不确定性-o：用于揭示大型多模态模型不确定性的一个与模型无关的框架

Ruiyang Zhang [1] Hu Zhang[2] Hao Fei[3] Zhedong Zheng[1]*
[1] FST and ICI, University of Macau, China
[2] CSIRO Data61, Australia [3] National University of Singapore

Abstract

大型多模态模型（LMMs）利用多种模态之间的互补性，通常被认为比纯语言大型模型（LLMs）更具有鲁棒性；然而，LMMs 知道他们不知道的内容吗？仍然存在三个关键的开放性问题：(1) 如何以统一的方式评估多样化的 LMMs 的不确定性，(2) 如何提示 LMMs 显示其不确定性，以及 (3) 如何量化下游任务中的不确定性。为了解决这些挑战，我们引入了 Uncertainty-o：(1) 一个与模型无关的框架，旨在揭示 LMMs 的不确定性，无论其模态、架构或功能如何，(2) 对多模态提示扰动的实证探索，以揭示 LMMs 的不确定性，提供洞察和发现，以及 (3) 推导多模态语义不确定性的公式，使得能够从多模态响应中量化不确定性。跨越各种模态的 18 个基准和 10 个 LMMs（包括开源和闭源）的实验表明，Uncertainty-o 能够有效估计 LMMs 的不确定性，从而增强下游任务，例如幻觉检测、幻觉缓解和不确定性敏感的连锁思维推理。

1. 介绍

大型多模态模型（LMMs） [9, 11] 显著拓展了人造智能的边界。(1) 丰富的模态不断被整合。除了最初成功的“文本输入-文本输出”范式 [1] 之外，LMMs 现在还包含了图像 [46, 56] 、视频 [43, 47] 、音频 [39, 54] 、点云 [50, 74] 、惯性测量单元 [28] 、功能性磁共振成像 [30] ，以及更多 [67] 。(2) 多样的模型架构正在迅速演变。大型语言模型（LLMs） [1, 8] 和扩散模型（DMs） [31, 56] 的基础设计已被广泛改编 [46] 并集成 [68] ，以创造更先进和复杂的结构。(3) 独特的模型能力显著提升。LMMs 现在能够理解复杂的多模态上下文和指令 [30, 48] ，同时生成多样化和高保真度的内容 [67, 77] 。

然而，大型模型 [1, 46] 仍然遵循简单的”提示输入-响应输出”流程，与人类的思维过程相比，这一流程显得原始和粗糙 [4, 52] 。当被问及问题或给定任务时，人类不仅会提供答案或采取行动，还会经历内在的感受：信心或不确定性，轻松或困难 [18, 51] 。这些感受指导我们预测结果并相应地调整计划 [5, 10] 。为了

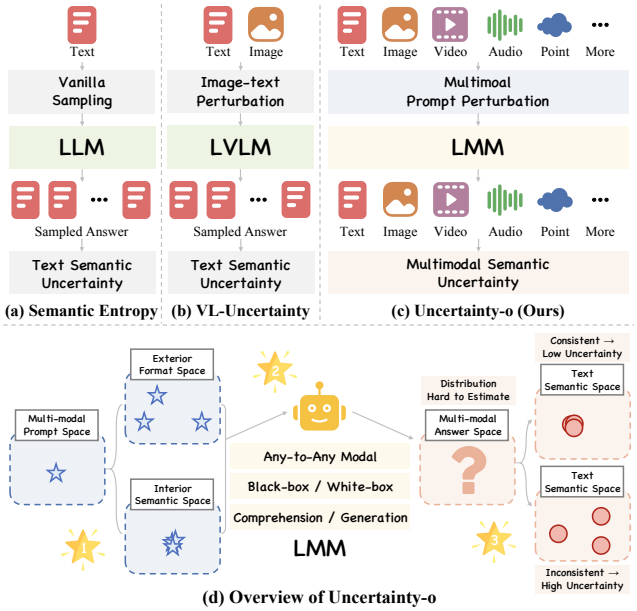

Figure 1. 与先前研究的比较及 Uncertainty-o 的概述。(a,b) 现有研究关注单一模态或双模态 [81] ；(c) 相比之下，我们的 Uncertainty-o 以模型无关的方式捕捉大型多模态模型中的不确定性。它通过多模态提示扰动实现可靠的不确定性估计。(d) 特别是，我们利用多模态语义不确定性，将多模态答案空间映射到统一的文本语义空间，然后根据文本一致性来估计不确定性。

弥合这一差距并探索 LMM 的心理学 [29, 37] ，我们专注于一个关键特性：LMM 的不确定性 [2, 57] 。

尽管 LMMs 表现强劲，但仍有三个关键开放性问题：(1) 如何以统一的方式评估不同 LMMs 的不确定性？对于不同的 LMMs 保持相同的评估标准具有挑战性。传统的解决方案是根据输入模态设计特定方案，但可能导致复杂且不灵活的框架，限制了引入新模态的适应性 [9, 11] 。(2) 如何提示 LMM 展示其不确定性？直接的方法是通过提示输入询问 LMM 的不确定性，但这通常会失败。这是因为包括 LLMs 在内的 LMMs 通常对其回应过于自信。(3) 我们如何为下游任务量化不确定性？线索嵌入在多模态响应 [39, 50, 56, 67] ，e.g. ，生成的图像 [56] 或点云 [40] 中，需要有效的技术来提取和量化 LMM 不确定性，以促进下游任务。

*Correspondence to zhedongzheng@um.edu.mo.

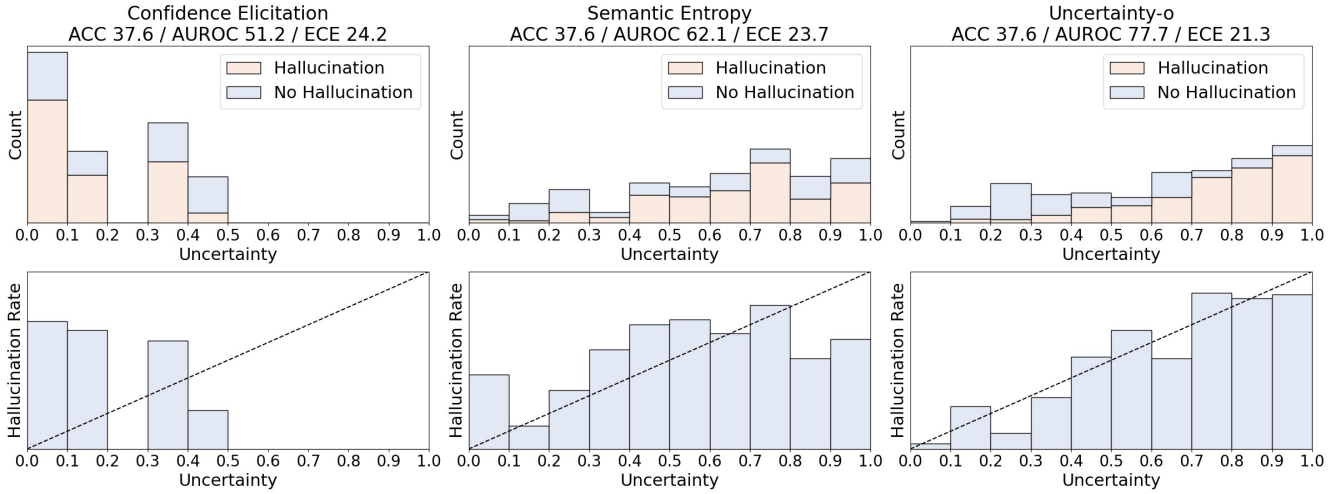

Figure 2. 我们估计不确定性的可靠性。通过与以前的方法比较，我们观察到：(1) 与信心引导 [71] 相比，Uncertainty-o 有效地避免了过度自信的问题，即对虚假的响应错误地分配了低不确定性（或高信心）。(2) 与语义熵 [24] 相比，Uncertainty-o 提供了更可靠的不确定性估计，e.g. ，不确定性更接近于每个区间中的错误率。来自 OneLLM [30] 在 ClothoV2 [21] 上的结果。

