基于共享表征学习的模态缺失条件下鲁棒有害表情包检测

Felix Breiteneder 林茨约翰·开普勒大学，奥地利
felix.breiteneder@aon.at

Mohammad Belal 阿尔托大学 埃斯波，芬兰
mohammad.belal@aalto.fi

Muhammad Saad Saeed 密歇根大学弗林特分校 密歇根州，美国
msaads@umich.edu

Shahed Masoudian 林茨约翰·开普勒大学，奥地利
shahed.masoudian@jku.at

Usman Naseem 麦考瑞大学 悉尼，澳大利亚
usman.naseem@mq.edu.au

Kulshrestha Juhi 阿尔托大学，埃斯波，芬兰
juhi.kulshrestha@aalto.fi

Markus Schedl 计算感知研究所，林茨约翰·开普勒大学与林茨理工学院，林茨，奥地利 markus.schedl@jku.at

Shah Nawaz 林茨约翰·开普勒大学，奥地利
shah.nawaz@jku.at

arXiv:2602.01101v1 [cs.CV] 1 Feb 2026

摘要

互联网梗图是强大的沟通工具，能够传播政治、心理和社会文化观点。然而，它们也可能造成伤害，并被用来传播针对特定个人或群体的仇恨。尽管以往研究侧重于设计新的检测方法，但这些方法往往依赖模态完整数据，例如文本和图像。然而，在现实场景中，文本等模态可能因光学字符识别质量不佳等问题而缺失，导致现有方法对信息缺失十分敏感，性能随之下降。为弥补这一空白，本文首次全面研究了在模态不完整数据条件下有害表情包检测方法的表现。具体而言，我们提出一种新的基线方法，通过对多种模态分别进行投影来学习共享表示。在数据模态不完整时，便可利用这些共享表示。在两个基准数据集上的实验结果表明，当文本缺失时，我们的方法优于现有方法。此外，结果显示，我们的方法能够更好地融合视觉特征，降低对文本的依赖，并提升文本信息缺失场景下的鲁棒性。本研究为推动有害表情包检测的现实应用迈出了重要一步，尤其适用于某种模态缺失的情况。

CCS 概念

• **计算方法** → **神经网络**。

本作品采用知识共享署名 4.0 国际许可协议授权。*WWW Companion '26*，迪拜，阿拉伯联合酋长国© 2026 版权所有归作者或著作权所有者所有。ACM
ISBN 979-8-4007-2308-7/2026/04
https://doi.org/10.1145/3774905.3795610

关键词

多模态学习、视觉与语言、有害梗图、缺失模态

ACM 参考格式：
Felix Breiteneder、Mohammad Belal、Muhammad Saad Saeed、Shahed Ma- soudian、Usman Naseem、Kulshrestha Juhi、MarkusSchedl 和 Shah Nawaz。2026 年。基于共享表征学习的模态缺失条件下鲁棒有害表情包检测。载于《*ACM Web Conference 2026* 伴随论文集》（*WWW Companion '26*），*2026* 年 *4* 月 *13* 日至 *17* 日，迪拜，阿拉伯联合酋长国。ACM，纽约，纽约州，美国，6 页。https://doi.org/10.1145/3774905. 3795610

1 引言

在数字时代，表情包已成为一种普遍的交流形式，将图像与文本融合，用以传达幽默、讽刺或社会文化观点。它们在社交媒体平台上的迅速传播，凸显了其对公共话语的影响。尽管许多表情包无伤大雅，或带有讽刺意味，但越来越多的表情包被用来传播仇恨、虚假信息或骚扰，针对个人、社会群体或机构 [1]。例如，有害梗图可能会嵌入极端主义意识形态、宣扬自我伤害或延续刻板印象，并利用视觉与文本线索之间的相互作用，规避传统的内容审核方法 [3, 22]。此类内容的激增表明，开发稳健的检测机制迫在眉睫，尤其是在各平台努力平衡言论自由与用户安全之际 [1]。由于有害梗图具有多模态特性，其检测面临诸多挑战：这类内容融合了图像和文本组件，往往传达隐含或依赖语境的含义 [24]。

传统方法利用机器学习技术，分别或结合起来分析这两个组成部分，以识别有害内容 [10, 25, 31, 36]。现有的有害表情包检测方法广泛依赖多模态框架来分析视觉和文本元素，例如 VisualBERT [13],

