arXiv:2603.02519v1 [cs.MM] 3 Mar 2026

# 基于自适应测试时扩展的智能体混合来源多模态虚假信息检测

Wei Jiang 昆士兰大学 布里斯班，澳大利亚 wei.jiang@uq.edu.au

Tong Chen 昆士兰大学 布里斯班，澳大利亚 tong.chen@uq.edu.au

Wei Yuan 昆士兰大学 布里斯班，澳大利亚 w.yuan@uq.edu.au

Quoc Viet Hung Nguyen 格里菲斯大学 黄金海岸，澳大利亚 henry.nguyen@griffith.edu.au

Hongzhi Yin*昆士兰大学 布里斯班，澳大利亚 h.yin1@uq.edu.au

## 摘要

视觉语言模型（VLMs）已被证明能够有效检测社交平台上的多模态虚假信息，尤其是在无法获取标注或标注延迟的零样本设置下。然而，单个视觉语言模型的能力不足以应对更复杂的混合来源多模态虚假信息检测（M³D）任务。以带说明文字的图像为例，在 M³D 中，虚假信息可能源于不真实文本、伪造图像，或两种模态之间的不匹配。尽管近期的智能体系统能够通过连接特定模态的视觉语言模型智能体来处理零样本 M³D，但其效能仍受限于智能体架构。在现有的智能体 M³D 方案中，对于任何输入样本，每个智能体都只进行单次前向推理，因此在具有挑战性的案例中，其决策容易受到固有模型随机性和推理错误的影响。此外，由于缺乏对替代推理路径的探索，现代视觉语言模型的推理能力未能得到充分发挥。本文提出 AgentM³D，一种用于零样本 M³D 的多智能体框架。为增强视觉语言模型的推理能力，我们提出一种创新的自适应测试时扩展范式：AgentM³D 通过 best-of-*N* 机制扩展每个特定模态视觉语言模型智能体的推理，并设计了评判智能体以实现任务对齐评分。所有智能体均置于一个按模态划分的级联决策链中，从而减少不必要的智能体调用并限制错误传播。此外，为确保 AgentM³D 的可扩展性，规划智能体会根据样本难度动态确定所需推理路径的最大数量，并采用自适应停止机制，进一步防止每个智能体过度思考。在两个 M³D 基准测试上的大量实验表明，与.....相比，AgentM³D 的零样本检测性能达到了最先进水平

与各种基于视觉语言模型的方法和智能体基线进行比较。

## 关键词

多模态虚假信息检测；多智能体系统；测试时扩展

*通讯作者。

*Conference'17*, 美国华盛顿特区2026年. ACM
ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

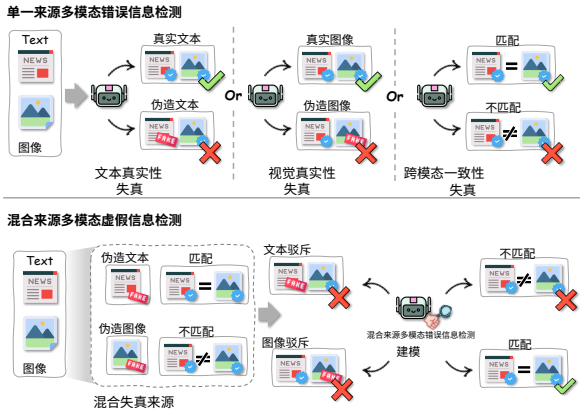



**图 1：单一来源与混合来源多模态虚假信息检测的比较。**

## 1 引言

社交媒体上的多模态虚假信息迅速演变，新的操纵模式不断涌现 [18]，包括假新闻、图像伪造、叙事篡改等。在这类动态环境中，收集涵盖新出现伪造类型的高质量标注数据往往难以实现，或会严重延迟 [22, 29]。因此，监督式多模态虚假信息检测（MMD）系统难以在现实部署中持续发挥效用。这使得零样本多模态错误信息检测成为一项切实且迫切的需求，相关解决方案也日益受到关注 [1, 14] 。

近年来，基于大规模图文语料库 [40], 预训练的视觉语言模型（VLM）展现出强大的跨模态推理和泛化能力，因此被广泛用于零样本多模态虚假信息检测 [16,27]。然而，现有的基于视觉语言模型的多模态错误信息检测方法大多隐含地基于对虚假信息来源的简化假设。具体而言，这些方法主要关注虚假信息具有相同操纵模式的场景，忽视了现实世界中的虚假信息往往呈现混合来源特征这一事实，如图 1 所示。在混合来源多模态虚假信息检测（M³D）场景中，虚假信息可能来自各种操纵来源。如图 1 所示，在包含图像的新闻中，虚假信息可能源自不实文本主张、伪造或被操纵的图像，或真实性模态之间细微的不匹配

。[102]一实际场景进一步将检测任务从二分类扩展到多分类 [5, 16],，以便精确归因伪造内容的具体来源。因此，这些来源之间的强相互依赖关系使混合来源检测对单一视觉语言模型解决方案而言极具挑战。具体而言，混合来源检测要求进行分阶段条件推理：应独立于视觉内容评估文本主张，必须根据客观事实核实视觉证据，并依据先前决策的结果评估跨模态对齐。当单一视觉语言模型被迫同时对所有这些相互耦合的方面进行推理时，即使某一维度上的细微错误也可能传播开来，导致最终预测错误 [28, 43]。

为解决这一问题，近期研究探索了用于零样本混合来源检测的智能体框架，将检测分解为特定模态的推理阶段 [16]。通过连接多个专门的视觉语言模型智能体，这些系统引入了结构化推理，与单体模型相比，部分降低了决策耦合。然而，现有的智能体混合来源检测方法仍受到其推理范式的根本限制。对于每个输入样本，各个智能体通常只执行一次前向推理过程，使其预测容易受到固有模型随机性以及困难案例中脆弱推理的影响。更重要的是，缺乏对替代推理路径的系统性探索，使现代视觉语言模型无法充分发挥其潜在推理能力；尤其是在复杂的混合来源场景中，单条推理轨迹往往不足以应对问题。一些近期的智能体系统尝试通过引入规划模块和大量工具调用来缓解这一局限 [5]。例如，T$^2$Agent 使用规划器将每个输入分解为多个可由工具解决的子任务，并确定不同信息来源的优先级。然而，这些子任务仍通过单次视觉语言模型推理来执行，因此基本上没有解决一次性决策的根本推理瓶颈。此外，这类设计高度依赖第三方组件，而这些组件的可用性和可靠性往往难以确定，可能会损害系统在实际部署中的稳健性。因此，尽管近期取得了进展，构建有效的零样本视觉语言模型混合来源检测系统仍是一项尚待解决的挑战。

在本研究中，我们旨在充分利用视觉语言模型在推理时的推理能力，通过测试时扩展（TTS） [4]提升 M$^3$D 系统的有效性。TTS 无需额外训练，即可通过探索并评估多条推理轨迹来提高推理可靠性；事实证明，它能够增强模型在具有挑战性的决策场景中的表现 [41]。然而，将 TTS 直接应用于 M$^3$D 会带来若干不可忽视的挑战。首先，何时扩展：不同样本的推理难度差异很大，而一些简单案例只需一次前向传播即可准确解决。因此，对所有输入一律采用测试时扩展会造成不必要的计算开销，却未必带来明显收益。其次，扩展多少：即使对于需要增强推理的样本，其难度也会因虚假信息来源和模态而异。对简单的推理路径过度扩展，不仅会增加推理成本，还可能导致过度思考并降低决策质量，因此必须制定合理的停止准则。第三，如何可靠地评分：测试时扩展依赖奖励模型评估多条推理轨迹，但通用奖励模型并未针对 M$^3$D 的复杂决策需求进行调整，在混合来源场景中可能会误判推理质量。迄今为止，尚无现有框架能以可扩展且具成本效益的方式，针对零样本 M$^3$D 同时解决这些挑战。