针对这些重要问题，我们提出了 Uncertainty-o（见图 1 ）：(1) Uncertainty-o 以模型无关的方式揭示了 LMMs 中的不确定性，不论所涉及的模态、本身的架构或能力重点。值得注意的是，Uncertainty-o 也是可扩展的，可以合并新兴的模态、架构和能力。(2) 我们通过实验证实了多模态提示扰动在揭示提示变异性不确定性方面的作用。通过在提示端造成差异，我们可以观察到 LMM 响应的波动。更高程度的波动表明更高水平的变异性不确定性，体现了提示的复杂性和挑战。我们提供了对 5 种常用模态的扰动综合实证分析，包括文本、图像、视频、音频和点云。最后，我们提供了我们的发现和洞察，这些都是模态无关的，并且可以在为其他潜在模态设计扰动时作为指导。(3) 我们提出了多模态语义不确定性，可以从多模态响应中挖掘 LMM 自身的认知不确定性。通过将多模态响应映射到文本语义空间，答案语义分布的观察变得可行。文本语义空间中的答案分布熵作为认知不确定性的有效指示器。此外，各模态的不确定性被合并以揭示总体的 LMM 不确定性。

通过对 18 个多模态基准（包括 5 种模态）和 10 个 LMM 模型（包括开源和闭源）的广泛实验，我们验证了 Uncertainty-o 在准确捕捉 LMM 模型不确定性方面的有效性。可靠的不确定性估计（见图 2 ）促进了各种下游任务，包括幻觉检测、幻觉缓解和不确定性感知的 CoT。总之，我们的贡献如下：

- 揭示 LMM 不确定性的统一框架。我们提出了 Uncertainty-o，它以与模型无关的方式揭示了 LMM 心理学中的一个关键属性：不确定性。Uncertainty-o 可以无缝处理具有各种模态、复杂模型架构和不同容量重点的 LMMs。
- 多模态提示扰动手册。我们提出了多模态提示扰动，并从经验上探索如何扰动提示，以最大化其对揭示提示不确定性贡献的作用。
- 语义空间中的多模态不确定性。我们提出多模态语义不确定性，通过语义空间映射，明确地实现从多模态生成内容中挖掘内在的认知不确定性。

2. 不确定性-o

2.1. 多模态提示扰动

定义 3.1（响应语义距离）。LMM M 为提示 x_i 和 x_j（从原始提示 x 中扰动）生成响应。对于扰动提示，LMM 的预测分别为 $y_i = M(x_i)$ 和 $y_j = M(x_j)$ 。我们关注 LMM 的 Epistemic Uncertainty ，即模型参数的不确定性，$\mathrm{Var}(\theta|x)$ 。定义预测差异 $D(x)$ 为：

$$D(x) = \|y_i - y_j\|, \tag{1}$$

其中 $\|\cdot\|$ 表示 ℓ_2 范数。

命题 3.2。来自提示扰动的响应语义距离 $D(x)$ 与 LMM 不确定性的平方根成正比：因此，预测语义距离 $D(x)$ 可以作为在给定上下文下 LMM 不确定性的度量。

基于命题 3.2，我们提出了一种语义等效扰动，用于多模态提示，以诱导 LMM 响应中的变化，从而实现对 LMM 不确定性的捕获（见图 3 ）。这种方法生成跨多种模态的原始提示的语义等效变体，并将其输入 LMM。采样响应中的方差可以有效地说明 LMM 的不确定性。具体而言，我们提出了跨五种模态的多模态提示扰动，包括文本、图像、音频、视频和点云。值得注意的是，我们采用渐进扰动策略，其中每个提示从低到高程度逐步扰动，向 LMM 引入更高水平的挑战。不同模态的扰动同时应用，以进一步增强不确定性估计。

在实际操作中，不同的扰动技术会根据模态的不同而应用。对于文本提示，使用基于 LLM 的措辞和基于规则的方法，如单词替换。对于图像提示，应用旋转等空间变换以及模糊或亮度调节等属性失真。对于音频提示，扰动关键元素如音量、音高和音色，并进行额外的时间调整，如时间偏移。对于视频提示，同时应用空间和时间扰动。时间扰动包括时间裁剪和帧丢弃技术，而空间扰动涉及对每帧视频应用图像扰动技术。对于点云提示，扰动点密度和形状等 3D 特性，使用随机采样和点云抖动等技术。

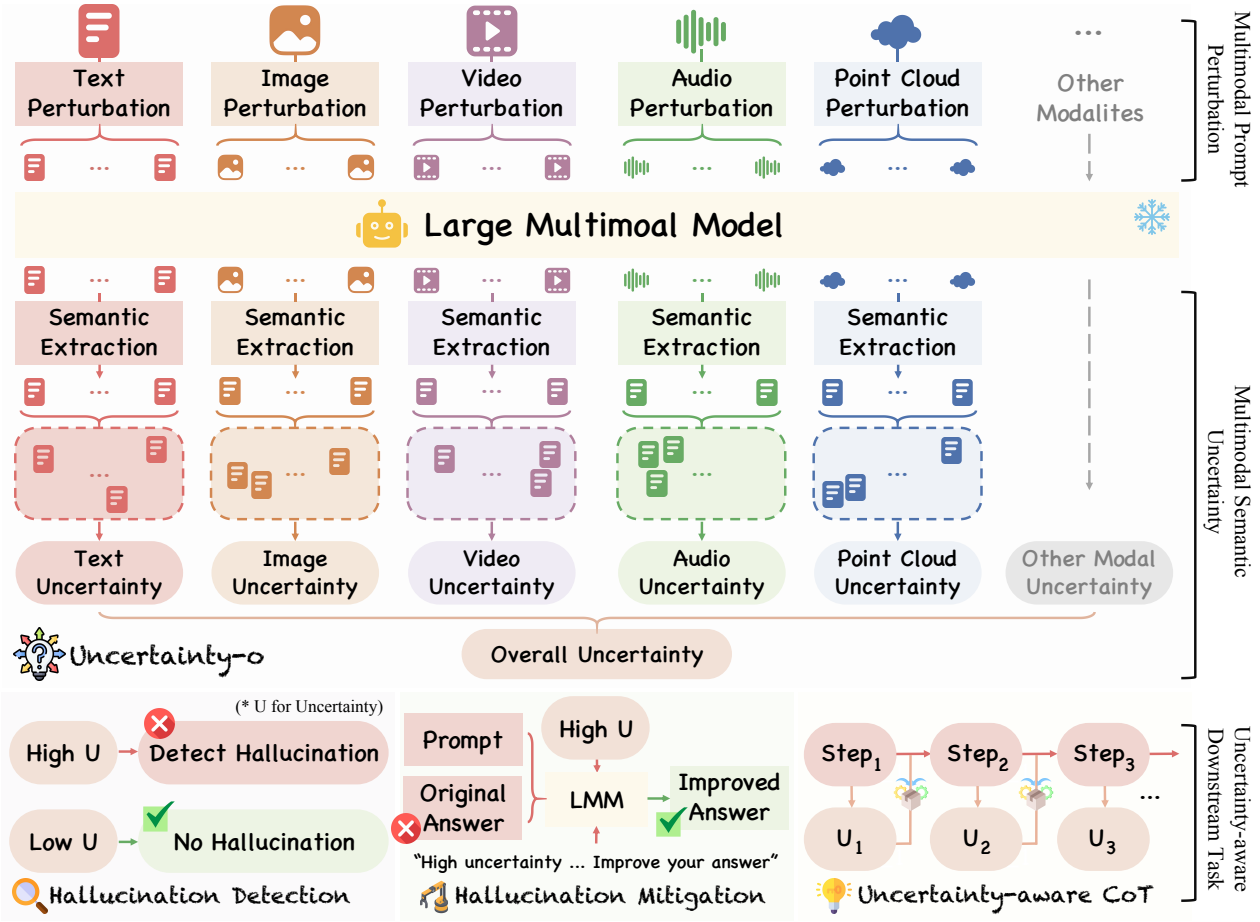

Figure 3. 我们的不确定性 o 流程。给定一个多模态提示和大型多模态模型，我们执行多模态提示扰动以生成多样化的响应。由于这些模型在扰动下固有的认识不确定性，通常会获得不同的响应。为了量化这种不确定性，我们对收集的响应进行语义聚类并计算其熵。具体来说，响应被分组到语义相似的群组中，并计算这些群组之间的熵作为最终的不确定性测量。更高的熵表示响应的变异性更大，表明信心较低，而较低的熵反映出更高的一致性，因此信心更高。最后，我们可以利用这种不确定性进行幻觉检测、幻觉缓解，并促进思维链（CoT）。