表 1：基于词错误率（WER）和双语评估替补（BLEU）分数，对合并的 Harm-C 和 Harm-P 数据集测试集上的光学字符识别模型性能进行比较。这些模型在真实世界表情包数据上的表现凸显了实际应用中文本可用性的不可靠性。

模型	词错误率（WER）	BLEU
EasyOCR	0.5522	0.4617
PaddleOCR	0.5197	0.4711
Tesseract	0.8129	0.1969

基于 CLIP 的模型或跨模态注意力融合技术。例如，MOMENTA [25] 框架采用多模态网络，从全局和局部视角分析迷因，以评估其潜在危害。这种方法凸显了考虑文本与图像之间相互作用的重要性，有助于理解表情包传达的信息。正如仇恨梗图挑战赛 [11], 等基准测试所表明的那样，这些方法在受控环境中往往表现出色；该挑战赛强调了联合模态分析的优势。近期研究已开始通过将大语言模型纳入检测流程来应对这些挑战 [6, 7, 16]。

然而，在现实场景中存在一个关键局限：某种模态（例如文本）可能缺失、损坏，或被刻意混淆以规避检测。例如，表情包可能使用经过大幅修改且文本极少的图像，或依赖自动化系统无法识别的文化隐喻 [28]。假设模态完整的现有方法对这类缺失十分敏感，导致性能和可靠性下降。表 1 展示了初步分析：光学字符识别系统往往无法提取表情包中嵌入的文本，这证实了我们的假设，即现实场景中的文本可用性往往并不可靠。

本文提出一种用于有害表情包检测的新型基线方法，通过分别投影各个模态来学习多种模态的共享表示。该方法着眼于实际应用场景：表情包往往含义模糊、信息缺失或模态缺失，旨在为多样化的数字环境中的有害内容检测提供一个适应性更强的框架。本文的贡献如下：

- 据我们所知，这是首项全面研究在文本模态缺失这一现实条件下有害表情包检测问题的工作。
- 我们提出一种基线方法，通过学习多种模态的共享表示，即使面对模态不完整数据，也能实现稳健的性能。
- 我们展示了详尽的实验结果与分析，证明所提方法在处理文本缺失问题上的有效性，并凸显其稳健性和实际应用价值。

2 相关工作

2.1 有害表情包检测

以往的有害表情包检测研究主要聚焦于在多模态环境中检测虚假信息、仇恨言论和冒犯性内容设置。包括巨魔模因分类[29] 和仇恨梗图检测 [11, 36] 在内的若干共享任务，推动了这一任务的发展。早期方法曾探索逻辑回归和基于 BERT 的分类器 [5], ，近期研究则聚焦于联合处理文本和图像的多模态网络。VisualBERT [13], 、UNITER [4], 、ViLBERT [18], 和 LXMERT[30] 等模型通过利用联合视觉-语言表征，展现出更优的性能。后续研究还采用了多模态深度集成 [27] 来提升模型性能。然而，这些方法更关注仇恨言论检测和情感分析，而非全面解决有害表情包检测问题。

在有害表情包检测中，模态对齐仍是一项重大挑战，因为文本和视觉组件往往提供互补信息。为解决这一问题，人们开发了不同的融合技术。一些方法将不同模态的编码表征组合到共享特征空间[23, 32],中，而另一些方法则先独立处理各个模态，再合并这些表征 [36]。MOMENTA [25] 引入了跨模态注意力机制，以改善梗图图像与背景文本之间的对齐效果，其性能优于传统融合技术。

其他研究则探讨了将模态迁移作为直接融合的替代方案。Pro-Cap [2] 在零样本视觉问答框架中使用预训练视觉语言模型，为有害内容检测生成图像说明文字。同样， [8]提出将多模态迷因内容转换为文本说明文字，并在基准数据集上表现良好。然而，这些方法依赖高质量的说明文字生成，可能无法很好地泛化到复杂的多模态交互场景。

CapAlign [7] 通过生成信息丰富的说明文字，而非泛泛的文本描述，拓展了这一研究方向。CapAlign 通过提示大语言模型向预训练视觉语言模型提出针对性问题，提取语境相关细节，然后将生成的说明文字与迷因文本内容对齐。该方法提升了跨模态对齐效果，并在有害表情包检测基准测试中展现出最先进的性能。

2.2 模态缺失

多模态学习整合来自不同模态的信息，例如自然语言、视觉信号和音频，以提升各种任务的表现。该领域的一项关键挑战是融合，即将不同模态结合起来，以增强模型的理解能力。特征拼接等简单融合方法已得到广泛应用 [23, 32], ，而张量融合 [33] 和低秩融合 [17] 等更复杂的方法则解决了可扩展性问题。然而，这些方法假设所有模态均可用，而这在现实场景中很少成立 [14]。