为应对这些挑战，我们提出**AgentM$^3$D**，一种多智能体框架，利用自适应测试时扩展，增强视觉语言模型的内在推理能力，以实现零样本混合来源多模态虚假信息检测。AgentM$^3$D 将模态专属检测智能体组织成层级级联结构，从而缓解决策耦合并限制推理阶段之间的错误传播。具体而言，为了确定何时需要增强推理，AgentM$^3$D 引入一个轻量级规划智能体，根据样本难度有选择地启用测试时扩展。为确保对扩展后的推理候选项进行可靠评分，该框架集成了批评感知的 Best-of-$N$ 推理 [26]，其中模态专属批评信号与通用奖励模型相辅相成，提供任务对齐评估。为控制需要多少计算资源，该框架采用自适应提前停止机制，动态控制推理探索的程度。这些组件共同将测试时扩展从蛮力启发式方法转变为结构化、决策感知的推理过程，从而实现稳健且高效的零样本 M$^3$D。

我们的贡献有三个方面：

- 据我们所知，我们首次系统刻画了将测试时扩展应用于M$^3$D的关键挑战，包括何时扩展、如何赋予可靠分数，以及多大的计算量是合理的。
- 我们提出AgentM$^3$D，一种具有自适应测试时扩展的多智能体系统，在零样本设置下联合解决何时扩展、如何可靠评分以及需要多少计算量的问题。
- 我们在多个多模态虚假信息基准上对AgentM$^3$D进行了实证验证，结果表明，相较于强大的基于视觉语言模型的基线和智能体基线，其有效性和测试时效率均有所提升。

## 2 预备知识

**问题形式化**。我们研究混合来源多模态虚假信息检测（M$^3$D）问题。形式上，一组新闻表示为$\mathrm{D}_{\text{news}} = \{(T_i, V_i, y_i)\}_{i=1}^{N}$，其中每个条目由文本主张 $T_i$、关联的新闻图像 $V_i$以及真实标签 $y_i \in \{0,1,2,3\}$组成。标签空间对应四个互斥类别，分别为：原始（真实）、文本真实性失真、视觉真实性失真和跨模态一致性失真。

我们将 M$^3$D 形式化为零样本多分类任务。给定输入样本 $(T_i, V_i) \in \mathrm{D}_{\text{news}}$，单个智能体 $a$ 会生成类别预测以及相应的自然语言解释：

$$(y_{i,a}, r_{i,a}) \sim \phi_a(T_i, V_i, \mathcal{P}_a), \tag{1}$$

其中$r_{i,a} \in \mathrm{R}$ 为相应的自然语言解释，$\mathrm{P}_a$ 为智能体 $a$的提示。

我们进一步将这一表述扩展到多智能体系统 $\mathrm{A} = \{a_1, a_2, \ldots\ldots, a_m\}$，其中智能体可以承担不同角色并协同工作。智能体通过将其他智能体的输出作为额外输入进行交互。

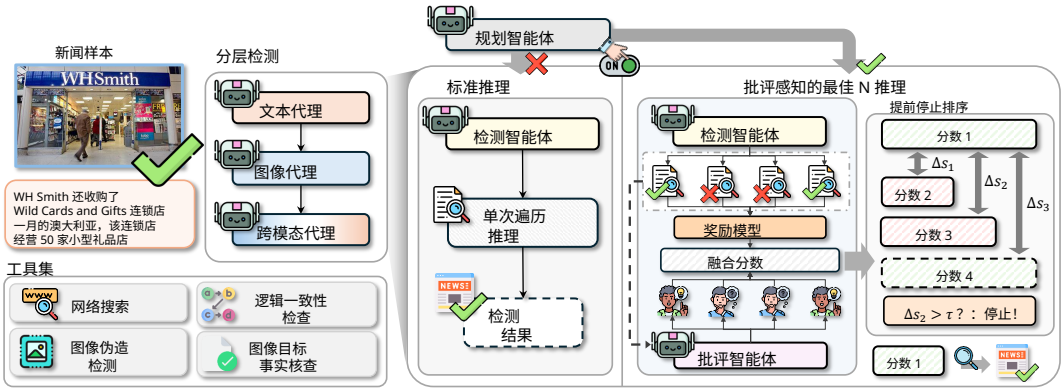



**图2：AgentM³D的整体结构。规划智能体将每个输入路由至标准推理或批判感知的Best-of-$N$推理。后者探索多条推理轨迹，整合奖励与批判信号以进行候选项选择，并应用自适应提前停止排序，以便在出现高置信分数时动态终止推理。**

如果智能体$a_j$与一组智能体 N ($j$) ⊆A 相关联，那么其推理既取决于输入 ($T_i$, $V_i$)，也取决于所有相关智能体的输出：

$$(y_{i,a_j}, r_{i,a_j}) \sim \phi_{a_j}\Big(T_i, V_i, \{(y_{i,a_k}, r_{i,a_k}) \mid a_k \in \mathcal{N}(j)\}, \mathcal{P}_{a_j}\Big). \quad (2)$$

**基于Best-of-$N$的测试时扩展（BoN）**。除单次推理之外，可以考虑采用测试时扩展（TTS）策略来增强单个智能体的推理鲁棒性。具体而言，给定输入样本（$T_i$, $V_i$）和智能体 $a$，该方法允许该智能体在相同提示P$_a$下执行多次独立推理试验。形式上，智能体被激发 $N$ 次，以产生一组候选输出：

$$\{(y_{i,a}^{(m)}, r_{i,a}^{(m)})\}_{n=1}^{M} \sim \phi_a(T_i, V_i, \mathcal{P}_a), \quad (3)$$

其中，每次试验对应于该智能体推理过程的一次独立随机实现。

BoN 策略根据预定义的选择标准 S，从这些候选项中选出单个输出：

$$(y_{i,a}^{*}, r_{i,a}^{*}) = \mathcal{S}\Big(\{(y_{i,a}^{(m)}, r_{i,a}^{(m)})\}_{n=1}^{M}\Big). \quad (4)$$

直观而言，BoN 是一种测试时扩展，使智能体能够探索多种推理路径并保留最可靠的预测，从而在无需额外训练的情况下提升鲁棒性。

## 3 方法

本节介绍 AgentM³D 的技术细节。AgentM³D 是一个用于混合来源多模态错误信息检测的多智能体框架。AgentM³D 采用分层自适应设计，通过依次调用专用智能体，以结构化方式评估文本真实性、视觉真实性和跨模态一致性。我们首先在第 3.1 节介绍分层的模态专用检测智能体；这些智能体通过单次前向推理，将错误信息检测分解为文本、图像和跨模态推理阶段。尽管这种推理范式行之有效，但其本质上容易受到随机推理差异的影响，也可能无法充分发挥现代视觉语言模型的全部推理能力。为解决这一局限，第 3.2 节介绍了一种基于选择性 Best-of-$N$ (BoN) 推理与批评感知排序的测试时扩展机制，使智能体能够探索多条推理轨迹并提高鲁棒性。最后，为兼顾检测性能与推理效率，第 3.3 节提出一种轻量级规划策略，用于动态确定何时分配测试时计算。AgentM³D 的整体结构如图2所示。

### 3.1 分层模态专用检测

受 MMD-Agent 启发，AgentM [16], D 将多模态虚假信息分析分解为多个分层推理阶段。然而，在 MMD-Agent 中，这种分解是通过单个视觉语言模型内固定的提示链实现的，其中间推理结果被隐式嵌入后续提示中。相比之下，AgentM³D 将每个推理阶段明确设定为独立的检测智能体，并为其定义清晰的输入输出接口。这些智能体通过条件激活组成分层级联，逐步细化检测决策，从单模态真实性评估推进至跨模态一致性检查。具体而言，级联从文本真实性检测开始，因为文本主张通常提供主要的语义信号，且能够以相对较低的计算成本高效评估，从而及早筛除失真内容。每个检测智能体不仅会给出类别决策，还会将明确的自然语言推理轨迹作为输出的一部分。这种显式的智能体抽象将各阶段的推理与提示工程解耦，并使得能够直接在智能体推理层面，以有原则的方式建模智能体间依赖关系和条件激活。