2.2. 多模态语义不确定性

我们量化这些采样答案的方差，以捕捉 LMM 固有的不确定性。我们并不只是根据格式差异简单地计算方差，而是着重于回答语义的变化。为了估计采样答案语义的方差，最关键的步骤在于对两个生成答案进行语义检测。鉴于这一挑战，特别是对于更高级别的模态，e.g. ，视频，我们建议利用现成的 LMM 字幕工具来提取答案语义并将关键点总结为文本标题。因此，复杂的多模态语义检测被转换为相对容易处理的文本语义检测（见图 3 ）。

对于那些文本说明，我们直接提示大型语言模型迭代地检查成对的文本答案。对于语义相似的文本，我们将它们聚集在一个组中。在获取采样答案的语义集群后，我们分析答案语义的离散分布并计算分布熵。此外，我们将计算得到的熵与采样时间决定的最大熵进行归一化，获得估计的 LMM 不确定性，较高的数值表示更大的 LMM 不确定性：

$$u_m = -\sum_{i=1}^{n} p_i \log(p_i) \quad \text{where} \quad p_i = \frac{c_i}{C}, \tag{2}$$

其中 c_i 是第 i 组的答案数量，C 是答案的总数，u_m 是某个模态的估计不确定性 m 。最后，将所有答案模态的不确定性进行平均得到最终的 LMM 不确定性 u 。

2.3. 不确定性感知的下游任务

幻觉检测。我们将不确定性低的答案视为非幻觉，将不确定性高的答案视为幻觉。在实践中，我们首先使用原始提示推断 LMM，以获得初始答案。然后将此答案与真实值进行比较，以确定它是否包含幻觉。接下来，我们应用提出的多模态提示扰动和多模态语义不确定性来估计 LMM 对其初始答案的不确定性。最后，我们使用 3 种不同的指标评估估计的不确定性与实际幻觉之间的一致性：接收者操作特征曲线下的面

Image Comprehension									
Method	LLaVABench [46]			MMVet [76]			CoCoCap [13]		
	AUROC ↑	AURAC ↑	ECE ↓	AUROC ↑	AURAC ↑	ECE ↓	AUROC ↑	AURAC ↑	ECE ↓
Confidence Elicitation [71]	54.0	78.1	11.3	52.3	55.8	13.1	51.2	71.0	23.1
GAVIE [45]	45.3	75.2	23.8	55.4	45.2	9.4	57.7	79.6	28.9
Semantic Entropy [24]	65.8	87.4	14.9	60.6	56.9	8.8	60.1	78.4	6.9
Uncertainty-o (ours)	68.0	91.3	10.1	63.5	66.4	8.2	65.5	82.5	5.9
Video Comprehension									
Method	MSRVTTQA [72]			MSVDQA [72]			NextQA [70]		
	AUROC ↑	AURAC ↑	ECE ↓	AUROC ↑	AURAC ↑	ECE ↓	AUROC ↑	AURAC ↑	ECE ↓
Confidence Elicitation [71]	53.5	32.6	12.2	51.0	32.8	16.1	51.5	65.9	10.9
GAVIE [45]	50.1	33.9	18.2	39.6	30.1	20.5	47.4	60.8	11.4
Semantic Entropy [24]	51.6	35.1	14.7	47.1	32.7	18.6	59.1	62.4	12.6
Uncertainty-o (ours)	55.4	39.2	10.8	52.4	36.3	14.5	67.4	69.0	7.4
Audio Comprehension									
Method	ClothoV2 [21]			ClothoAQA [44]			AudioCaps [38]		
	AUROC ↑	AURAC ↑	ECE ↓	AUROC ↑	AURAC ↑	ECE ↓	AUROC ↑	AURAC ↑	ECE ↓
Confidence Elicitation [71]	51.2	34.1	24.2	49.2	43.5	29.0	50.9	30.2	18.6
GAVIE [45]	51.1	31.9	30.6	53.1	46.7	23.9	47.7	34.8	19.7
Semantic Entropy [24]	62.1	38.5	23.7	56.7	50.2	17.9	50.9	42.4	16.1
Uncertainty-o (ours)	77.7	44.1	21.3	68.1	51.0	17.7	56.7	46.6	12.4
Point Cloud Comprehension									
Method	ModelNet [69]			ShapeNet [12]			Objaverse [19]		
	AUROC ↑	AURAC ↑	ECE ↓	AUROC ↑	AURAC ↑	ECE ↓	AUROC ↑	AURAC ↑	ECE ↓
Confidence Elicitation [71]	47.1	78.0	36.1	46.4	72.1	39.7	49.5	28.0	23.1
GAVIE [45]	46.0	84.5	43.5	51.1	75.4	37.5	49.8	35.1	29.4
Semantic Entropy [24]	52.7	79.2	39.8	54.6	73.6	32.6	50.1	38.5	26.0
Uncertainty-o (ours)	66.5	92.1	31.9	60.6	78.0	31.7	56.9	43.8	18.7

Table 1. 理解幻想检测结果。在 LMM 理解幻想检测中，Uncertainty-o 以明显的优势在众多强大的基准测试中表现优异。对于 LMM，我们利用 InternVL [16] 、VideoLLaMA [78] 、OneLLM [30] 、PointLLM [74] 来进行图像、视频、音频、点云理解。最佳结果用粗体显示，次优结果为 underlined 。

Method	MMVet [76] w. GPT4o [34]			Method	Medical Diagnosis		
	AUROC ↑	AURAC ↑	ECE ↓		AUROC ↑	AURAC ↑	ECE ↓
Confidence Elicitation [71]	49.5	67.8	11.7	Confidence Elicitation [71]	42.1	30.9	24.5
GAVIE [45]	49.0	65.6	10.4	GAVIE [45]	44.2	33.6	31.5
Semantic Entropy [24]	52.8	64.9	11.1	Semantic Entropy [24]	52.4	34.5	24.1
Uncertainty-o (ours)	56.1	75.8	8.6	Uncertainty-o (ours)	59.0	40.6	15.3
Method	MSRVTTQA [72] w. QwenVLMax [6]			Method	Embodied Robot		
	AUROC ↑	AURAC ↑	ECE ↓		AUROC ↑	AURAC ↑	ECE ↓
Confidence Elicitation [71]	54.7	39.9	11.5	Confidence Elicitation [71]	40.1	49.6	18.9
GAVIE [45]	55.9	40.1	15.6	GAVIE [45]	45.2	53.6	23.5
Semantic Entropy [24]	56.8	41.5	12.8	Semantic Entropy [24]	56.2	55.9	21.1
Uncertainty-o (ours)	60.1	47.5	8.9	Uncertainty-o (ours)	58.2	66.5	12.2

Table 2. 封闭源 LMM 的幻觉检测。我们分别利用 GPT4o [34] 和 QwenVLMax [6] 进行图像和视频理解。

Table 3. 幻觉检测用于安全关键任务。我们利用 MIMIC-CXR [36] 进行医学图像诊断和 OpenEQA [49] 进行视频体现问答。

积 (AUROC) [24] 、â€˜ 拒绝精度 â€™ 曲线下的面积 (AURAC) [24] 和预期校正误差 (ECE) [71] 。AUROC 和 AURAC 关注排名能力（值越高越好），而 ECE 则测量校准（值越低越好）。

幻觉缓解。我们将幻觉缓解表述为一个感知不确定性的两阶段修订过程。在第一阶段，我们提示 LMM 生成一个初始答案，然后用 Uncertainty-o 估计的分数来分配其不确定性分数。在构建了含有对应不确定性的答案池后，我们选择不确定性分数最高的前 K 个答案，因为这些最可能包含幻觉。在第二阶段，我们通过用原始上下文、初始答案和不确定性分数提示 LMM 来改进这些高不确定性的答案。我们明确告知 LMM 其初始答案具有高不确定性分数，并指示其由于潜在幻觉而修订响应。修订后的答案然后作为 LMM 的最终输出：

$$y_{\text{final}} = M(x', y_{\text{initial}}, u), \quad y_{\text{initial}} \in \boldsymbol{Y}, \tag{3}$$