为提升多模态学习方法的鲁棒性， [19]研究了Transformer对模态缺失的敏感性，并提出了一种自适应搜索优化方法，以动态选择最有效的融合策略。该方法根据训练条件调整融合机制，从而减轻性能下降。其他方法则致力于确保在模态不完整时的鲁棒性 [15, 20, 21, 26, 34, 35]。SMIL [21] 通过贝叶斯元学习估计潜在特征，

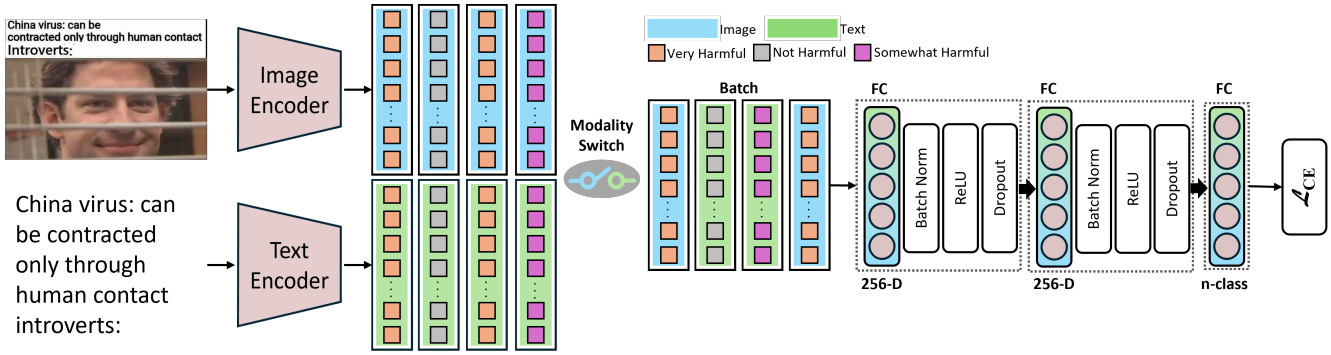


图1：我们的模型概览。每种模态均通过一个开关独立选择，并经过多层全连接网络，以学习共享表示。

而 [34] 引入了标签编码模块，以动态适配Transformer。MMIN [35] 根据可用数据预测缺失模态的表示，Ma等人 [20] 则探索了多任务优化，以改进不完整场景下的融合策略。Lee等人 [12] 提出了模态感知提示，作为全面模型重新训练的替代方案。该方法利用提示学习高效适配Transformer，其中输入级提示可提升性能，而注意力级提示则能增强模型在不同数据集上的稳定性。

3 方法

3.1 问题定义

有害表情包检测是一项多模态任务，需要联合分析文本和视觉信息，以确定表情包是否包含有害内容。形式化地，给定数据集 $D = \{(x_t, x_v, y)\}_{i=1}^{N}$，其中 x_t 和 x_v 分别表示文本和图像模态，且 $y \in Y$ 为目标标签，目标是学习函数 $f: (x_t, x_v) \rightarrow y$，以准确预测表情包的类别。现有多模态方法通常采用双分支网络，分别处理文本和图像数据，然后将二者结合以学习融合表示，用于分类任务。在全模态条件下，这些模型具有良好的效果和模块化特性，但其高度依赖模态的可用性；当某一模态缺失时，分类性能会显著下降。由于现实场景中经常会遇到不完整或噪声数据，例如文本难以辨认或缺失的表情包，因此必须设计能够有效处理这类情况的模型。在本文中，我们通过引入一种多模态学习方法来应对这一挑战，该方法独立处理图像和文本，以学习共享表示。因此，即使文本不可用，模型仍能保持鲁棒性，并利用跨模态交互的共享表示来补偿模态缺失数据。我们的方法提升了模态缺失场景下的分类性能，同时在两种模态均存在时保持具有竞争力的性能。

我们提出的模型（图1）采用了一种方法，将文本和视觉模态嵌入分别映射到共享的潜在空间中。不同于依赖各模态独立分支的传统多模态Transformer，我们的方法确保文本和视觉特征被投影到共享表示空间中，从而提升在模态缺失条件下的鲁棒性。

3.2 模态编码器

我们对两种模态均使用预先提取的 CLIP嵌入，其中$x_t \in \mathbb{R}^d$ 表示文本嵌入，$x_v \in \mathbb{R}^d$ 表示视觉嵌入。这些嵌入作为共享处理模块的输入。