具体而言，AgentM³D 采用三类智能体来检测文本、视觉和跨模态错误信息。

**文本真实性检测智能体**。文本真实性检测智能体负责评估文本主张 $T$ 是否与可信的客观证据相矛盾。其推理过程可表述为

$$(y_{\text{text}}, r_{\text{text}}) \sim \phi_{\text{text}}(T, \mathcal{P}_{\text{text}}, \mathcal{T}_{\text{text}}), \quad (5)$$

其中， $y_{\text{text}}$ 表示智能体给出的预测文本真实性标签， $r_{\text{text}}$ 表示相应的显式自然语言推理。$\mathbb{T}_{\text{text}}$ 表示从外部文本信息工具检索到的辅助证据（详情见附录A.3）。**视觉真实性检测智能体**。视觉真实性检测智能体用于评估图像 $V$ 是否与可信的客观证据相矛盾，或是否违反常识约束。其推理过程可表述为

$$(y_{\text{image}},\ r_{\text{image}}) \sim \phi_{\text{image}}\big(V,\ \mathcal{P}_{\text{image}},\ \mathcal{T}_{\text{image}}\big), \tag{6}$$

其中， $y_{\text{image}}$ 表示预测的视觉真实性标签， $r_{\text{image}}$ 表示相应的显式自然语言推理。$\mathbb{T}^{\text{image}}$表示通过图像分析工具获得的辅助观察结果（详情见附录A.3）。

**跨模态一致性检测智能体**。跨模态一致性检测智能体用于评估文本主张 $T$ 与图像 $V$ 在语义上是否一致，即它们是否指向相同的实体、事件和语境。与文本真实性和视觉真实性智能体不同，该智能体并不旨在验证各模态信息的事实正确性。相反，它针对的是欺骗性配对：每种模态单独来看都可能显得合理，但组合起来却形成误导性叙事。其推理过程可表示为

$$(y_{\text{cross}},\ r_{\text{cross}}) \sim \phi_{\text{cross}}\big(T,\ V,\ \mathcal{P}_{\text{cross}},\ \mathcal{T}_{\text{image}}\big), \tag{7}$$

其中， $y_{\text{cross}}$ 表示预测的跨模态一致性标签， $r_{\text{cross}}$ 表示相应的显式自然语言推理。我们引入$\mathbb{T}^{\text{image}}$，这是一种由图像分析工具生成的辅助观察结果，可提供基于视觉内容的描述。该基于视觉依据的摘要将图像内容明确呈现出来，便于与文本主张进行比较，从而有助于可靠地评估二者是否一致。请注意，我们在此处未使用文本工具$\mathbb{T}_{\text{text}}$，因为它用于真实性验证，而跨模态一致性检测并不需要进行此项验证。

**分层级联执行**。智能体M³D将上述三个检测智能体组织成具有条件激活的分层级联。其核心思想是将检测决策从单模态真实性评估逐步细化到跨模态一致性检验，同时避免不必要的计算。我们首先将失真指示函数定义为：

$$\delta(y) = \mathbb{I}[y \neq \text{original}], \tag{8}$$

其中 $y \in \{y_{\text{text}}, y_{\text{image}}, y_{\text{cross}}\}$表示检测智能体产生的类别预测。指示量 $\delta(y) = 1$ 表示检测到失真，而 $\delta(y) = 0$ 表示内容被归类为原始内容。

基于 $\delta(\cdot)$，三个检测智能体的激活指示量定义如下：

$$\begin{aligned} \alpha_{\text{text}} &= 1, \\ \alpha_{\text{image}} &= 1 - \delta(y_{\text{text}}), \\ \alpha_{\text{cross}} &= (1 - \delta(y_{\text{text}})) \cdot \big(1 - \delta(y_{\text{image}})\big), \end{aligned} \tag{9}$$

其中， $\alpha_k \in \{0,\ 1\}$表示第 $k$个检测智能体的激活状态， $\alpha_k = 1$ 表示调用该智能体， $\alpha_k = 0$ 表示跳过该智能体。直观而言，式 9 定义了一个条件级联推理过程。AgentM³D 从文本真实性检测开始。只有当文本内容被判定为原始内容时，框架才会继续进行视觉真实性检测。同样，只有当文本和视觉内容都被判定为原始内容时，才会激活跨模态一致性检测智能体。一旦任何智能体检测到失真，级联过程便会终止，并将输入判定为错误信息。只有通过所有已激活检测阶段的输入才被视为可信。

## 3.2 结合批判感知排序的N选优推理

上一节介绍的分层模态特定检测框架为多模态虚假信息检测任务提供了结构化且有效的分解。然而，迄今所述的检测智能体采用单次前向推理范式，每个智能体在固定提示条件下仅产生单一推理结果。尽管计算效率较高，单次推理往往不足以充分挖掘现代视觉语言模型的推理能力。由于视觉语言模型推理具有随机性，单次推理可能导致预测不稳定，并偶尔出现推理失败，尤其是在模糊或具有挑战性的情形中。因此，即使整体框架设计合理，分层级联的检测性能仍可能不够理想。

为解决这一局限，我们引入一种测试时扩展机制，通过在推理过程中鼓励生成多条推理轨迹，增强检测的稳健性并激发视觉语言模型的潜力。具体而言，我们采用 Best-of-$N$ (BoN) 推理 [26] 来生成并评估多个候选解，使检测智能体能够减少随机误差，并更好地发挥其内在推理能力。我们采用 BoN，是因为它与 AgentM³D 的分层级联天然契合。在每个检测阶段选择最可靠的候选项，BoN 能够得出更清晰的中间决策，并避免聚合异构模态证据 [26]所造成的噪声累积。此外，其并行且非迭代的特性也使其能够与我们的提前停止和规划机制无缝集成，从而实现高效、稳健的测试时推理。采用批判感知排序的 Best-of-$N$ 推理整体流程见附录 A.2

**批判感知的 Best-of-$N$ 推理**。对于每个检测智能体$k \in$ {text、图像、跨模态}，我们为其配置 BoN 测试时扩展策略，以提高推理的稳健性。具体而言，在固定提示配置 P 下，会并行多次调用该智能体，以生成一组多样化的推理候选项：

$$\{(y_k^{(n)}, r_k^{(n)})\}_{n=1}^{N} \sim \phi_k(\mathcal{P}). \tag{10}$$

每个候选首先由预训练奖励模型 R（例如 [15, 37]）进行评估，该模型根据预测标签及其相关推理分配一个标量分数：

$$u_k^{(n)} = \mathcal{R}\Big(y_k^{(n)}, r_k^{(n)}\Big), \tag{11}$$

其中， $u_k^{(n)}$表示归一化奖励分数，数值越高，表示预测越可靠。

尽管奖励模型能够对推理质量提供反馈，但它们通常基于通用问答或指令跟随数据进行训练，可能无法完全满足虚假信息检测对不同模态有效性的特定要求。具体而言，这些奖励模型可能无法识别文本或视觉声明中的事实矛盾，只能对标签预测与其对应推理解释之间的语义连贯性作出表面判断。

为解决这一局限，我们进一步引入用于文本和视觉检测的模态特定批评智能体，利用与任务相关的外部信号明确验证模态层面的有效性。形式上，批评智能体定义如下：

$$\begin{aligned} q_{\text{text}}^{(n)} &\sim g_{\text{text}}\left(y_{\text{text}}^{(n)}, r_{\text{text}}^{(n)}, \mathcal{T}_{\text{logic}}, \mathcal{P}_{\text{text}}^{\text{crit}}\right), \\ q_{\text{image}}^{(n)} &\sim g_{\text{image}}\left(y_{\text{image}}^{(n)}, r_{\text{image}}^{(n)}, \mathcal{T}_{\text{forgery}}, \mathcal{P}_{\text{image}}^{\text{crit}}\right), \end{aligned} \tag{12}$$

其中，$q_k^{(n)}$ 表示分配给第 $n$个候选项的批评分数。$\mathrm{T}_{\text{logic}}$ 和$\mathrm{T}_{\text{forgery}}$ 分别是逻辑一致性检查工具和图像伪造检测工具的输出（详情见附录A.3），而$\mathrm{P}_k^{\text{crit}}$ 表示批评提示。每个批评智能体都会评估预测标签及其对应的推理，提供一种正交的、模态特定的事实有效性信号，以补充基于奖励的评估。