其中 $\boldsymbol{Y}$ 包含具有最高不确定性分数的选定答案。x' 、

Image Generation			
Method	Flickr [53] AUROC ↑	AURAC ↑	ECE ↓
GAVIE [45]	52.3	67.9	12.7
Semantic Entropy [24]	50.8	70.1	21.5
Uncertainty-o (ours)	59.5	74.5	8.3
Video Generation			
Method	MSRVTT [73] AUROC ↑	AURAC ↑	ECE ↓
GAVIE [45]	51.6	24.1	21.7
Semantic Entropy [24]	54.7	30.9	23.0
Uncertainty-o (ours)	61.1	38.9	15.5
Audio Generation			
Method	VCTK [75] AUROC ↑	AURAC ↑	ECE ↓
GAVIE [45]	46.0	74.6	21.5
Semantic Entropy [24]	48.1	72.8	20.6
Uncertainty-o (ours)	53.5	81.4	11.3
Point Cloud Generation			
Method	Pix3D [59] AUROC ↑	AURAC ↑	ECE ↓
GAVIE [45]	27.9	34.2	39.5
Semantic Entropy [24]	43.6	38.8	43.5
Uncertainty-o (ours)	51.1	44.2	32.0

Table 4. 生成幻觉检测结果。我们使用 StableDiffusion [56] 、VideoFusion [47] 、AnyGPT [77] 、RGB2point [40] 进行图像、视频、音频、点云的生成。

y_{initial} 和 u 分别指更新的提示、初始答案和估计的不确定性。y_{final} 是修订后的答案。这个迭代过程帮助 LMM 改进其响应并提高准确性。

不确定性感知的链式思维。我们利用不确定性-o 来增强基础链式思维过程。自我反思对于链式思维过程中的大型语言模型至关重要，我们估计的不确定性作为引导信号用于触发自我反思。具体来说，在每个推理步骤中，除了生成响应之外，我们还估计答案的不确定性。在下一步骤中，我们将前一个答案及其不确定性纳入上下文中，明确提示大型语言模型在发现不确定性高时反思其推理过程，如：

$$y_t, u_t = M(C_t), \quad C_{t+1} = C_t \cup \{y_t, u_t\}, \tag{4}$$

其中 y_t 和 u_t 是答案和不确定性，给定步骤 t 的上下文 C_t 。这种不确定性感知的程序促进了更详细和谨慎的推理过程。

3. 实验

3.1. 与现有技术的比较

理解幻觉检测。我们将 Uncertainty-o 与图像、视频、音频和点云理解幻觉检测任务中的之前的最新方法进行比较（参见表 1 ）。Uncertainty-o 在强大的基线之上表现出显著和一致的性能提升。这验证了所提出的多模态提示扰动在更好地引发和捕获不确定性方面的一般有效性，从而促进了可靠的幻觉检测。值得注意的是，Uncertainty-o 在 ClothoV2 上超过了基线 15.6 %（AUROC），在 MMVet 上超过 9.5 %（AURAC），在 ModelNet 上超过 7.9 %（ECE），在 NextQA 上超过 4.1 %（AURAC）。我们还观察到闭源 LMM 的一致改进，这验证了 Uncertainty-o 的稳健性（参见表 2 ）。我们进一步展示了医疗诊断和实体机器人领域的检测功效比较（参见表 3 ）。我们观察到 Uncertainty-o 有效地处理了现实世界应用中的复杂性，并以明显的优势超越了基线。

生成幻觉检测。我们也展示了在图像、视频、音频和点云模态生成幻觉检测任务上的结果（见表 4 ）。在这里我们调整了 GAVIE [45] ，并在 [24] 中的仅文本语义熵的基础上扩展了我们的多模态语义不确定性。我们观察到，不确定性-o 在这些基准上始终表现出明显的改进。例如，在点云生成任务中，不确定性-o 的 AUROC 提高了 23.2 %，AURAC 提高了 10 %。这些结果验证了多模态提示扰动与多模态语义不确定性之间的相互作用可以有效地实现更准确的幻觉检测。

幻觉缓解。我们报告了在四个视频和音频基准上的理解幻觉缓解结果（见表 5 ）。“直接推理” 准确率是没有应用我们提议的不确定性方法的结果，并由我们自行复现。通过我们的不确定性引导的修正过程，我们观察到了一致的准确率提升。这些结果验证了我们估计的不确定性在幻觉缓解过程中带来的效益。

Video Comprehension		
VideoLLaMA [78]	MSRVTTQA [72] Acc.	MSVDQA [72] Acc.
Direct Inference	62.9	70.5
+ Uncertainty-o	67.2	72.1
Audio Comprehension		
OneLLM [30]	ClothoAQA [44] Acc.	ClothoV2 [21] Acc.
Direct Inference	57.1	45.5
+ Uncertainty-o	59.2	51.9

Table 5. 幻觉缓解结果。我们的不确定性引导修订有效地缓解了答案幻觉。

Image Comprehension		
InternVL [16]	MMVet [76] # Step	LLaVABench [46] # Step
Vanilla CoT	3.1	2.6
Uncertainty-Aware CoT	5.4	4.1
Point Cloud Comprehension		
PointLLM [74]	ModelNet [69] # Step	ShapeNet [12] # Step
Vanilla CoT	2.7	3.5
Uncertainty-Aware CoT	3.4	4.8

Table 6. 不确定性感知的思维链结果。我们估计的不确定性丰富了推理背景，促进了更全面的思考过程。我们报告平均推理步骤。

不确定性感知的 CoT。我们展示了不确定性感知的 CoT 与普通 CoT 的比较（见表 6 ）。通过将不确定性纳入思维背景中，LMMs 生成了更全面的思维过程。平均而言，不确定性感知的 CoT 导致整体思维长度显

著增加。

3.2. 消融研究及进一步讨论

多模态提示扰动的实证研究。我们对视频、音频和点云的 24 种不同提示扰动进行了实证比较（见图 4）。对于视频扰动，我们发现能够保持原视频语义的扰动，例如 e.g.，调整视频速度，效果更佳。相反，改变原始语义的扰动，例如 e.g.，空间裁剪，由于对不确定性估计过高，导致次优结果。对于音频扰动，观察到类似的模式。我们观察到能够保持原音频语义的扰动有利于可靠的不确定性估计，例如 e.g.，调整音量。在内部语义保持不变而外部呈现发生变化的情况下，采样答案的波动可以直接反映出 LMM 的不确定性。对于点云扰动，类似的模式也出现：保持点云语义的操作对准确的幻觉检测更有帮助。

Text	Video	Audio	Point	AUROC ↑	AURAC ↑	ECE ↓
✗	✗	-	-	51.6	35.1	14.7
✓	✗	-	-	53.0	36.1	13.9
✗	✓	-	-	54.1	37.5	12.9
✓	✓	-	-	55.4	39.2	10.8
✗	-	✗	-	50.9	42.4	16.1
✓	-	✗	-	53.0	42.6	15.5
✗	-	✓	-	52.6	44.9	12.9
✓	-	✓	-	56.7	46.6	12.4
✗	-	-	✗	50.1	38.5	26.0
✓	-	-	✗	52.4	40.2	20.9
✗	-	-	✓	54.1	41.6	19.1
✓	-	-	✓	56.9	43.8	18.7

Table 7. 扰动组合的消融研究。我们观察到，多模态提示的扰动组合能产生更好的结果。单模态提示扰动也能带来性能提升。来自 MSRVTTQA [72]、AudioCaps [38] 和 Objaverse [19] 的理解幻觉检测结果。带有 灰色阴影 的行是我们的默认设置。

> Cookbook 1: Try to use perturbation that largely preserves original prompt semantic.

扰动组合。我们在表格 7 中报告了扰动组合的消融。在多个模态提示同时发生扰动时，结果最佳。该方法有效地挖掘了每种模态带来的不确定性，有助于捕捉准确的 LMM 不确定性。

> Cookbook 2: Apply perturbation to prompt from different modalities simultaneously.

