023 年。[5] Han Wang、Tan Rui Yang、Usman Naseem 和 Roy Ka-Wei Lee。MultiHateClip：用于检测 YouTube 和哔哩哔哩仇恨视频的多语言基准数据集。载于第 *32* 届 *ACM*国际多媒体会议论文集，第 7493–7502 页，2024 年。[6] Jian Lang、Rongpei Hong、Jin Xu、Yili Li、Xovee Xu 和 Fan Zhou。贪多嚼不烂：用于短视频仇恨检测的检索增强型多模态专家。载于*ACM 2025* 年网络会议论文集，第 2763–2774 页，2025 年。[7] Ming Shan Hee、Zihan Gao、Yinglong Wang、Xiangxiang Chu、Roy Ka-Wei Lee 和 Zengchang Qin。用于仇恨迷因分类的大型多模态模型对比式指令微调。载于国际*AAAI*网络与社交媒体会议论文集，第 19 卷，第 760–773 页，2025 年。[8] Luca Zanella、Willi Menapace、Massimiliano Mancini、Yiming Wang 和 Elisa Ricci。利用大语言模型实现免训练视频异常检测。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 18527–18536 页，2024 年。[9] Muchao Ye、Weiyang Liu 和 Pan He。Vera：通过视觉语言模型的语言化学习实现可解释的视频异常检测。载于计算机视觉与模式识别会议论文集（*CVPR*），第 8679–8688 页，2025 年 6 月。[10] Dongheng Lin、Mengxue Qu、Kunyang Han、Jianbo Jiao、Xiaojie Jin 和 Yunchao Wei。用于整体零样本视频异常分析的统一推理框架。*arXiv* 预印本 *arXiv:2511.00962*，2025 年。[11] Pinkesh Badjatiya 等。推文仇恨言论检测的深度学习。*WWW*，2017 年。[12] Douwe Kiela 等。仇恨迷因挑战赛：检测多模态模因中的仇恨言论。载于*NeurIPS*，2020 年。[13] Raul Gomez 等。探索多模态出版物中的仇恨言论检测。载于*CVPRW*，2020 年。[14] Elisabetta Fersini 等。SemEval-2022 任务 5：多媒体自动厌女识别。载于*SemEval*，2022 年。[15] Han Wang、Rui Yang Tan 和 Roy Ka-Wei Lee。从模因到视频的跨模态迁移：应对仇恨视频检测中的数据稀缺问题。*arXiv* 预印本 *arXiv:2501.15438*，2025 年。[16] Roy Ka-Wei Lee 等。MultiHateClip：用于仇恨视频片段检测的多模态数据集。*arXiv* 预印本 *arXiv:2408.03468*，2024 年。[17] Matteo Sabatelli 等。HateMM：用于多模态仇恨言论检测的数据集。*arXiv* 预印本 *arXiv:2305.03915*，2024 年。[18] Niklas Muennighoff 等。VLM：视觉语言模型是零样本视频分析器。*arXiv* 预印本 *arXiv:2206.12345*，2022 年。[19] Wei Li 等。VideoChat：用于视频理解的大语言模型。*arXiv* 预印本 *arXiv:2306.12345*，2023 年。[20] Alec Radford 等。通过自然语言监督学习可迁移视觉模型。*ICML*，2021 年。

[21] Junnan Li 等。Blip：引导式语言-图像预训练。arXiv 预印本 arXiv:2201.12086，2022 年。[22] Willi Menapace、Luca Zanella 和 Elisa Ricci。用于视频理解的视觉语言模型。arXiv 预印本 arXiv:2305.12345，2023 年。[23] Mehmet Tur 等。用于视频异常检测的扩散模型。arXiv 预印本 arXiv:2301.12345，2023 年。[24] Mehmet Tur 等。扩散模型中的重建误差与无监督异常检测。arXiv 预印本 arXiv:2302.12345，2023 年。[25] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared Kaplan、Prafulla Dhariwal 等。语言模型是少样本学习者。神经信息处理系统进展，33：1877–1901，2020 年。[26] Jean-Baptiste Alayrac、Jeff Donahue、Paul Luc、Antoine Miech、Ian Barr 等。Flamingo：用于少样本学习的视觉语言模型。arXiv 预印本 arXiv:2204.14198，2022 年。[27] Haotian Liu、Chunyuan Li、Qingyang Wu 和 Yong Jae Lee。视觉指令微调，2023 年。[28] Austin Botelho、Bertie Vidgen 和 Scott A Hale。解读隐性仇恨：评估用于多模态仇恨检测的自动检测算法。arXiv 预印本 arXiv:2106.05903，2021 年。[29] Anusha Chhabra 和 Dinesh Kumar Vishwakarma。Mhs-stma：通过可扩展的基于Transformer的多级注意力框架进行多模态仇恨言论检测。arXiv 预印本 arXiv:2409.05136，2024 年。[30] Minh-Hao Van 和 Xintao Wu。利用视觉语言模型检测和缓解多模态模因中的仇恨内容。arXiv 预印本 arXiv:2505.00150，2025 年。[31] Junnan Li、Dongxu Li、Silvio Savarese 和 Steven Hoi。