3.3 共享表示网络

该网络由两个堆叠的模块组成，每个模块均包含一个全连接（FC）层，后接批归一化（BN）、ReLU激活和随机失活。该设计使模型能够通过使用同一组参数独立处理文本和图像嵌入来学习共享表示。在测试时两种模态均可用的情况下，模型会先将它们对应的逻辑值相加，再应用 Softmax 进行最终分类。如果测试时仅存在一种模态，模型会直接使用可用模态的逻辑值，而无需进行特定于模态的调整。

3.4 损失函数

我们使用预测类别分布上的交叉熵损失对模型进行优化。通过在共享表示空间中处理两种模态，我们的模型能够有效学习跨模态嵌入，即使缺少文本数据，也能保持较强的分类性能。

4 实验设置

4.1 数据集

我们使用 HarMeme 数据集进行实验，该数据集包含两个子集：Harm-C（与新冠肺炎相关的迷因）和 Harm-P（政治迷因）。数据集将每个迷因标注为三个类别之一：危害极大、有一定危害或无害。这些类别对应二分类任务和多分类任务。据我们所知，本研究首次报告了修正后的 Harm-P 数据集的性能指标。原始数据集由 [25], 引入，其中训练

表 2：Harm-C 和 Harm-P 上二分类的 F1分数。最高分以粗体显示，第二高分带下划线。

	Harm-C（二分类）							Harm-P（二分类）						
模型	100	90	70	50	30	10	0	100	90	70	50	30	10	0
VisualBERT [13]	77.6	76.0	71.2	67.5	61.2	55.0	53.9	61.0	60.0	58.0	54.5	48.2	39.3	11.8
MMBT [9]	78.0	76.0	71.0	68.0	61.0	57.0	55.0	64.9	61.9	55.9	50.1	38.4	21.7	11.8
MOMENTA [25]	**80.8**	**77.5**	70.4	63.5	55.7	43.0	38.1	59.8	58.4	56.9	55.5	52.3	47.9	45.5
Ours	78.7	75.6	**71.7**	**68.1**	**65.3**	**61.0**	**59.1**	**67.5**	**67.4**	**67.3**	**67.0**	**66.7**	**66.5**	**66.3**

表 3：Harm-C 和 Harm-P 多分类任务的 F1分数。最高分以粗体显示，次高分加下划线。

	Harm-C（多分类）							Harm-P（多分类）						
模型	100	90	70	50	30	10	0	100	90	70	50	30	10	0
VisualBERT [13]	49.1	47.4	45.6	42.3	40.6	38.6	38.1	35.6	35.1	34.8	34.7	32.5	28.9	28.4
MMBT [9]	**54.4**	53.4	50.7	44.7	37.4	31.7	29.0	**47.1**	42.7	40.3	38.9	31.7	25.8	22.4
MOMENTA [25]	52.1	49.3	43.6	37.6	30.6	21.9	16.6	39.0	37.9	35.1	33.0	31.1	26.6	25.9
Ours	54.2	**53.4**	**53.4**	**52.9**	**52.0**	**51.3**	**51.1**	46.1	**45.6**	**45.6**	**43.8**	**43.1**	**42.4**	**42.4**

集与测试集/验证集之间存在重复样本，导致数据泄漏。作者后来意识到了这一问题，并发布了更新版本，我们在实验中使用的正是该版本。

4.2 基线方法

我们将模型与三种成熟的多模态方法进行比较：MMBT [9], VisualBERT [13], 和 MOMENTA [25]。

- **MMBT** [9] 使用预训练的 BERT 模型处理文本，并使用 ResNet 主干网络处理图像。提取的特征经过拼接后输入 Transformer 编码器，从而实现多模态数据的后期融合。
- **VisualBERT** [13] 通过将图像区域嵌入注入基于 BERT 的 Transformer，整合文本和视觉表示。该模型针对视觉语言任务进行预训练，通过早期融合学习对齐表示。
- **MOMENTA** [25] 专为有害表情包检测而设计，对全局和局部多模态特征进行编码。它在基于 Transformer 的多模态分类器中加入辅助对比损失，以增强对有害和无害表情包的区分能力。

4.3 实现细节

我们使用 NVIDIA A100 GPU 训练模型 5 个周期，批量大小为 8。训练采用 AdamW优化器，初始学习率为 $1e^{-4}$，并将梯度裁剪阈值设为 2.0。学习率调度包括在前 20% 的训练步数中进行线性预热，随后线性衰减至初始学习率的 10%。所有实验均采用 5 折交叉验证，并使用 3 个不同的随机种子重复进行。我们采用既定的训练集、验证集和测试集划分 [25]。