配对顺序。我们在表 8 中展示了配对顺序的消融实验。对于不同程度扰动的多模态提示，以渐进的方式与相似程度配对能产生最佳结果。这种方法创建了一个从低到高挑战水平的提示集，有效地引发了 LMM 的不确定性。

> Cookbook 3: Perturb prompts from each modality to varying degrees and pair them in the progressive order.

Pairing Order	AUROC ↑	AURAC ↑	ECE ↓
Progressive	55.4	39.2	10.8
Random	51.1	37.2	13.9
Shifted	52.4	37.8	12.5

Table 8. 扰动提示的配对顺序。将来自不同模态的具有相似程度扰动的提示进行配对可获得最佳结果。来自 MSRVTT-TQA 的理解幻觉检测结果。我们的默认设置为 灰色 。

Sampling Time	AUROC ↑	AURAC ↑	ECE ↓
2	50.6	35.1	15.9
3	54.2	36.9	12.4
5	55.4	39.2	10.8
8	55.1	38.9	11.6
10	54.9	39.2	11.2

Table 9. 采样时间消融。适度的采样时间能够实现最有效的不确定性估计。我们报告了 MSRVTTQA 理解幻觉检测的结果。我们的默认设置用 灰色阴影 强调。

采样时间。我们说明了在不确定性估计过程中采样时间的消融（见表 9）。使用适中的采样时间，e.g.，5 次，我们获得了最佳结果。过小的采样时间无法有效地建模 LMM 的不确定性，而过高的采样时间不可避免地导致不确定性被高估。

> Cookbook 4: Try with a moderate perturbation time (sampling time), e.g., 5 times.

成功幻觉检测的定性结果。我们在图 5 中展示了理解和生成场景中成功的幻觉检测案例。通过提出的多模态提示扰动，Uncertainty-o 实现了可靠的不确定性估计，从而实现准确的幻觉检测。

4. 相关工作

大型多模态模型。大型多模态模型（LMMs）[9] 可以在多种模态中感知和生成内容 [9, 65, 79, 80]，例如文本 [1]、图像 [26, 56]、音频 [77]、视频 [25, 47]、点云 [74] 等。基础性工作在构建一种通用模态编码器方面率先进行，通过将各种模态与语言对齐来实现 [28, 48]。一方面，各种研究 [30, 43, 46, 74] 专注于增强多模态理解能力 [27]，强调理解来自不同模态的提示之间的复杂性 [66] 和交互性 [82]。另一方面的研究则专注于生成多模态内容 [40, 47, 56, 77]，重视多模态响应的保真度 [41, 58] 和多样性 [22, 62]。此外，最近的一些研究 [67, 77] 试图在大型模型中构建“任意到任意”能力 [60, 61]，使模型能够同时理解复杂的上下文并根据这些上下文生成高保真内容。LMM 幻觉检测。LMM 幻觉检测 [7, 33] 旨在检测生成的多模态响应中的上下文错误 [32] 或事实错误 [3]。对于基于外

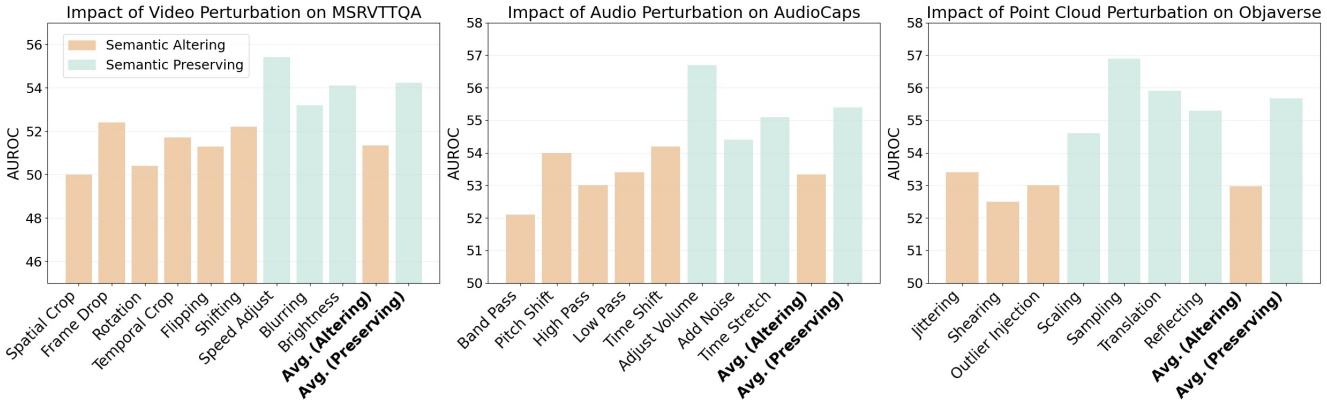

Figure 4. 不同提示扰动的经验比较。平均而言，语义保留的扰动比语义改变的扰动更有效地引发 LMM 的不确定性。来自视频、音频、点的理解幻觉检测结果。

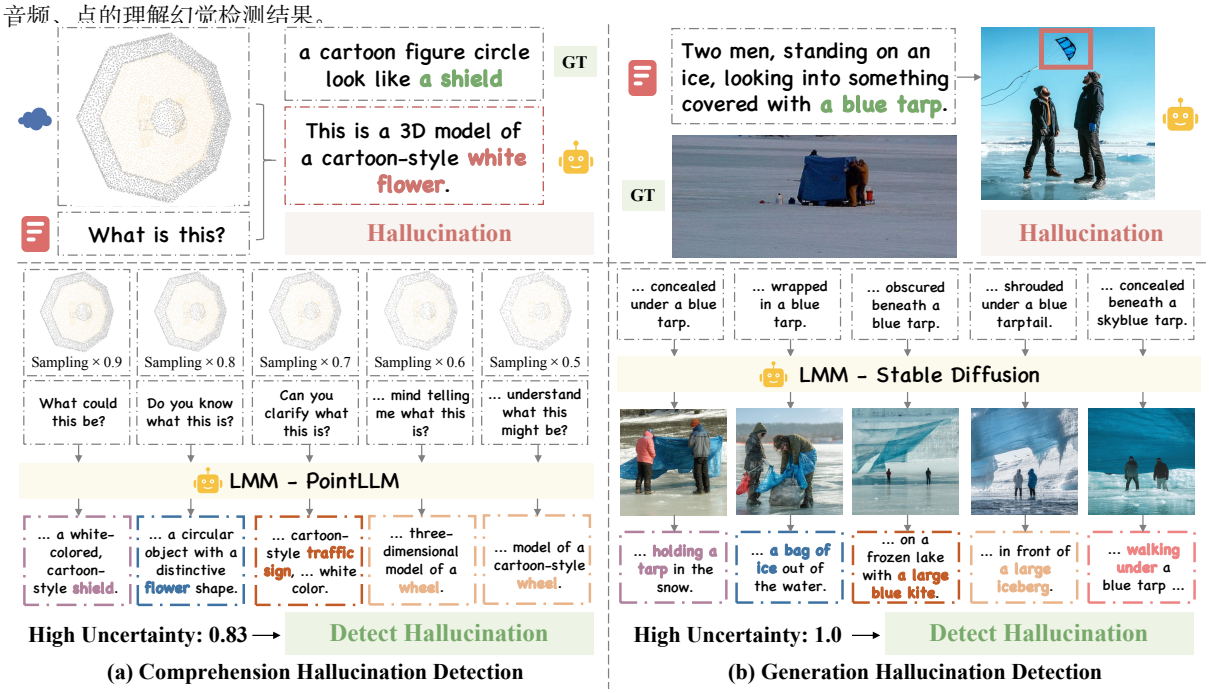

Figure 5. 成功幻觉检测的定性结果。Uncertainty-o 在基于不确定性值准确检测多种幻觉方面表现出色，包括理解（点云问答 [19] ）和生成任务（图像生成 [53] ）。