请注意，我们没有为跨模态一致性检测引入批评智能体。这一设计选择源于跨模态一致性关注的是模态之间的语义兼容性，而不是验证单条内容的事实正确性。因此，可以通过评估预测标签与其对应推理之间的语义连贯性，有效评估跨模态预测的可靠性；对此，通用奖励模型能够提供足够的偏好信号。我们还在第4.4节通过实证研究考察了跨模态检测是否需要使用批评智能体，结果显示性能有所下降，这可能是由于引入了冗余或噪声信号。

最后，我们将基于奖励的信号与基于批评的信号整合到统一的评分函数中，以指导BoN选择：

$$s_k^{(n)} = \begin{cases} u_k^{(n)} + q_k^{(n)}, & k \in \{\text{text}, \text{image}\}, \\ u_k^{(n)}, & k = \text{cross}, \end{cases} \tag{13}$$

其中 $u_k^{(n)}$ 和 $q_k^{(n)}$ 在融合前被独立归一化到同一尺度。所得分数 $s_k^{(n)}$ 能够对推理候选项进行一致且具备事实感知的排序，从而引导BoN推理趋向于更能满足特定模态虚假信息检测标准的预测。

**基于Top-$m$ 平均间隔的提前停止排序**。 在BoN推理下评估全部 $N$ 个候选项可能计算效率低下，尤其是当一个明显占优的解在排序早期就出现时。为了在不牺牲可靠性的前提下提升测试时效率，我们引入一种自适应提前停止机制，该机制基于排名最高的候选项与其竞争者之间的相对分数间隔。

令$\mathbf{s}_k = (s_{(1)k}, \ldots, s^{(N)}{}_k)$ 表示为智能体 $k$ 的推理候选项生成的融合分数。我们将这些分数按降序排列，得到 $\mathbf{s}_{k,\downarrow}$，并将 $\mathbf{s}_k^{(1:m)}{}_{,\downarrow}$ 表示为排名前 $m$ 的分数。

为量化首位候选项的置信度，我们将 Top-$m$ 平均差距定义为

$$\Delta_m(\mathbf{s}_k) = s_{k,\downarrow}^{(1)} - \frac{1}{m-1}\sum_{j=2}^{m} s_{k,\downarrow}^{(j)}, \qquad m \geq 2. \tag{14}$$

该指标衡量排名最高的候选项与其余头部竞争者的分离程度。

停止点由满足以下条件的最小 $m$确定，即排名最高的候选项充分优于其他候选项：

$$m_k^* = \min\left\{m \in \{2, \ldots, N\} \mid \Delta_m(\mathbf{s}_k) > \tau\right\}, \tag{15}$$

其中 $\tau$是预定义的置信度阈值。若不存在满足该条件的 $m$，则令 $m_k^* = N$，对应于完整评估。 $\tau$ 对检测性能和推理效率的影响在第4.5.1节中进行了系统分析。

最后，从排名前$m_k^*$ 个候选项中选出融合分数最高的候选项，作为智能体 $k$的输出：

$$n_k^* = \arg\max_{n \in \{1, \ldots, m_k^*\}} s_{k,\downarrow}^{(n)}, \qquad (y_k^*, r_k^*) = (y_k^{(n_k^*)}, r_k^{(n_k^*)}). \tag{16}$$

选定的配对 $(y_k^*,\ r_k^*)$ 作为智能体 $k$的最终输出，随后用于式 9 中的分层激活与决策。

## 3.3 自适应测试时扩展规划

尽管感知批评的 Best-of-$N$ 推理提升了检测性能，但将其统一应用于所有输入在计算上效率低下，而且往往没有必要，因为许多多模态样本只需一次前向传播即可可靠地完成判断。因此，我们引入自适应规划机制，根据输入难度有选择地分配测试时计算。

具体而言，AgentM$^3$D 纳入了一个轻量级规划智能体，用于判断是否应针对给定输入样本启用增强推理。由于规划智能体只需粗略评估样本难度，而非进行精细的事实核验，因此将其实现为一种无需工具调用的轻量级、基于提示的决策模块。形式上，给定一个文本-图像对 $(T,\ V)$，规划智能体生成一个推理动作 A，如下所示：

$$\mathcal{A} \sim \phi_{\text{plan}}\left(T, V, \mathcal{P}_{\text{plan}}\right), \tag{17}$$

其中，$\phi_{\text{plan}}$ 表示由提示 $\mathrm{P}_{\text{plan}}$参数化的规划智能体，A 则指定下游检测智能体所采用的推理策略（即标准单次推理或增强型 Best-of-$N$ 推理）。

## 3.4 检测智能体推理的概率解释

从概率视角来看，每个检测智能体的推理都可以解释为在给定输入和上游检测结果的条件下进行的近似后验推理。具体而言，对于一个检测智能体 $k \in$ {text、图像、跨模态}，我们将其输出建模为一个条件分布：

$$p_k(y, r \mid x_k, \mathbf{y}_{<k}) \propto \exp\left(\beta S_k(y, r \mid x_k, \mathbf{y}_{<k})\right), \tag{18}$$

其中，$x_k$表示智能体特定输入（例如 $T$、$V$ 或 $(T, V)$），$\mathbf{y}_{<k}$ 表示该层级中所有先前智能体产生的检测结果。指数形式会诱导出一个由评分函数参数化的偏好分布。参数 $\beta > 0$ 控制该分布向高分候选项集中的程度，从而在利用与探索之间进行权衡。

评分函数 $S_k(\cdot)$ 综合了基于奖励和基于批评的信号：

$$S_k(y, r \mid x_k, \mathbf{y}_{<k}) = \mathcal{R}(y, r) + q_k(y, r, \mathcal{T}_k), \tag{19}$$

其中 $\mathcal{R}$表示奖励模型分数，而 $q_k$ 是相应批评智能体根据工具观测 $\mathcal{T}_k$产生的批评分数。

在这一表述下，BoN 对应于从诱导的后验分布中进行随机采样，而批判感知排序则对采样得到的候选项执行近似最大后验估计。分层激活机制进一步施加硬门控先验，确保每个智能体只有在被先前的推理结果激活时，才会对最终决策作出贡献。

## 4 实验

在本节中，我们开展实验以验证 AgentM$^3$D 的有效性，并旨在回答以下研究问题（RQs）。（RQ1）与强大的基于视觉语言模型的基线和近期用于多模态虚假信息检测的智能体方法相比，AgentM$^3$D 的表现如何？（RQ2）AgentM$^3$D 在检测性能和推理效率之间的平衡效果如何？与现有的多智能体方法相比，它能否实现更优的准确率—延迟权衡？（RQ3）AgentM$^3$D 的核心组件，包括自适应测试时扩展、批评感知推理和结构化智能体协同，分别对整体检测性能有何贡献？（RQ4）AgentM$^3$D 中的超参数如何影响性能和效率？（RQ5）通过对正确分类和错误分类实例的案例研究，所提出的批评智能体和自适应 Best-of-$N$ 推理在实践中如何影响决策过程？

### 4.1 实验设置

**数据集**。我们在两个混合来源多模态虚假信息检测（M$^3$D）基准测试上评估 AgentM$^3$D。首先，我们采用*MMFakeBench*，其中包含超过 11,000 个图文对，涵盖四个类别：真实样本、文本真实性失真（TVD）、视觉真实性失真（VVD）和跨模态一致性失真（CMM），图像既来自人工生成，也来自机器生成 [16]。遵循先前基于智能体的工作 MMD-Agent 和 T$^2$Agent [5, 16], ，我们使用相同的评估协议，并抽取相同的 1,000 个验证实例子集，均衡分布为 300 个真实样本、300 个 TVD、100 个 VVD 和 300 个 CMM 样本。

为进一步评估混合来源场景下的泛化能力，我们通过整合三个互补数据集中的样本，构建了一个组合 基准：*Mocheg*，一个以文本为中心的虚假信息数据集，包含经记者核验的声明 [38]；*Fakeddit-M*，一个基于 Reddit、聚焦于篡改视觉内容的数据集 [17]；以及*VERITE*，一个真实世界基准，包含模态均衡的图像–文本配对 [19]。我们从每个数据集中随机抽取 100 个真实样本实例和 200 个虚假实例，最终构建出一个统一基准，其中包含 300 个真实样本、200 个 TVD 样本、200 个 VVD 样本和 200 个 CMM 样本。这种组合设置能够在多样的内容分布下系统评估 M$^3$D。