4.4 评估任务

我们在两项分类任务中评估了所提出的方法：二分类和多分类，以及文本输入缺失的情况。

对于部分样本缺少文本输入的情况。在二分类任务中，非常有害和有些有害这两个类别被合并为一个有害类别。多分类任务保留了原有的三个标签：非常有害、有些有害和无害。

为评估测试集中部分样本缺少文本输入时模型的性能，我们进行了模态缺失实验。我们在不同文本可用性水平下评估模型：100%、90%、70%、50%、30%、10% 和 0%。例如，在文本可用性为 10% 时，只有 10% 的测试样本同时包含文本和图像数据，其余 90% 的样本仅包含图像模态。我们对含有文本的样本进行选择，以保持其标签分布与完整测试集一致，从而确保评估的一致性。为应对类别不平衡，我们采用 F1 分数作为主要评估指标。模型依据验证集性能进行选择，所有结果均在修正后的 HarMeme 数据集测试集上报告。

5 结果与分析

5.1 二分类

表 2 显示，在所有文本可用性水平下，我们的模型在 Harm-P 上均优于所有基线方法。在 Harm-C 上，文本完全可用时 MOMENTA 得分最高，但随着文本可用性降低，其性能急剧下降。当文本可用性为 70% 或更低时，我们的模型优于所有基线方法，展现出对文本缺失的鲁棒性。

5.2 多分类

在表 3 中，当文本可用性为 90% 或更低时，我们的模型在 Harm-C 和 Harm-P 上均取得最佳表现。尽管 MMBT 在文本完全可用时表现最佳，但随着文本被移除，其性能显著下降。VisualBERT 的排名始终低于其他基线方法，尤其难以应对文本缺失的情况。相比之下，我们的模型表现更加稳定

表4：Harm-C 上多分类的 F1分数。最高分以粗体显示，第二高分加下划线。

模型	Harm-C（多分类） 原始	噪声
MOMENTA [25]	<u>52.1</u>	<u>40.8</u>
Ours	**54.2**	**50.0**

在不同文本可用性水平下，凸显了权重共享提升模态鲁棒性的能力。

5.3 文本缺失的影响

我们观察到，当文本可用性为30%或更低时，所有模型的性能都会急剧下降，在某些情况下F1分数最多下降42%。MOMENTA和MMBT的下降最为陡峭，表明它们对文本有很强的依赖。相比之下，我们的模型保持了更高的稳定性，即使在文本可用性为0%时仍能维持具有竞争力的F1 分数。这表明我们提出的模型能够更好地融合视觉特征，降低对文本的依赖，并在文本信息不可靠的场景中提高鲁棒性。

5.4 噪声文本的影响

我们还在Harm-C多类别配置下，比较了我们的模型与使用PaddleOCR提取的噪声文本的MOMENTA的性能。在表4中，MOMENTA表现出急剧下降，表明其对高质量文本描述有很强的依赖。相比之下，我们的模型保持了更高的稳定性。

5.5 消融实验

现有用于检测有害梗图的多模态学习方法通过映射多种模态的融合嵌入来学习共享表示。当模态可用时，这些融合表示（FR）提供了适合执行任务的重叠信息和互补信息。相比之下，我们的基线方法通过在嵌入空间中分别映射多种模态，为其学习共享表示（SR）。当数据存在模态缺失时，可以利用这些共享表示。表5比较了采用融合表示和共享表示的配置与最先进方法的表现。结果表明，当所有模态均可用时，共享表示的表现略低于融合表示。然而，当文本模态缺失时，共享表示的表现显著优于融合表示，证明了学习映射单个模态的有效性。

6 结论

我们通过实证发现，有害表情包检测方法对模态不完整数据十分敏感。基于这些发现，我们提出了一种新颖的基线方法，通过在嵌入空间中对每种模态进行独立映射，学习跨多种模态的共享表示。事实证明，这些共享表示能够有效应对数据完整和缺失的场景。我们认为，我们的工作为

表5：Harm-C 上二分类的 F1分数。最高分以粗体显示，第二高分以下划线标出。

模型	Harm-C（二分类） 100	70	50	10
VisualBERT [13]	77.6	<u>71.2</u>	67.5	55.0
MMBT [9]	78.0	71.0	<u>68.0</u>	<u>57.0</u>
MOMENTA [25]	**80.8**	70.4	63.5	43.0
Ours_FR	**80.8**	66.8	56.3	17.3
本文方法 SR_	<u>78.7</u>	**71.7**	**68.1**	**61.0**