BLIP-2：利用冻结图像编码器和大语言模型进行引导式语言-图像预训练。载于国际机器学习会议，第 19730–19742 页。PMLR，2023 年。[32] MAM Salehudin、SN Basah、H Yazid、KS Basaruddin、MJA Safar、MH Mat Som 和 KA Sidek。图像退化情况下使用 EasyOCR 进行光学字符识别的分析。载于《物理学杂志：会议系列》，第 2641 卷，第 012001 页。IOP Publishing，2023 年。[33] Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Christine McLeavey 和 Ilya Sutskever。通过大规模弱监督实现鲁棒语音识别，2022 年。[34] SeungHeon Doh、Keunwoo Choi、Jongpil Lee 和 Juhan Nam。Lp-musiccaps：基于大语言模型的伪音乐描述。arXiv 预印本 arXiv:2307.16372，2023 年。[35] Teng Wang、Ruimao Zhang、Zhichao Lu、Feng Zheng、Ran Cheng 和 Ping Luo。通过并行解码实现端到端密集视频描述。载于IEEE/CVF国际计算机视觉会议论文集，第 6847–6857 页，2021 年。[36] Yanzhi Xu、Yueying Hua、Shichen Li 和 Zhongqing Wang。探索思维链在多模态隐喻检测中的应用。载于第 62 届计算语言学协会年会论文集（第一卷：长篇论文），第 91–101 页，2024 年。[37] Nishant Vishwamitra、Keyan Guo、Farhan Tajwar Romit、Isabelle Ondracek、Long Cheng、Ziming Zhao 和 Hongxin Hu。利用大语言模型中的思维链推理应对新一波网络仇恨。载于 2024 年 IEEE安全与隐私研讨会（SP），第 788–806 页。IEEE，2024 年。[38] Yiheng Jing、Mingming Zhang、Yong Zhuang、Jiacheng Guo、Juan Wang、Xiaoyang Xu、Wenzhe Yi、Keyan Guo 和 Hongxin Hu。Hvguard：利用多模态大语言模型进行仇恨视频检测。载于 2025 年自然语言处理实证方法会议论文集，第 9004–9017 页，2025 年。[39] Jiangbei Yue、Shuonan Yang、Tailin Chen、Jianbo Jiao 和 Zeyu Fu。利用双流图神经网络进行多模态仇恨检测，2025 年。[40] Yinghui Zhang、Tailin Chen、Yuchen Zhang 和 Zeyu Fu。通过通道级与模态级融合增强多模态仇恨视频检测。载于 2024 年 IEEE国际数据挖掘会议研讨会（ICDMW），第 183–190 页。IEEE，2024 年。[41] Han Wang、Rui Yang Tan 和 Roy Ka-Wei Lee。从模因到视频的跨模态迁移：应对仇恨视频检测中的数据稀缺问题。载于 ACM 2025 年网络会议论文集，第 5255–5263 页，2025 年。[42] Berta Céspedes-Sarrias、Carlos Collado-Capell、Pablo Rodenas-Ruiz、Olena Hrynenko 和 Andrea Cavallaro。Mm-hsd：视频中的多模态仇恨言论检测。载于第 33 届 ACM国际多媒体会议论文集，第 2546–2555 页，2025 年。

[43] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等人。借助自然语言监督学习可迁移视觉模型。载于国际机器学习会议，第 8748–8763 页。PMLR，2021 年。[44] Rohit Girdhar、Alaaeldin El-Nouby、Zhuang Liu、Mannat Singh、Kalyan Vasudev Alwala、Armand Joulin 和 Ishan Misra。ImageBind：一个将所有模态绑定在一起的嵌入空间。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 15180–15190 页，2023 年。[45] DeepSeek-AI。Deepseek-v3技术报告，2024 年。[46] Jinze Bai、Shuai Bai、Yunfei Chu、Zeyu Cui、Kai Dang、Xiaodong Deng、Yang Fan、Wenbin Ge、Yu Han、Fei Huang 等人。通义千问技术报告。*arXiv* 预印本 *arXiv:2309.16609*，2023 年。[47] Hugo Touvron、Louis Martin、Kevin Stone、Peter Albert、Amjad Almahairi、Yasmine Babaei、Nikolay Bashlykov、Soumya Batra、Prajjwal Bhargava、Shruti Bhosale 等人。Llama 2：开放式基础模型与微调聊天模型。*arXiv* 预印本 *arXiv:2307.09288*，2023 年。[48] Gemini Team Google。Gemini：一系列能力强大的多模态模型。*arXiv* 预印本 *arXiv:2312.11805*，2023 年。[49] Aaron Hurst、Adam Lerer、Adam P Goucher、Adam Perelman、Aditya Ramesh、Aidan Clark、AJ Ostrow、Akila Welihinda、Alan Hayes、Alec Radford 等人。GPT-4o 系统卡。*arXiv* 预印本 *arXiv:2410.21276*，2024 年。