**基线方法**。我们将 AgentM$^3$D 与一组多样化的基线方法进行比较，这些基线涵盖三种代表性范式：基于视觉语言模型的单模型推理、测试时扩展策略，以及智能体式多模态检测框架。这些基线使我们能够在可比设置下同时评估绝对检测性能，以及不同推理与扩展机制的有效性。

- 标准：一种基础视觉语言模型基线，通过单次前向传播，直接根据文本-图像输入预测标签。
- Best-of-$N$ (BoN) [8]：一种并行的测试时扩展策略，从视觉语言模型中采样多个检测输出，并使用奖励模型选择最终预测。
- T$^2$智能体 [5]：一种智能体方法，将多模态虚假信息检测建模为树搜索问题，并应用蒙特卡洛树搜索来引导多步推理。
- MMD-Agent[16]：一种智能体框架，将多模态虚假信息检测分解为多个具有层级推理能力、专用于特定任务的智能体。
- MMD-Agent+BoN：MMD-Agent 的一种变体，其中所有检测智能体都通过 BoN 采样进行统一增强。

**实现细节**。 所有实验均在零样本设置下进行，未针对特定任务进行微调。我们报告标准分类指标，包括准确率、精确率、召回率和 F1 分数。我们采用两种视觉语言模型骨干网络 *Qwen3-VL-4B-Instruct*[1] 和 *Qwen3-VL-8B-Instruct*[2], 进行性能比较，其他所有实验则均使用 *Qwen3-VL-8B-Instruct*。遵循近期的智能体框架 [2,34], , 我们只使用 Qwen-VL 系列的模型，因为在本研究开展时，它们是可用的最强开源视觉语言模型，且参数规模对环境友好 [13]。所有实验均在单张 NVIDIA RTX 4090 GPU 上完成。对于 Best-of-$N$推理，我们设置 $N = 5$，并在第 4.5.2 节分析超参数 $N$对检测性能的影响。其他所有视觉语言模型相关的推理参数均采用 MMD-Agent [16]的默认配置。对于所有涉及 Best-of-$N$的实验，我们使用 *GRM-Gemma2-2B-rewardmodel-ft* [37] 作为候选项选择的奖励模型。AgentM$^3$D 中所提智能体的提示设计见附录 A.1。

### 4.2 性能比较（RQ1）

我们在零样本设置下，使用两个 M$^3$D 基准测试评估所提出的 AgentM$^3$D，并将其与强大的、基于视觉语言模型和智能体的最先进方法进行比较。定量结果汇总于表 1。根据这些结果，我们得出以下观察：

- 与基于视觉语言模型的基线和现有智能体方法相比，AgentM$^3$D 在几乎所有评估指标和基准测试中都取得了最佳表现，展现了其在不同内容分布下进行零样本 M$^3$D 的有效性。

[1]https://huggingface .co/Qwen/Qwen3 -VL-4B-指令[2]https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct

**表1：在MMDFake和组合基准测试上的性能比较。所有结果均以准确率、F1值、召回率（召回率）和精确率（精确率）报告。"–"表示结果不可用。**

| 骨干网络 | 方法 | MMFakeBench | | | | 组合 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Acc | F1 | Rec | Pre | Acc | F1 | Rec | Pre |
| Qwen3-VL-4B | 标准 | 42.9 | 29.2 | 35.8 | 35.8 | 30.3 | 23.6 | 31.0 | 42.3 |
| | BoN | 43.7 | 31.1 | 36.4 | 47.9 | 28.2 | 21.9 | 29.5 | 40.5 |
| | T²智能体 | 50.1 | 50.3 | 49.4 | 54.6 | 35.4 | 35.6 | 38.2 | 45.7 |
| | MMD-Agent | 55.2 | 55.4 | 55.8 | 57.1 | 41.9 | 40.9 | 44.1 | 48.5 |
| | MMD-Agent + BoN | 57.4 | 57.8 | 58.6 | **58.5** | 40.6 | 39.7 | 42.7 | 48.6 |
| | **智能体M³D（我们的方法）** | **58.1** | **58.0** | **60.0** | 57.1 | **45.4** | **45.6** | **47.3** | **49.0** |
| Qwen3-VL-8B | 标准 | 46.9 | 37.0 | 39.4 | 59.9 | 33.6 | 28.9 | 36.0 | 40.6 |
| | BoN | 45.7 | 35.6 | 38.4 | 62.5 | 33.6 | 28.4 | 36.3 | 42.9 |
| | T²智能体 | 54.3 | 54.0 | 52.0 | 61.3 | 36.2 | 36.1 | 38.8 | 45.5 |
| | MMD-Agent | 59.4 | 60.2 | 60.3 | 62.5 | 43.3 | 43.5 | 45.2 | 50.5 |
| | MMD-Agent + BoN | 60.1 | 60.7 | 60.4 | **62.9** | 42.3 | 42.6 | 44.3 | 48.7 |
| | **AgentM³D（我们的方法）** | **62.0** | **62.6** | **64.2** | 62.1 | **48.1** | **48.3** | **50.5** | **52.4** |

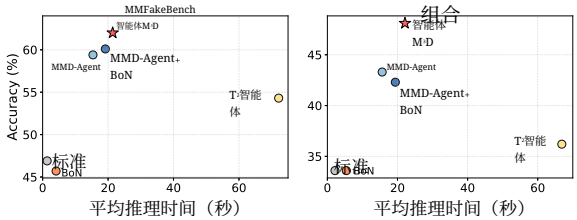



**图3：AgentM³D 与基线方法的推理效率对比。**

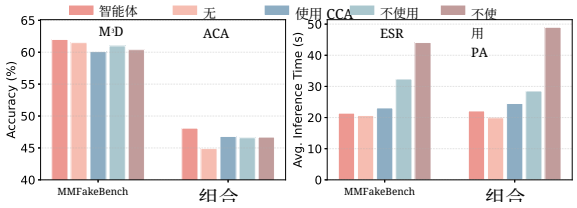



**图 4：AgentM³D 各关键组件的影响。**

- 仅通过增强标准视觉语言模型推理或现有智能体框架并采用 BoN 采样，并不能稳定提升性能，而且经常会出现性能下降。这表明，将 BoN 直接与视觉语言模型推理或智能体系统相结合，不足以保证获得可靠的收益。相比之下，AgentM³D 引入了批评智能体，明确评估并筛选候选预测，从而弥补了简单使用 BoN 的不足，并使其在不同设置下均能稳定提升性能。

## 4.3 推理效率分析（RQ2）

我们在零样本设置下，从准确率—延迟权衡的角度，将AgentM³D 的推理效率与基线方法进行比较，如图3所示。在MMFakeBench 和 Combined基准上，基于朴素测试时扩展的方法（包括 T²Agent）相比其原始版本，推理延迟均显著更高，而准确率提升有限或不稳定。相比之下，AgentM³D 以适度增加推理时间为代价取得了最高准确率，因此相较于标准视觉语言模型推理和现有智能体基线，处于更有利的权衡区域。值得注意的是，所提出的规划智能体仅针对部分样本选择性地触发 Best-of-$N$ 推理，

其中，MMFakeBench上69.1%的实例以及Combined基准上77.2%的实例被识别为需要增强推理。这些结果表明，自适应测试时扩展与批判感知候选选择相结合，使AgentM³D能够提升在不过度增加计算开销的情况下提高检测可靠性，使其非常适用于实际的 M³D。