推动有害表情包检测在现实应用中的落地迈出了重要一步，尤其是在某种模态缺失的场景中。

7 致谢

本研究全部或部分由奥地利科学基金（FWF）资助：卓越集群双边人工智能 (https://doi.org/10.55776/COE12) 、doc.funds.connect 项目 以人为中心的人工智能 (https://doi.org/10. 55776/DFH23)以及 PI 项目意图感知音乐推荐系统 (https://doi.org/10.55776/P36413)。为实现开放获取，作者已对本次投稿所产生的任何经作者认可的稿件版本采用知识共享署名许可（CC BY）公共版权许可。

参考文献

[1] Firoj Alam、Stefano Cresci、Tanmoy Chakraborty、Fabrizio Silvestri、Dimiter Dimitrov、Giovanni Da San Martino、Shaden Shaar、Hamed Firooz 和 Preslav Nakov。2022年。多模态虚假信息检测综述。载于第*29*届国际计算语言学大会论文集，Nicoletta Cal- zolari、Chu-Ren Huang、Hansaem Kim、James Pustejovsky、Leo Wanner、Key-Sun Choi、Pum-Mo Ryu、Hsin-Hsi Chen、Lucia Donatelli、Heng Ji、Sadao Kurohashi、Patrizia Paggio、Nianwen Xue、Seokhwan Kim、Younggyun Hahm、Zhong He、Tony Kyungil Lee、Enrico Santus、Francis Bond 和 Seung-Hoon Na（编）。国际计算语言学委员会，庆州，大韩民国，6625–6643。https://aclanthology.org/2022.coling-1.576/ [2] Rui Cao、Ming Shan Hee、Adriel Kuek、Wen-Haw Chong、Roy Ka-Wei Lee 和 Jing Jiang。2023年。Pro-Cap：利用冻结的视觉语言模型进行仇恨梗图检测。第*31*届*ACM*国际多媒体会议论文集（ 2023 260925510 ）。[3] 2024 https://api.semanticscholar.org/CorpusID: Shijing Chen、Usman Naseem、Imran Razzak 和 Flora Salim。未注明年份。揭示厌女模因：对模态效应及其识别的多模态分析。载于*2024*年*ACM*网络会议伴随论文集。1864–1871。[4] Yen-Chun Chen、Linjie Li、Licheng Yu、Ahmed El Kholy、Faisal Ahmed、Zhe Gan、Yu Cheng 和 Jingjing Liu。2019年。UNITER：通用图像文本表示学习。载于欧洲计算机视觉会议。https://api.semanticscholar.org/CorpusID:[5] 216080982

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019年。BERT：用于语言理解的深度双向Transformer预训练。载于北美计算语言学协会分会。https://api.semanticscholar.org/CorpusID:52967399[6] Jianzhao Huang、Hongzhan Lin、Liu Ziyan、Ziyang Luo、Guang Chen 和 Jing Ma。2024年。借助大型多模态模型智能体实现低资源有害模因检测。载于*2024*年自然语言处理经验方法会议论文集，Yaser Al-Onaizan、Mohit Bansal 和 Yun-Nung Chen（编）。计算语言学协会，美国佛罗里达州迈阿密，2269–2293。doi: 10 18653 1 2024 136 。/v / .emnlp-main。Junhui Ji、Xuanrui [7] Lin 和 Usman Naseem。2024年。CapAlign：通过信息丰富的图像描述改进有害模因检测中的跨模态对齐。载于*2024*年*ACM*网络会议论文集（新加坡，新加坡）*(WWW '24)*。计算机协会，美国纽约州纽约市，4585–4594。doi:10.1145/3589334.3648146 [8] Junhui Ji、Wei Ren 和 Usman Naseem。2023年。通过基于提示的方法识别创意有害模因。计算机协会，美国纽约州纽约市。