部评估者的方法，强大的响应评分器 [45, 63] 或检索到的真实世界事实 [14, 20] 促进了幻觉检测。对于基于离散规则的检查，CHAIR [55] 使用响应中呈现的对象相对于标准对象列表的比例来识别幻觉答案，其局限在于固定的对象类别。在 CHAIR 的基础上，POPE [42] 通过专注于是非问题来增强提示技术，简化流程并提高稳定性。与现有工作不同，Uncertainty-o 利用 LMM 不确定性作为幻觉响应的内在指标。LMM 中的不确定性。MAP [35] 提出了一种概率分布编码器用于在大型模型预训练期间建模不确定性。VL-Uncertainty [81] 利用视觉和文本提示的语义等价扰动来更好地捕捉 MLLM 不确定性。DropoutDecoding [23] 将与平均标记分布的差异视为不确定性，丢弃不确定性高的标记。Calibration-MLLM [15] 利用 MLLM 不同阶段之间的校准，例如在视觉微调之前和之后，揭示不确定性。IDK [17] 引入一个额外的特殊 ‘ I Dont Know ’ 标记，并基于此标记的预测概率量化不确定性。UAL [64] 利用不确定性进行大型模型的对齐，并改善特征空间中的标记收敛。

5. 结论

在本文中，我们研究 LMMs 的可靠性并探索其内在不确定性。我们提出了 Uncertainty-o，这是一个与模型无关的框架，可以揭示 LMMs 中的不确定性，并同时处理五种不同的模态（视觉、听觉、文本、视频和点云）。具体来说，我们首先引入多模态提示扰动，并通过实验证明语义保持的扰动更有效于捕捉 LMM 的不确定性。然后，我们基于施加的扰动引入多模态语义不确定性，有效地从多模态生成的响应中挖掘不确定性。大量实验和分析验证了 Uncertainty-o 在准确估计 LMM 不确定性和促进各种下游任务中的优越性。

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv:2303.08774, 2023.

[2] Lukas Aichberger, Kajetan Schweighofer, Mykyta Ielanskyi, and Sepp Hochreiter. How many opinions does your llm have? improving uncertainty estimation in nlg. In ICLR 2024 Workshop on Secure and Trustworthy Large Language Models, 2024.

[3] Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. Multimodal automated fact-checking: A survey. arXiv:2305.13507, 2023.

[4] Alfonso Amayuelas, Kyle Wong, Liangming Pan, Wenhu Chen, and William Wang. Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models. arXiv:2305.13712, 2023.

[5] Fabrizia Auletta, Rachel W Kallen, Mario di Bernardo, and Michael J Richardson. Predicting and understanding human action decisions during skillful joint-action using supervised machine learning and explainable-ai. Scientific Reports, 13(1):4992, 2023.

[6] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023.

[7] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. arXiv:2404.18930, 2024.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33: 1877–1901, 2020.

[9] Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: a survey. arXiv:2402.12451, 2024.

[10] Frederick Callaway, Bas van Opheusden, Sayan Gul, Priyam Das, Paul M Krueger, Thomas L Griffiths, and Falk Lieder. Rational use of cognitive resources in human planning. Nature Human Behaviour, 6(8): 1112–1125, 2022.

[11] Kilian Carolan, Laura Fennelly, and Alan F Smeaton. A review of multi-modal large language and vision models. arXiv:2404.01322, 2024.

[12] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. arXiv:1512.03012, 2015.

[13] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325, 2015.

[14] Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. Unified hallucination detection for multimodal large language models. arXiv:2402.03190, 2024.

[15] Zijun Chen, Wenbo Hu, Guande He, Zhijie Deng, Zheng Zhang, and Richang Hong. Unveiling uncertainty: A deep dive into calibration and performance of multimodal large language models. arXiv:2412.14660, 2024.

[16] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 24185–24198, 2024.

[17] Roi Cohen, Konstantin Dobler, Eden Biran, and Gerard de Melo. I don't know: Explicit modeling of uncertainty with an [idk] token. Advances in Neural Information Processing Systems, 37:10935–10958, 2024.

[18] Katherine Maeve Collins, Matthew Barker, Mateo Espinosa Zarlenga, Naveen Raman, Umang Bhatt, Mateja Jamnik, Ilia Sucholutsky, Adrian Weller, and Krishnamurthy Dvijotham. Human uncertainty in concept-based ai systems. In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, pages 869–889, 2023.

[19] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In CVPR, pages 13142–13153, 2023.

[20] Hanxing Ding, Liang Pang, Zihao Wei, Huawei Shen, and Xueqi Cheng. Retrieve only when it needs: Adaptive retrieval augmentation for hallucination mitigation in large language models. arXiv:2402.10612, 2024.

[21] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 736–740. IEEE, 2020.

[22] Rongyao Fang, Chengqi Duan, Kun Wang, Hao Li, Hao Tian, Xingyu Zeng, Rui Zhao, Jifeng Dai, Hongsheng Li, and Xihui Liu. Puma: Empowering unified mllm with multi-granular visual generation. arXiv:2410.13861, 2024.

[23] Yixiong Fang, Ziran Yang, Zhaorun Chen, Zhuokai Zhao, and Jiawei Zhou. From uncertainty to trust: Enhancing reliability in vision-language models with uncertainty-guided dropout decoding. arXiv:2412.06474, 2024.

[24] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. Nature, 630(8017): 625–630, 2024.

[25] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. arXiv preprint arXiv:2501.03230, 2024.

[26] Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. arXiv preprint arXiv:2412.19806, 2024.

[27] Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. Enhancing video-language representations with structural spatio-temporal alignment. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.

[28] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In CVPR, pages 15180–15190, 2023.

[29] Lewis Griffin, Bennett Kleinberg, Maximilian Mozes, Kimberly Mai, Maria Do Mar Vau, Matthew Caldwell, and Augustine Mavor-Parker. Large language models respond to influence like humans. In Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023), pages 15–24, 2023.

[30] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In CVPR, pages 26584–26595, 2024.

[31] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33:6840–6851, 2020.

[32] Jiaxing Huang and Jingyi Zhang. A survey on evaluation of multimodal large language models. arXiv:2408.15769, 2024.

[33] Wen Huang, Hongbin Liu, Minxin Guo, and Neil Zhenqiang Gong. Visual hallucinations of multi-modal large language models. arXiv:2402.14683, 2024.

[34] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. arXiv preprint arXiv:2410.21276, 2024.

[35] Yatai Ji, Junjie Wang, Yuan Gong, Lin Zhang, Yanru Zhu, Hongfa Wang, Jiaxing Zhang, Tetsuya Sakai, and Yujiu Yang. Map: Multimodal uncertainty-aware vision-language pre-training model. arXiv:2210.05335, 2022.

[36] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data, 6(1):317, 2019.

[37] Luoma Ke, Song Tong, Peng Cheng, and Kaiping Peng. Exploring the frontiers of llms in psychological applications: A comprehensive review. arXiv:2401.01519, 2024.

[38] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 119–132, 2019.

[39] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. arXiv:2209.15352, 2022.

[40] Jae Joong Lee and Bedrich Benes. Rgb2point: 3d point cloud generation from single rgb images. arXiv:2407.14979, 2024.

[41] Wei Li, Xue Xu, Jiachen Liu, and Xinyan Xiao. Unimo-g: Unified image generation through multimodal conditional diffusion. arXiv:2401.13388, 2024.

[42] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. arXiv:2305.10355, 2023.

[43] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. arXiv:2311.10122, 2023.

[44] Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. Clotho-aqa: A crowdsourced dataset for audio question answering. In 2022 30th European Signal Processing Conference (EUSIPCO), pages 1140–1144. IEEE, 2022.

[45] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. arXiv:2306.14565, 2023.

[46] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. NeurIPS, 36:34892–34916, 2023.

[47] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. arXiv:2303.08320, 2023.

[48] Yuanhuiyi Lyu, Xu Zheng, Jiazhou Zhou, and Lin Wang. Unibind: Llm-augmented unified and balanced representation space to bind them all. In CVPR, pages 26752–26762, 2024.

[49] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In CVPR, pages 16488–16498, 2024.

[50] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. arXiv:2212.08751, 2022.

[51] Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In ICCV, pages 9617–9626, 2019.

[52] Renjie Pi, Tianyang Han, Jianshu Zhang, Yueqi Xie, Rui Pan, Qing Lian, Hanze Dong, Jipeng Zhang, and Tong Zhang. Mllm-protector: Ensuring mllm's safety without hurting performance. arXiv:2401.02906, 2024.