## 4.4 消融研究（RQ3）

我们在MMFakeBench和Combined基准上对AgentM³D进行消融研究，以评估其关键组件的贡献（图4）。移除提前停止排序（w/o ESR）会显著增加推理延迟，而准确率仅有微小变化，证明了其对效率提升的有效性。禁用所有批判智能体（w/o ACA）会持续降低准确率，尤其是在Combined基准上，凸显了批判感知评估对于可靠M³D的重要性。相比之下，额外增加一个跨模态批判智能体（w/ CCA）会导致性能略有下降，表明过度批判的收益递减。最后，移除规划智能体（w/o PA）会大幅增加推理时间且没有准确率提升，强调了规划在高效测试时推理中的作用。总体而言，这些结果验证了AgentM³D中规划、批判感知最佳选择与提前停止的联合设计。

## 4.5 超参数分析（RQ4）

*4.5.1* 超参数的影响 $\tau$ 在提前停止排序中的影响。我们分析了早停阈值 $\tau$ 对检测性能和推理行为的影响。如图5所示，提高 $\tau$ 会减少触发提前停止的样本比例，从而进行更全面的 BoN评估。随着 $\tau$ 从 0.1 增至 0.5，准确率有所提升，这表明此时处于一种平衡状态：既能抑制低质量候选项，又不会过度限制候选探索；而进一步将 $\tau$ 提高至 0.7，则不会带来额外收益，

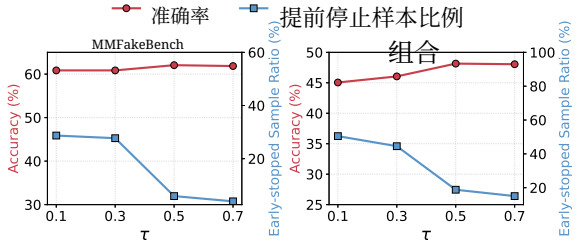



**图 5：早停阈值 $\tau$ 对检测准确率和提前停止样本比例的影响。**

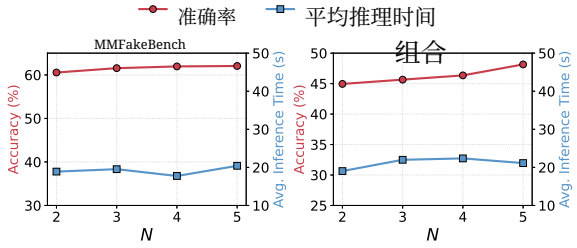



**图 6：BoN推理中的 $N$ 对检测准确率和推理时间的影响。**

还可能引入排序噪声。总体而言， $\tau$ 能够有效控制准确率–效率权衡，适中的取值在实践中可实现最理想的平衡。

*4.5.2 $N$ 在 BoN 推理中的影响。* 如图 6 所示，我们对 BoN 推理中使用的候选项数量 $N$ 进行了敏感性分析。与先前研究一致，增加 $N$ 通常能够通过扩大对推理轨迹的探索范围 [26] 来提升检测准确率。遵循采用 BoN 推理的智能体方法 [12], 中的常见做法，我们选择$N = 5$ 作为一种平衡设置，在带来可靠性能提升的同时，也能满足我们实验环境下的计算成本要求。由于规划智能体只对一部分样本启用 BoN 推理，因此增加 $N$ 不会导致计算量成比例增加。

## 4.6 案例研究（RQ5）

图7展示了组合基准中一个具有代表性的文本真实性失真案例。尽管 MMD-Agent+BoN 和 AgentM$^3$D 都生成了多个预测各异的推理候选项，但二者的决策机制存在显著差异。对于 MMD-Agent+BoN，最终预测由一条高分推理路径主导，而这条路径错误地支持了该主张；因此，尽管存在反驳性候选项，系统仍做出了错误判断。相比之下，AgentM$^3$D 采用批判感知评估，降低缺乏依据的推理路径的权重，并突出兼具高置信度且符合批评意见的候选项。AgentM$^3$D 通过汇总可靠的推理轨迹，而非仅依赖原始 BoN 评分，在此案例中正确识别出文本真实性失真。

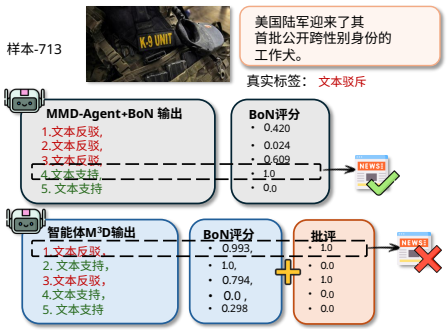



**图7：Combined基准上的案例研究 .**

## 5 相关工作

### 5.1 多模态虚假信息检测

多模态虚假信息检测（MMD）已得到广泛研究，相关方法要么分析不同模态间的内部不一致，要么借助外部证据核验主张。以往研究通过建模跨模态相关性、特定模态语义或视觉伪影，检测文本、视觉及跨模态失真 [9–11, 21, 30, 31, 39], ；另一些研究则利用网络检索、实体匹配或经验证的资料库开展基于证据的核验 [20, 35]。近期进展进一步利用多模态大语言模型（LLMs）和分层推理来提升鲁棒性 [3, 24]。然而，大多数方法都假设操纵源单一，难以泛化到复杂的混合型虚假信息。为此，近期出现了面向零样本 M$^3$D 的智能体框架。MMD-agent [16] 采用级联式多智能体设计，依次执行文本、视觉和跨模态检测；T$^2$Agent [5] 则将智能体检测框架构建为搜索问题，并采用蒙特卡洛树搜索。尽管这些方法提升了零样本能力，但仍存在测试时成本高、推理模块耦合松散的问题。与此同时，TRUST-VL [36] 通过在大规模标注数据上进行微调来检测混合来源虚假信息，因此难以应用于本研究所关注的零样本设置。为此，我们提出 AgentM$^3$D，这是一种决策感知多智能体框架，通过自适应测试时扩展增强零样本 MMD。AgentM$^3$D 融合轻量级规划、具备批判感知能力的 Best-of-$N$以及自适应提前停止，从而有选择地分配计算资源，实现稳健且高效的推理。

### 5.2 基于LLM的多智能体系统 s

近期研究主张从单体模型转向由多个专用组件构成的复合系统，多智能体框架正逐渐成为这一理念的重要实践形式。基于大语言模型的多智能体系统已在多个领域得到探索，包括叙事生成、金融交易和协作式问题求解；与单智能体系统相比，它们展现出更强的推理、规划和决策能力 [6, 7, 23, 33]。为组织智能体间的协作，研究人员探索了多种组织拓扑结构，例如顺序管线、集中式辐射结构设计、层次树结构以及基于图的交互模式 [25, 32]。近期框架则进一步

强调采用分层或按角色专门化的组织方式，以有效协调智能体行为。尽管取得了成功，现有多智能体系统仍未能为零样本多模态虚假信息检测提供令人满意的智能体式解决方案，无法妥善应对跨模态错误传播的缓解，以及以有原则的方式调控测试时计算等挑战。在这项工作中，我们提出 AgentM³D，这是一种专为多模态虚假信息检测设计的多智能体框架，旨在解决现有多智能体系统在混合来源和零样本设置下的局限性。

## 6 结论

在本研究中，我们提出了 AgentM³D，这是一种用于混合来源多模态虚假信息检测的多智能体框架，兼顾可靠性与测试时效率。通过整合自适应测试时扩展与具备批评意识的 Best-of-$N$ 推理，AgentM³D 可依据置信度信号有选择地分配计算资源，在维持稳健检测性能的同时避免不必要的候选探索。在 MMFakeBench 和新构建的混合来源基准上开展的大量实验表明，在零样本设置下，AgentM³D 始终优于强大的基于视觉语言模型的基线和智能体基线，同时实现了良好的准确率–效率权衡。总体而言，AgentM³D 为在现实计算约束下部署可靠的多模态虚假信息检测系统提供了实用且可扩展的基础。