[9] Douwe Kiela、Suvrat Bhooshan、Hamed Firooz、Ethan Perez 和 Davide Testug- gine。2019年。用于图像和文本分类的有监督多模态双向变换器。*arXiv*预印本 *arXiv:1909.02950*（2019 年）。[10] Douwe Kiela 、 Hamed Firooz 、 Aravind Mohan、Vedanuj Goswami、Aman- preet Singh、Pratik Ringshia 和 Davide Testuggine。2020年。仇恨梗图挑战赛：检测多模态梗图中的仇恨言论。载于神经信息处理系统进展，H. Larochelle、M. Ran- zato、R. Hadsell、M.F. Balcan 和 H. Lin（编），第33卷。Curran Associates, Inc.，2611–2624。https://proceedings.neurips.cc/paper_files/paper/2020/file/1b84c4cee2b8b3d823b30e2d604b1878-Paper.pdf[11] Douwe Kiela、Hamed Firooz、Aravind Mohan、Vedanuj Goswami、Amanpreet Singh、Pratik Ringshia 和 Davide Testuggine。2020年。仇恨梗图挑战赛：检测多模态梗图中的仇恨言论。神经信息处理系统进展33（2020年），2611–2624。[12] Yi-Lun Lee、Yi-Hsuan Tsai、Wei-Chen Chiu 和 Chen-Yu Lee。2023年。面向视觉识别的模态缺失多模态提示。*2023*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）（2023年），14943–14952。https://api.semanticscholar.org/CorpusID:257985347[13] Liunian Harold Li、Mark Yatskar、Da Yin、Cho-Jui Hsieh 和 Kai-Wei Chang。2019年。VisualBERT：视觉与语言领域中简单且高效的基线模型。*arXiv*预印本 *arXiv:1908.03557*（2019年）。[14] Muhammad Irzam Liaqat、Qaiser Abbas、Shah Nawaz、Zaigham Zaheer、Marta Moscati、Yufang Hou、Muhammad Haris Khan、Salman Khan、Elisabeth Andre 和 Markus Schedl。2025年。不完美数据条件下的多模态学习：综述。*Authorea*预印本（2025年）。[15] Muhammad Irzam Liaqat、Shah Nawaz、Muhammad Zaigham Zaheer、Muham- mad Saad Saeed、Hassan Sajjad、Tom De Schepper、Karthik Nandakumar、Muham- mad Haris Khan、Ignazio Gallo 和 Markus Schedl。2025年。Chameleon：一种能够适应模态缺失的多模态学习框架。《国际多媒体信息检索期刊》14，2（2025年），21。[16] Hongzhan Lin、Ziyang Luo、Jing Ma 和 Long Chen。2023年。表象之下：从大语言模型蒸馏多模态推理以揭示有害梗图。载于计算语言学协会年会：*EMNLP 2023*论文集，Houda Bouamor、Juan Pino 和 Kalika Bali（编）。计算语言学协会，新加坡，9114–9128。doi:10.18653/v1/2023.findings- emnlp.611[17] Zhun Liu、Ying Shen、Varun Bharadhwaj Lakshminarasimhan、Paul Pu Liang、Amir Zadeh 和 Louis-Philippe Morency。2018年。利用模态特定因子实现高效低秩多模态融合。*arXiv*abs/1806.00064（2018年）。https://api.semanticscholar.org/CorpusID:44131943[18] Jiasen Lu、Dhruv Batra、Devi Parikh 和 Stefan Lee。2019年。ViLBERT：为视觉与语言任务预训练任务无关的视觉语言表征。载于神经信息处理系统。https://api.semanticscholar.org/CorpusID:199453025[19] Mengmeng Ma、Jian Ren、Long Zhao、Davide Testuggine 和 Xi Peng。2022年。多模态Transformer能否适应模态缺失？载于*2022*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）。18156–18165。doi:10.1109/CVPR52688.2022.01764[20] Mengmeng Ma、Jian Ren、Long Zhao、Davide Testuggine 和 Xi Peng。2022年。多模态Transformer能否适应模态缺失？载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集。18177–18186。[21] Mengmeng Ma、Jian Ren、Long Zhao、S. Tulyakov、Cathy Wu 和 Xi Peng。2021年。SMIL：严重模态缺失情况下的多模态学习。*arXiv*abs/2103.05677（2021）。https://api.semanticscholar.org/CorpusID:232170317[22] Usman Naseem、Jinman Kim、Matloob Khushi 和 Adam G Dunn。未注明年份。