[53] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In ICCV, pages 2641–2649, 2015.

[54] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In ICML, pages 28492–28518. PMLR, 2023.

[55] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. arXiv:1809.02156, 2018.

[56] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In CVPR, pages 10684–10695, 2022.

[57] Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. arXiv:2412.05563, 2024.

[58] Kunpeng Song, Yizhe Zhu, Bingchen Liu, Qing Yan, Ahmed Elgammal, and Xiao Yang. Moma: Multimodal llm adapter for fast personalized image generation. In ECCV, pages 117–132, 2024.

[59] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In CVPR, pages 2974–2983, 2018.

[60] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. Advances in Neural Information Processing Systems, 36:16083–16099, 2023.

[61] Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. Codi-2: In-context interleaved and interactive any-to-any generation. In CVPR, pages 27425–27434, 2024.

[62] Guy Tevet and Jonathan Berant. Evaluating the evaluation of diversity in natural language generation. arXiv:2004.02990, 2020.

[63] Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. Evaluation and analysis of hallucination in large vision-language models. arXiv:2308.15126, 2023.

[64] Yikun Wang, Rui Zheng, Liang Ding, Qi Zhang, Dahua Lin, and Dacheng Tao. Uncertainty aware learning for language model alignment. arXiv:2406.04854, 2024.

[65] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and S Yu Philip. Multimodal large language models: A survey. In 2023 IEEE International Conference on Big Data (BigData), pages 2247–2256. IEEE, 2023.

[66] Junda Wu, Zhehao Zhang, Yu Xia, Xintong Li, Zhaoyang Xia, Aaron Chang, Tong Yu, Sungchul Kim, Ryan A Rossi, Ruiyi Zhang, et al. Visual prompting in multimodal large language models: A survey. arXiv:2409.15310, 2024.

[67] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In ICML, 2024.

[68] Tsung-Han Wu, Long Lian, Joseph E Gonzalez, Boyi Li, and Trevor Darrell. Self-correcting llm-controlled diffusion models. In CVPR, pages 6327–6336, 2024.

[69] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In CVPR, pages 1912–1920, 2015.

[70] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In CVPR, pages 9777–9786, 2021.

[71] Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. arXiv:2306.13063, 2023.

[72] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In Proceedings of the 25th ACM international conference on Multimedia, pages 1645–1653, 2017.

[73] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In CVPR, pages 5288–5296, 2016.

[74] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In ECCV, pages 131–147, 2024.

[75] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). 2019.

[76] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv:2308.02490, 2023.

[77] Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. Anygpt: Unified multimodal llm with discrete sequence modeling. arXiv:2402.12226, 2024.

[78] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858, 2023.

[79] Ruiyang Zhang, Hu Zhang, Hang Yu, and Zhedong Zheng. Approaching outside: scaling unsupervised 3d object detection from 2d scene. In ECCV, pages 249–266, 2024.

[80] Ruiyang Zhang, Hu Zhang, Hang Yu, and Zhedong Zheng. Harnessing uncertainty-aware bounding boxes for unsupervised 3d object detection. arXiv:2408.00619, 2024.

[81] Ruiyang Zhang, Hu Zhang, and Zhedong Zheng. Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. arXiv:2411.11919, 2024.

[82] Yuechen Zhang, Shengju Qian, Bohao Peng, Shu Liu, and Jiaya Jia. Prompt highlighter: Interactive control for multi-modal llms. In CVPR, pages 13215–13224, 2024.

不确定性-o：用于揭示大型多模态模型不确定性的一个与模型无关的框架

Supplementary Material

6. 讨论

为什么语义等价扰动比不等价扰动更有效？在语义等价扰动中，所有被扰动的提示的语义保持不变。因此，LMM 回答中的任何波动直接反映了其不确定性。另一方面，语义不等价扰动改变了原始语义，答案的变化是不可避免的，这使得很难准确反映 LMM 的固有不确定性。

为什么提示扰动在捕捉 LMM 不确定性方面优于经典采样？多模态提示扰动通过在不同层次上变化的提示来挑战 LMM，答案的更大变化表示更高的不确定性。相比之下，经典采样通常在采样时使用高温度，这往往会导致所有样本的不确定性增高。例如，LMM 可能对其答案非常自信，但高温度不可避免地导致答案变化，从而导致人为的高不确定性。

多模态语义不确定性可以扩展到更多的模态吗？可以。Uncertainty-o 提出了一个用于 LMM 不确定性估计的通用流程。这个流程是正交的，并且可以从 LMM 的多次迭代中受益。通过一个更通用的 LMM，此 LMM 能够将附加模态的语义内容解析为文本描述，Uncertainty-o 可以无缝地估计在更多模态中的那些答案的不确定性。

为什么不直接利用 LMM 来检查两个生成的多模态响应在语义上是否等价？当前的 LMM 仍然缺乏针对多模态语义检查的熟练能力。我们尝试了两种方法利用 LMM 进行直接的多模态数据语义检查：(1) 将两个多模态答案合并为一个，并提示 LMM 检查答案是否相似。例如，将两个点云合并为一个并进行位移以避免重叠。然而，当前的 LMM 表现不尽如人意。(2) 直接将两个答案输入 LMM 并提示其进行检查。我们发现，许多 LMM 仍然不支持多输入，比如同时理解两个点云。

7. 设置

基准测试。我们使用 18 个基准进行实验。在理解幻觉任务中，我们采用的基准通常是问答类型。对于图像，我们使用 MMVet、LLaVABench 和 CoCoCap。对于视频基准，我们利用 MSRVTTQA、MSVDQA 和 NextQA。对于音频模态，我们使用 ClothV2 和 AudioCaps。对于点云基准，我们利用 PointCap 和 ModelNet。在生成幻觉检测中，我们使用标题类型的基准，并将标题作为生成提示。对于图像生成，我们使用 Flickr。对于视频生成，我们利用 MSRVTT。对于音频生成，我们使用 VCTK。对于点云生成，我们使用 Pix3D，旨在从图像生成到点云生成。对于安全关键任务，我们使用 MIMICCXR 和 OpenEQA。MIMICCXR 是一个胸部 X 光片基准，由放射科医生提供医学诊断注释。OpenEQA 是一个来自于具身机器人视角的内部视频基准，并附有基于文本的问题。

LMM。我们实验了 10 种不同的 LMM，它们能够理解和生成多种数据模式。对于闭源 LMM，我们使用 GPT-4o 和 QwenVLMax。对于开源 LMM，InternVL 是一个大型视觉-语言模型，它接收图文提示并以文本形式回答。VideoLLaMA 则通过视频作为提示并生成文本答案。OneLLM 擅长处理各种模式输入，e.g. 、音频和点云，并仅输出文本响应。PointLLM 专门用于现实世界点云的理解。StableDiffusion 是一个基于文本的图像生成模型。VideoFusion 是一个高效的视频生成模型，以文本作为提示。AnyGPT 能够生成音频和视觉数据。RGB2point 通过接收单一图像实现点云生成。