## 参考文献

[1] Sara Abdali、Sina Shaham 和 Bhaskar Krishnamachari。2024年。多模态虚假信息检测：方法、挑战与机遇。《计算调查》57, 3 (2024年), 1–29。[2] Hao Bai、Alexey Taymanov、Tong Zhang、Aviral Kumar 和 Spencer Whitehead。2026年。WebGym：通过现实任务扩展视觉网页智能体的训练环境。*arXiv* 预印本 *arXiv:2601.02439* (2026年)。[3] Alimohammad Beigi、Bohan Jiang、Dawei Li、Zhen Tan、Pouya Shaeri、Tharindu Kumarage、Amrita Bhattacharjee 和 Huan Liu。2024年。大语言模型能否通过提出相关问题改进多模态事实核查？*arXiv* 预印本 *arXiv:2410.04616* (2024年)。[4] Zhe Chen、Weiyun Wang、Yue Cao、Yangzhou Liu、Zhangwei Gao、Erfei Cui、Ji nguo Zhu、Shenglong Ye、Hao Tian、Zhaoyang Liu 等。2024年。通过模型、数据和测试时扩展，拓展开源多模态模型的性能边界。*arXiv* 预印本 *arXiv:2412.05271* (2024年)。[5] Xing Cui、Yueying Zou、Zekun Li、Peipei Li、Xinyuan Xu、Xuannan Liu 和 Huaibo Huang。2025年。Tˆ 2Agent：一种结合蒙特卡洛树搜索的工具增强型多模态虚假信息检测智能体。*arXiv* 预印本 *arXiv:2505.19768* (2025年)。[6] Yilun Du、Shuang Li、Antonio Torralba、Joshua B Tenenbaum 和 Igor Mordatch。2023年。通过多智能体辩论提升语言模型的事实性与推理能力。载于第四十一届国际机器学习大会。[7] Taicheng Guo、Xiuying Chen、Yaqi Wang、Ruidi Chang、Shichao Pei、Nitesh V Chawla、Olaf Wiest 和 Xiangliang Zhang。2024年。基于大语言模型的多智能体：进展与挑战综述。*arXiv* 预印本 *arXiv:2402.01680* (2024年)。[8] Robert Irvine、Douglas Boubert、Vyas Raina、Adian Liusie、Ziyi Zhu、Vineet Mudupalli、Aliaksei Korshuk、Zongyi Liu、Fritz Cremer、Valentin Assassi 等。2023年。通过与数百万用户的真实场景互动奖励聊天机器人。*arXiv* 预印本 *arXiv:2303.06135* (2023年)。[9] Wei Jiang、Tong Chen、Xinyi Gao、Wentao Zhang、Lizhen Cui 和 Hongzhi Yin。2025年。用于稳健谣言检测的流行病学信息网络。载于*2025*年 *ACM* 万维网大会论文集。3618–3627。[10] Wei Jiang、Tong Chen、Wei Yuan、Xiangyu Zhao、Quoc Viet Hung Nguyen 和 Hongzhi Yin。2025年。面向有监督社交媒体图分析的传播感知表示学习。*arXiv* 预印本 *arXiv:2509.01124* (2025年)。[11] Dhruv Khattar、Jaipal Singh Goud、Manish Gupta 和 Vasudeva Varma。2019年。Mvae：用于假新闻检测的多模态变分自编码器。载于万维网大会。2915–2921。[12] Bingxuan Li、Yiwei Wang、Jiuxiang Gu、Kai-Wei Chang 和 Nanyun Peng。2025年。Metal：一种采用测试时扩展的图表生成多智能体框架。*arXiv* 预印本 *arXiv:2502.17651*(2025年)。[13] Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu 等。2026年。Qwen3-VL-Embedding 和 Qwen3-VL-Reranker：用于最先进水平多模态检索与排序的统一框架。arXiv 预印本 arXiv:2601.04720（2026年）。[14] Hongzhan Lin, Pengyao Yi, Jing Ma, Haiyun Jiang, Ziyang Luo, Shuming Shi 和 Ruifang Liu。2023年。通过提示学习和传播结构进行零样本谣言检测。载于《人工智能促进协会年会论文集》第37卷。5213–5221。[15] Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, Yang Liu 和 Yahui Zhou。2025年。Skywork-Reward-V2：通过人机协同扩展偏好数据整理。arXiv 预印本 arXiv:2507.01352（2025年）。[16] Xuannan Liu, Zekun Li, Peipei Li, Huaibo Huang, Shuhan Xia, Xing Cui, Linzhi Huang, Weihong Deng 和 Zhaofeng He。2024年。MMFakeBench：面向大型视觉语言模型的混合来源多模态错误信息检测基准。arXiv 预印本 arXiv:2406.08772（2024年）。[17] Kai Nakamura, Sharon Levy 和 William Yang Wang。2020年。Fakeddit：用于细粒度假新闻检测的新型多模态基准数据集。载于《第十二届语言资源与评测会议论文集》。6149–6157。[18] Cailin O'Connor 和 James Owen Weatherall。2019年。《错误信息时代：错误信念如何传播》。耶鲁大学出版社。[19] Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos 和 Panagiotis C Petrantonakis。2024年。VERITE：一种考虑单模态偏差的稳健多模态错误信息检测基准。《国际多媒体信息检索期刊》13，1（2024年），4。[20] Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos 和 Panagiotis C Petrantonakis。2025年。Red-dot：通过相关证据检测进行多模态事实核查。《IEEE 计算社会系统汇刊》（2025年）。[21] Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos 和 Panagiotis C Petrantonakis。2025年。相似性优于事实性：我们在多模态脱离语境错误信息检测方面取得进展了吗？载于《2025年 IEEE/CVF冬季计算机视觉应用会议（WACV）》。IEEE，5041–5050。[22] Gordon Pennycook 和 David G Rand。2019年。利用众包对新闻来源质量的判断，打击社交媒体上的错误信息。《美国国家科学院院刊》116，7（2019年），2521–2526。[23] Aske Plaat, Max van Duijn, Niki van Stein, Mike Preuss, Peter van der Putten 和 Kees Joost Batenburg。2025年。智能体大语言模型综述。arXiv 预印本 arXiv:2503.23037（2025年）。[24] Peng Qi, Zehong Yan, Wynne Hsu 和 Mong Li Lee。2024年。Sniffer：用于可解释的脱离语境错误信息检测的多模态大语言模型。载于《IEEE/CVF计算机视觉与模式识别会议论文集》。13052–13062。[25] Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong 等。2024年。Chatdev：用于软件开发的交互式智能体。载于《计算语言学协会第62届年会论文集》（第一卷：长篇论文）。15174–15186。[26] Pier Giuseppe Sessa, Robert Dadashi, Léonard Hussenot, Johan Ferret, Nino Vieillard, Alexandre Ramé, Bobak Shariari, Sarah Perrin, Abe Friesen, Geoffrey Cideron 等。[n. d.]。Bond：通过优选蒸馏对齐大语言模型，2024年。网址：https://arxiv. org/abs/2407.14622（[n. d.]）。[27] Lanyu Shang, Yang Zhang, Bozhang Chen, Ruohan Zong, Zhenrui Yue, Huimin Zeng, Na Wei 和 Dong Wang。2024年。MMAdapt：一种用于早期健康错误信息检测的知识引导型多来源多类别领域自适应框架。载于《2024年ACM Web 大会论文集》。4653–4663。[28] Weizhou Shen, Chenliang Li, Hongzhan Chen, Ming Yan, Xiaojun Quan, Hehong Chen, Ji Zhang 和 Fei Huang。2024年。小型大语言模型是弱工具学习者：一种多大语言模型智能体。arXiv 预印本 arXiv:2401.07324（2024年）。[29] Lisa Singh, Shweta Bansal, Leticia Bode, Ceren Budak, Guangqing Chi, Korn- raphop Kawintiranon, Colton Padden, Rebecca Vanarsdall, Emily Vraga 和 Yanchen Wang。2020年。首次审视 Twitter 上新冠肺炎相关信息与错误信息的分享情况。arXiv 预印本 arXiv:2003.13907（2020年）。[30] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su 和 Jing Gao。2018年。Eann：用于多模态假新闻检测的事件对抗神经网络。载于《第24届ACM SIGKDD知识发现与数据挖掘国际会议论文集》。849–857。[31] Zongwei Wang, Min Gao, Junliang Yu, Tong Chen 和 Chenghua Lin。2026年。PAMAS：用于虚假信息检测、具备视角聚合能力的自适应多智能体系统。arXiv 预印本 arXiv:2602.03158（2026年）。[32] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu 等。2024年。Autogen：通过多智能体对话实现新一代大语言模型应用。载于首届语言建模会议。[33] Yijia Xiao, Edward Sun, Di Luo 和 Wei Wang。2024年。TradingAgents：多智能体大语言模型金融交易框架。arXiv 预印本 arXiv:2412.20138（2024年）。[34] Yong Xien Chng, Tao Hu, Wenwen Tong, Xueheng Li, Jiandong Chen, Haojia Yu, Jiefan Lu, Hewei Guo, Hanming Deng, Chengjun Xie 等。2025年。SenseNova-MARS：通过强化学习赋能多模态智能体推理与搜索。arXiv 电子印本（2025年），arXiv–2512。