用于识别Twitter上疫苗主题争议性梗图的多模态框架。载于第十六届*ACM*国际网络搜索与数据挖掘会议论文集。706–714。[23] Soujanya Poria、Iti Chaturvedi、E. Cambria 和 Amir Hussain。2016年。基于卷积多核学习的多模态情绪识别与情感分析。*2016*年电气电子工程师学会第*16*届国际数据挖掘会议（*ICDM*）（2016），439–448。https://api.semanticscholar.org/CorpusID:5749615[24] Shraman Pra manick、Dimitar Dimitrov、Rituparna Mukherjee、Shivam Sharma、Md Shad Akhtar、Preslav Nakov 和 Tanmoy Chakraborty。2021年。检测有害梗图及其目标对象。*arXiv* 预印本 *arXiv:2110.00413*（2021）。[25] Shraman Pramanick、Shivam Sharma、Dimitar Dimitrov、Md. Shad Akhtar、Preslav Nakov 和 Tanmoy Chakraborty。2021年。MOMENTA：用于检测有害梗图及其目标对象的多模态框架。载于计算语言学协会论文集：*EMNLP 2021*年会议论文集，Marie-Francine Moens、Xuanjing Huang、Lucia Specia 和 Scott Wen-tau Yih（编）。计算语言学协会，多米尼加共和国蓬塔卡纳，4439–4455。doi:10.18653/v1/2021.findings-emnlp.379[26] Muhammad Saad Saeed、Shah Nawaz、Muhammad Zaigham Zaheer、Muhammad Haris Khan、Karthik Nandakumar、Muhammad Haroon Yousaf、Hassan Sajjad、Tom De Schepper 和 Markus Schedl。2024年。用于处理模态缺失的模态不变多模态学习：单分支方法。*arXiv* 电子预印本（2024），arXiv–2408。[27] Vlad Sandulescu。2020年。使用多模态深度集成检测仇恨模因。*ArXiv*abs/2012.13235（2020）。https://api.semanticscholar.org/CorpusID: 229371500[28] Shivam Sharma、Firoj Alam、Md. Shad Akhtar、Dimitar Dimitrov、Giovanni Da San Martino、Hamed Firooz、Alon Halevy、Fabrizio Silvestri、Preslav Nakov 和 Tanmoy Chakraborty。2022年。检测和理解有害梗图：综述。载于第三十一届国际人工智能联合会议论文集，*IJCAI-22*，Lud De Raedt（编）。国际人工智能联合会议组织，5597–5606。doi:10.24963/ijcai.2022/781 综述专题。[29] Shardul Suryawanshi 和 Bharathi Raja Chakravarthi。2021年。泰米尔语巨魔模因分类共享任务的研究成果。载于*DRAVIDIANLANGTECH*。https://api.semanticscholar.org/CorpusID:235365114[30] 2019 Hao Hao Tan 和 Mohit Bansal。LXMERT：从 Transformer 中学习跨模态编码器表示。载于自然语言处理实证方法会议。https://api.semanticscholar.org/CorpusID: 201103729[31] Riza Velioglu 和 Jewgeni Rose。2020年。使用多模态深度学习方法检测表情包中的仇恨言论：仇恨梗图挑战赛获奖方案。*ArXiv* abs/2012.12975（2020）。https://api.semanticscholar.org/ CorpusID:229371541[32] Haohan Wang、Aaksha Meghawat、Louis-Philippe Morency 和 Eric P. Xing。2016年。选择加性学习：提升多模态情感分析中的泛化能力。载于*2017*年电气电子工程师学会国际多媒体与博览会会议（*ICME*）（2016），949 954 6765227 –。https://api.semanticscholar.org/CorpusID: [33] Amir Zadeh、Minghai Chen、Soujanya Poria、E. Cambria 和 Louis-Philippe Morency。2017年。用于多模态情感分析的张量融合网络。载于自然语言处理实证方法会议。https: // api.semanticscholar.org/CorpusID:[34] 950292 Jiandian Zeng、Tianyi Liu 和 Jiantao Zhou。2022年。不确定模态缺失情况下的标签辅助多模态情感分析。计算机协会，纽约，美国。[35] Jinming Zhao、Ruichen Li 和 Qin Jin。2021年。用于不确定模态缺失情况下情绪识别的缺失模态想象网络。载于计算语言学协会第*59*届年会暨自然语言处理国际联合会议第*11*届会议论文集（第*1*卷：长篇论文），Chengqing Zong、Fei Xia、Wenjie Li 和 Roberto Navigli（编）。计算语言学协会，线上，2608–2618。doi: 10 18653 1 2021 203 。/v / 。acl-long。Yi Zhou、Zhenhao Chen [36] 和 Huiyuan Yang。2021年。用于检测有害模因的多模态学习。载于*2021*年电气电子工程师学会国际多媒体与博览会研讨会（*IC MEW*）。电气电子工程师学会，1–6。