实施细节。幻觉检测过程包含三个关键阶段：(1) 初始答案获取。首先将上下文提供给被测试的 LMM 生成初始输出。在此阶段，我们将变异超参数保持在低水平，e.g. ，大型 LLM 模型使用低温度，Diffusion 模型使用高指导尺度。将初始答案与真实情况进行对比，以决定其是否包含幻觉。(2) 不确定性估计。根据它们包含的模态，我们提出的多模态提示扰动对初始上下文进行扰动。所有模态提示同时以不同程度进行扰动。不同层次的扰动提示启用了采样过程。对于采样的文本答案，现成的 LLM，Qwen2.5-7B，会按语义对它们进行聚类，并计算熵作为不确定性。对于其他模态的答案，如 e.g. ，图像，我们使用 OneLLM-7B 作为说明器，将这些答案转换为简洁的描述。然后，不确定性计算简化为普通文本语义不确定性。(3) 幻觉检测。借助于 (1) 中的幻觉标签和 (2) 中估计的不确定性，我们使用 AUROC、AURAC 和 ECE 作为幻觉检测指标。这三个指标全面评估了不确定性与幻觉之间的校准程度。为了缓解幻觉现象，我们遵循一个两阶段的程序：(1) 我们首先使用初始上下文提示 LMM 以获得初始答案，并使用 Uncertainty-o 计算其不确定性。(2) 然后我们选择不确定性最高的前 50 个% 答案，因为这些答案最可能包含幻觉。对于每个这些答案，我们使用以下提示：‘提示： $ X ，初始答案： $ Y ，你的答案具有 $ U 的高不确定性分数，范围从 0 到 1。你能改进你的答案并修正它以提高准确性吗？ $ X ， $ Y ，和 $ U 代表原始提示、初始答案和估计不确定性的实际内容。最后，我们使用修订后的答案作为 LMM 的最终输出。对于考虑不确定性的 CoT，我们将其视为一个多轮思维过程。在第一轮中，我们在原始提示后附加 “让我们逐步思考。现在，提供你的第一步答案：” 然后我们使用 Uncertainty-o 估计 LMM 的不确定性。从第二步开始，除了原始提示之外，我们还添加所有先前的答案及其对应的不确定性得分。在提示的末尾，我们添加 “当你认为已经解决了问题时，用 ‘完成’ 来回应。” 我们在每一步检查答案是否出现 “完成。”，一旦发现即退出循环。

Text Clustering	AUROC ↑	AURAC ↑	ECE ↓
Lexical	48.0	33.2	12.7
Semantic	55.4	39.2	10.8

Table 10. 文本聚类的消融研究。在进行语义不确定性计算时，以语义为基础的聚类比仅仅基于词汇表示的聚类更为有效。

Method	AUROC ↑	AURAC ↑	ECE ↓
Cycle Consistency	57.7	72.9	9.2
Uncertainty-o	59.5	74.5	8.3

Table 11. 消融循环一致性用于生成幻觉检测。不确定性-o 超过了循环一致性，因为循环一致性受到 LLM 的文本相似度评分能力的限制。我们在 Flickr 上使用 StableDiffusion 进行了实验。

8. 更多消融实验

文本聚类方式。我们报告了文本聚类方法的消融实验结果（见表格 10 ）。基于文本固有语义进行聚类，而不仅仅是其词汇表现，取得了最佳结果。

循环一致性。我们还在图像生成环境中尝试使用循环一致性来检测幻觉（见表 11 ）。我们观察到，Uncertainty-o 超越了这一基线方法，而这一基线方法受限于 LLM 的文字对相似性评分能力。

9. 比例定理的详细证明

9.1. 定义和假设

1. 采样过程：大型模型 M 同时对扰动后的提示 x_i 和 x_j 进行预测。对于这些输入提示，大型模型的预测结果分别是 $y_{\mathrm{i}} = M(x_{\mathrm{i}})$ 和 $y_{\mathrm{j}} = M(x_{\mathrm{j}})$ 。
2. 不确定性：我们关注大模型的认知不确定性，即大模型参数的不确定性。假设大模型参数 θ 是具有先验分布 $P(\theta)$ 的随机变量。
3. 预测差异：定义预测差异 $D(x)$ 为：

$$D(x) = \|y_{\mathrm{i}} - y_{\mathrm{j}}\|,$$

其中 $\|\cdot\|$ 表示语义空间距离。

9.2. 数学推导

大模型的预测分布。假设大模型的输出是一个概率分布 $P(y|x,\theta)$ ，其中 y 是响应，x 是提示，θ 是模型参数。

后验预测分布。根据贝叶斯定理，大模型的后验预测分布可以表示为：

$$P(y|x) = \int P(y|x,\theta)P(\theta|x)d\theta,$$

其中 $P(\theta|x)$ 是大模型参数的后验分布。

认知不确定性。参数的不确定性可以通过后验分布的方差来衡量：

$$\mathrm{Var}(\theta|x) = \mathbb{E}_{\theta|x}[(\theta - \mathbb{E}_{\theta|x}[\theta])^2] = \mathbb{E}_{\theta|x}[\theta^2] - (\mathbb{E}_{\theta|x}[\theta])^2.$$

预测差异和认知不确定性。为了将预测差异 $D(x)$ 与参数不确定性联系起来，我们需要考虑来自扰动提示的预测。假设在第 i 次和第 j 次提示扰动期间的参数分别是 θ_{i} 和 θ_{j} ，并且它们具有相同的先验分布，即 $P(\theta_{\mathrm{i}}) = P(\theta_{\mathrm{j}})$ 。

来自扰动提示的预测。来自扰动提示的预测可以表示为：

$$y_{\mathrm{i}} = \mathbb{E}_{\theta_{\mathrm{i}}|x}\left[P(y|x,\theta_{\mathrm{i}})\right],$$

$$y_{\mathrm{j}} = \mathbb{E}_{\theta_{\mathrm{j}}|x}\left[P(y|x,\theta_{\mathrm{j}})\right].$$

预测差异的表达式。假设预测差异可以用一阶泰勒展开式来近似：

$$y_{\mathrm{i}} - y_{\mathrm{j}} \approx \mathbb{E}_{\theta|x}\left[\nabla_\theta P(y|x,\theta)\cdot(\theta_{\mathrm{i}} - \theta_{\mathrm{j}})\right],$$

其中 $\nabla_\theta P(y|x,\theta)$ 是 $P(y|x,\theta)$ 对 θ 的梯度。

因此，预测差异 $D(x)$ 可以表示为：

$$D(x) = \|y_{\mathrm{i}} - y_{\mathrm{j}}\| \approx \mathbb{E}_{\theta|x}\left[\|\nabla_\theta P(y|x,\theta)\cdot(\theta_{\mathrm{i}} - \theta_{\mathrm{j}})\|\right].$$

来自提示扰动的预测差异与认知不确定性之间的关系。为了简化分析，假设 θ_{i} 和 θ_{j} 是独立同分布的（i.i.d.）。然后：

$$\mathbb{E}_{\theta|x}\left[\|\nabla_\theta P(y|x,\theta)\cdot(\theta_{\mathrm{i}} - \theta_{\mathrm{j}})\|^2\right] \approx$$

$$\mathbb{E}_{\theta|x}\left[\|\nabla_\theta P(y|x,\theta)\|^2\right]\cdot\mathbb{E}_{\theta|x}\left[(\theta_{\mathrm{i}} - \theta_{\mathrm{j}})^2\right].$$

使用 $\mathbb{E}_{\theta|x}\left[(\theta_{\mathrm{i}} - \theta_{\mathrm{j}})^2\right] = 2\left(\mathbb{E}_{\theta|x}[\theta^2] - (\mathbb{E}_{\theta|x}[\theta])^2\right) = 2\cdot\mathrm{Var}(\theta|x)$ ，我们有：

$$\mathbb{E}_{\theta|x}\left[\|\nabla_\theta P(y|x,\theta)\cdot(\theta_{\mathrm{i}} - \theta_{\mathrm{j}})\|^2\right] \approx$$

$$2\cdot\mathbb{E}_{\theta|x}\left[\|\nabla_\theta P(y|x,\theta)\|^2\right]\cdot\mathrm{Var}(\theta|x).$$

假设 $k = \mathbb{E}_{\theta|x}\left[\|\nabla_\theta P(y|x,\theta)\|^2\right]$ 为正，我们得到：

$$\mathbb{E}_{\theta|x}\left[\|\nabla_\theta P(y|x,\theta)\cdot(\theta_{\mathrm{i}} - \theta_{\mathrm{j}})\|^2\right] \approx 2k\cdot\mathrm{Var}(\theta|x).$$

因此，来自提示扰动的预测差异 $D(x)$ 可以表示为：

$$D(x) \approx \sqrt{2k}\cdot\sqrt{\mathrm{Var}(\theta|x)}.$$

进一步简化，我们得到：

$$D(x) \propto \sqrt{\mathrm{Var}(\theta|x)}.$$

9.3. 最终定理

从上述推导中，我们已展示出由提示扰动引起的预测差异 $D(x)$ 与模型认知不确定性 $\sqrt{\mathrm{Var}(\theta|x)}$ 的平方根成正比。因此，预测差异 $D(x)$ 可以作为衡量给定样本模型认知不确定性的方法。