[35] Kaiying Yan, Moyang Liu, Yukun Liu, Ruibo Fu, Zhengqi Wen, Jianhua Tao 和 Xuefei Liu。2025年。揭穿与推断：通过扩散生成证据和大语言模型推理进行多模态假新闻检测。*arXiv* 预印本 *arXiv:2506.21557*（2025年）。[36] Zehong Yan, Peng Qi, Wynne Hsu 和 Mong-Li Lee。2025年。TRUST-VL：用于通用多模态错误信息检测的可解释新闻助手。载于*2025*年自然语言处理实证方法会议论文集。5588–5604。[37] Rui Yang, Ruomeng Ding, Yong Lin, Huan Zhang 和 Tong Zhang。2024年。通过正则化隐藏状态，学习适用于大语言模型的泛化奖励模型。载于神经信息处理系统进展。[38] Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho 和 Lifu Huang。2023年。端到端多模态事实核查与解释生成：一个具有挑战性的数据集和模型。载于第*46*届*ACM*国际信息检索研究与发展会议论文集。2733–2743。[39] Junwei Yin, Min Gao, Kai Shu, Wentao Li, Yinqiu Huang 和 Zongwei Wang。2025年。图与序列：面向假新闻的广域语义建模检测。在*ACM Web Conference 2025*论文集中。2838–2849。[40] Jingyi Zhang、Jiaxing Huang、Sheng Jin 和 Shijian Lu。2024年。视觉语言模型在视觉任务中的应用：综述。*IEEE*模式分析与机器智能汇刊46, 8（2024年），5625–5644。[41] Qiyuan Zhang、Fuyuan Lyu、Zexu Sun、Lei Wang、Weixu Zhang、Wenyue Hua、Haolun Wu、Zhihan Guo、Yufei Wang、Niklas Muennighoff 等。2025年。大型语言模型中的测试时扩展综述：是什么、如何实现、应用于何处，以及效果如何？*arXiv* 预印本 *arXiv:2503.24235*（2025年）。[42] Nan Zhong、Yiran Xu、Sheng Li、Zhenxing Qian 和 Xinpeng Zhang。2023年。Patchcraft：探索纹理块，实现高效的AI 生成图像检测。*arXiv* 预印本 *arXiv:2311.12397*（2023年）。[43] Ziyi Zhou、XiaomingZhang、Litian Zhang、JiachengLiu、Senzhang Wang、Zheng Liu、Xi Zhang、Chaozhuo Li 和 Philip S Yu。2024年。Finefake：用于细粒度多领域假新闻检测的知识增强数据集。*arXiv* 预印本 *arXiv:2404.01336*（2024年）。

# A 附录

## A.1 所提出智能体的详细提示

**规划智能体** 给定一个多模态虚假信息样本，其中包含一条新闻标题和一张新闻图像。新闻标题为：<新闻标题>你的任务是判断此样本应采用标准推理，还是在后续阶段使用最佳 N 选一（BON）提升推理级别。最佳 N 选一（BON）是指使用相同提示对多个独立的检测响应进行采样，并选出其中最可靠的响应。请从以下方面分析给定的新闻标题和图像：- 标题与图像之间的关系是明显一致、明显不一致，还是模棱两可。- 标题是否提出了需要明确且具体的视觉证据才能证实的主张。- 仅凭图像内容是否足以验证这些主张。根据分析，从以下选项中选择一个操作：1. [BON level-0]：无需进行最佳 N 选一扩展。2. [BON level-1]：使用最佳 N 选一扩展。仅按以下确切格式返回操作： [BON level-n]

**文本评议智能体** 给定一条新闻标题，该新闻标题为：<新闻标题>你的任务是仅根据检测结果自身的推理，为其可信度打一个 0 到 1 之间的分数。检测结果：<检测结果>逻辑一致性检查工具的结果：<逻辑一致性检查结果>为检测结果打分，范围为 0 到 1：- 1.0：完全可信、逻辑严密，没有过度推断。- 0.5：部分可信，存在一些逻辑漏洞或支撑不足。- 0.0：不可信，存在逻辑矛盾或明显过度推断。仅输出 0 到 1 之间的分数。

**视觉评议智能体** 根据给定的新闻图像，你的任务是仅根据检测结果自身的推理，为其说服力打一个 0 到 1 之间的分数。检测结果：

<检测结果>

图像伪造检测工具的结果：<图像伪造检测结果>请将检测结果评分，范围为 0 到 1：- 1.0：图像取证结果有力支持该检测结果，没有明显的逻辑漏洞或过度解读。- 0.5：图像取证结果部分支持该检测结果，但存在不确定性、证据薄弱或轻微过度解读。

- 0.0：检测结果不受图像
取证结果支持，或显示出明显的逻辑不一致或严重的过度解读。
仅输出介于 0 和 1 之间的数字分数。

## A.2 算法

采用批评感知排序的 Best-of-$N$推理整体流程算法如算法 1 所示。

**算法 1：** 结合批判感知的Best-of-$N$ 推理排序。

**输入：** 智能体 $k$输入 $\mathrm{X}_k$，候选预算 $N$，阈值 $\tau$
**输出：** 选定的预测 $y^*$ 和推理 $r^*$
// 生成BoN候选项
示例 $N$推理候选项 $\{(y^{(n)},\ r(n))\}_{n=1}^{N} \sim \phi_k(\mathrm{X}_k)$
// 奖励与批评评分
**对于** $n \leftarrow 1$ **到** $N$ **执行**
  $u(n) \leftarrow \mathcal{R}(y^{(n)}, r(n))$
  **如果** $k \in$ {文本，图像} **那么**
    $q^{(n)} \sim g_k(y^{(n)}, r(n), \mathrm{T}_k)$
    $s^{(n)} \leftarrow u^{(n)} + q^{(n)}$
  **else**
    $s^{(n)} \leftarrow u^{(n)}$
// 基于 Top-$m$ Average Gap 的提前停止排序
按降序排列 $\{s^{(n)}\}_{n=1}^{N}$，得到 $\{s(j)\downarrow\}_{j=1}^{N}$
**对于** $m \leftarrow 2$ **到** $N$，**执行**
  $\Delta_m \leftarrow s_{\downarrow}^{(1)} - \frac{1}{m-1}\sum_{j=2}^{m} s(j)\downarrow$
  **if** $\Delta_m > \tau$ **则**
    **跳出**
// 在前$m$名中选择最佳候选项
$n^* \leftarrow \arg\max_{n\in\{1,\ldots,m\}} s_{\downarrow}^{(n)}$
**返回** $(y^{(n^*)}\ r(n^*))$

## A.3 工具集详情

为支持 AgentM³D 中的模态特定验证，我们采用一个轻量且可扩展的工具集，为虚假信息检测提供互补的外部信号。

**网络搜索**。 我们使用维基百科[3] 检索新闻文本中提及实体的百科知识，将其作为文本真实性评估的外部证据。

**文本逻辑一致性检查**。我们采用 Qwen3-4B-Instruct-2507[4] 作为逻辑一致性检查工具，分析文本主张是否存在逻辑矛盾或不合理的推理。

**图像伪造检测**。我们采用经过充分训练的图像伪造检测模型 [42]，识别视觉内容中潜在的数字篡改，为视觉真实性验证提供客观依据。

[3]https://www.wikipedia.org/ [4]
https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507

**图像分析**。我们使用 Qwen3-VL-8B-Instruct[5] 提取基于视觉依据的描述，用于视觉真实性评估和跨模态一致性检查。

[5]https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct