基于归纳保准预测的大型视觉语言模型预测集的数据驱动校准

A PREPRINT

Yuanchang Ye
School of Data Sciences
Zhejiang University of Finance & Economics
HangZhou, China

Yanwen Wei
School of Data Sciences
Zhejiang University of Finance & Economics
HangZhou, China

April 25, 2025

ABSTRACT

本研究通过一个分割一致性预测（SCP）框架，解决了大型视觉语言模型（LVLMs）在视觉问答（VQA）任务中的幻觉缓解这一关键挑战。尽管 LVLMs 在多模态推理方面表现出色，但其输出经常呈现出高置信度的幻觉内容，在安全关键应用中构成风险。我们提出了一种与模型无关的不确定性量化方法，该方法整合了动态阈值校准和跨模态一致性验证。通过将数据分为校准集和测试集，该框架计算偏离分数，以构建在用户定义的风险水平（α）下具有统计保证的预测集。主要创新包括：（1）严格控制边缘覆盖，以确保经验错误率严格低于 α；（2）动态调整预测集大小与 α 反比，过滤低置信度输出；（3）消除对先验分布假设和重新训练要求。在八个 LVLM 上对基准（ScienceQA, MMMU）的评估表明，SCP 在所有 α 值上强制实现理论保证。该框架在不同的校准到测试分割比例中实现了稳定的性能，凸显了其在医疗保健、自动驾驶系统和其他安全敏感领域中实际应用的稳健性。此项工作在多模态人工智能系统中弥合了理论可靠性与实际应用性之间的差距，为幻觉检测和不确定性感知决策提供了一种可扩展的解决方案。

随着多模态模型的快速发展，大型视觉语言模型（LVLMs）已经被广泛应用于医疗保健和自动驾驶等关键领域。然而，关于视觉语言问答（VQA）任务的研究表明，与单模态语言模型相比，这些多模态系统更容易受到明显幻觉现象的影响。尽管生成的响应往往显得可信并展现出高度自信，这些模型却可能产生不准确的输出。依赖此类幻觉结果可能引入决策偏差，甚至带来重大安全风险。在此背景下，开发高效自动的幻觉检测机制成为确保多模态 AI 系统可靠性的核心挑战。此外，研究表明在 VQA 任务中同时处理视觉和文本信息会增加幻觉的风险。这些问题凸显了需要自动化的检测框架，以便在不依赖于先验知识的情况下适应多模态的不确定性。我们的方法结合动态阈值校准和跨模态一致性验证，旨在为安全敏感的应用提供实时强健的可靠性。

先前的研究集中于量化模型输出并为用户提供评估自然语言生成（NLG）可靠性的措施 Liang et al. [2024], Li et al. [2023] 。当前的不确定性量化方法，如校准技术和语言化不确定性方法，旨在标识预测的可信度。然而，这些方法——经常是启发式的——未能提供任务特定的性能保证，限制了其实际应用。例如，语言化不确定性常常表现出过度自信，削弱了其可靠性。虽然校准使概率与经验真实性率保持一致，但它需要昂贵的重新训练，并且容易受到分布变化的影响。这些限制突显出需要更强大且更具普适性的框架来确保在 NLG 中进行可信的不确定性估计。

保序预测（CP）是一种不确定性量化框架，其主要优势在于仅基于数据的可交换性假设 Romano et al. [2019], Cresswell et al. [2024], Ke [2025] ，提供关于真实结果覆盖的严谨统计保证。与依赖于启发式近似或复杂先验分布的方法相比，CP 是模型无关、无分布假设且计算高效的，使其能够直接应用于预训练系统而无需重新训练。在这项工作中，我们采用了划分保序预测（SCP）方法，并将其扩展到封闭式视觉问答（Vision-VQA）任务中的多选情境。具体来说，目标数据集的候选答案集首先使用 LVLMs 生成，然后基于校准集样本的真实标签设计一个非一致性评分（NS），以量化模型输出的不确定性。通过计算校准集中 NS 的分位数并结合用户指定的风险水平（表示为 δ），最终在测试集中实现了对边缘覆盖的严格控制。这种方法不仅避免了传统方法内在分布假设的依赖，还为多模态场景中的可靠决策提供了理论支持。我们的实验使用了 MMMU 和 ScienceQA 作为基准数据集，并评估了来自四个不同模型组的八个 LVLMs，包括 LLaVA1.5、LLaVA-NeXT、Qwen2VL 和 InternVL2。广泛的实证结果表明，我们的框架在各种用户指定的风险水平（表示为 α）下，实现了对误覆盖率的严格控制。例如，在 ScienceQA 基准测试中，即使对错误概率（$\alpha \geq 0.6$）有较高的容忍

度，Qwen2-VL-7B-Instruct 模型的经验错误率仍保持在 $\alpha = 0.6$ 以下。值得注意的是，随着 α 的增加，生成的答案集的平均预测集大小系统地收紧——这是减轻 LVLMs 幻觉的重要特性。这种 α 与预测集大小之间的反比关系确保了较高的风险容忍度会产生更紧凑的预测集，有效过滤掉低置信度或虚假的输出。此外，无论校准到测试数据的划分比例如何，平均经验错误率始终遵循用户定义的风险水平。结合可控的预测集粒度，这种稳健性突显了该方法的双重能力：在确保统计学上有效覆盖的同时，通过自适应集约束动态抑制幻觉式反应。这种能力对于在安全关键场景中部署 LVLMs 至关重要，在这些场景中，可靠性和精确性同等重要。

1 相关工作

大型视觉语言模型。早期研究主要集中于从图像和文本输入生成文本响应。在此基础上，随后研究显著扩展了 LVLMs 的能力和应用领域。最近的进展进一步增强了细粒度解析能力，使其能够在整体图像理解之外，对局部区域（例如边界框或关键点）进行精确控制。这些发展促进了 LVLMs 在医疗诊断、具身机器人交互和自动驾驶等关键领域的广泛部署。然而，多模态交互的复杂性引入了新的挑战——例如，跨模态信息融合中的不一致可能会降低输出可靠性。在诸如医疗保健和自治系统等高风险场景中，不可靠的模型响应可能导致严重的安全隐患，这凸显了准确幻觉检测的必要性。与传统依赖外部验证的方法不同，本工作提出量化 LVLMs 的内在不确定性来识别幻觉，建立一个构建安全可靠的人机协作系统的新理论基础。

大语言模型中的幻觉。在自然语言处理领域，幻觉指的是生成的内容看似合理但偏离来源材料或事实准确性的现象，这种现象源于心理学中关于感知不存在现实的概念 Lin et al. [2023], Kuhn et al. [2023], Farquhar et al. [2024], Wang et al. [2025a] 。这种现象主要表现为两种类型：内在幻觉（与源上下文直接矛盾）和外在幻觉（通过原始上下文或外部知识库无法验证的内容）。对大型视觉-语言模型（LVLMs）的研究表明，它们对以用户为中心的互动和指令对齐的高度关注导致事实的扭曲，这些扭曲被分类为事实幻觉（偏离可验证真相）和忠实幻觉（违背用户指令、上下文一致或逻辑一致性）。检测方法遵循两种途径。(1) 基于外部模型的评估：这种方法利用先进的 LVLMs 作为评分判别器来评估响应质量，但受限于对合成注释的依赖。(2) 基于离散规则的检查：基于离散规则的系统通过 CHAIR、MME 和 POPE 等基准专注于对象幻觉（OH）的评估。缓解策略采用对比解码（CD）和后处理技术：CD 通过视觉区域比较、自我对比分析和偏好模型比较来解决感知偏差，但也存在敏感性和过于简化的问题；后处理通过迭代提示优化响应，但面临计算开销和任务适应性的限制。该框架为系统评估 LVLM 输出可靠性提供了多维见解。

分割保序预测（SCP）。SCP 作为一种理论支持的不确定性量化框架在大规模视觉语言模型（LVLMs）中展现了独特的优势。其核心机制利用可交换数据校准来生成具有统计保证的预测集，保证涵盖真实答案，适用于处理开放式自然语言生成任务的黑箱模型 Campos et al. [2024], Angelopoulos et al. [2023], Wang et al. [2024], Ye et al. [2024], Angelopoulos et al. [2024], Wang et al. [2025b,c] 。与传统的不确定性框架不同，SCP 需要的假设最少，同时提供可验证的覆盖保证。该方法对模型无关且不依赖分布，仅在可交换数据条件下运行。最近的扩展通过使用置信阈值（例如，在问答任务中筛选候选答案）或基于似然的生成序列停止规则，适应于多模态场景的动态预测集构建。针对开放式生成的局限，先进的实现采用黑箱不确定性量化策略，将不确定性度量与正确性标准严格关联，从而实现对不同模型架构和数据复杂性下的强大覆盖保证。尽管存在非交换数据适应和实时计算需求等挑战，SCP 的模型独立性、无分布依赖性和偏差控制能力使其成为评估 LVLM 输出可靠性的理论上严格且实际可行的解决方案。

2 方法

我们的方法主要解决两个挑战。(1) 如何识别模型输出中满足用户需求的响应分布。(2) 如何严格证明所识别的输出分布符合模型的统计保证。我们首先开发了一种基于不一致性分数的不确定性量化方法，以建立模型生成响应的可靠性度量。此外，我们采用分割一致性预测系统地将不确定性量化结果的启发式近似转化为统计上的严格结果。此方法确保了预测集的鲁棒性和更强的统计保证，从而为模型的输出分布提供理论上的保障。

2.1 预备知识

考虑一个预测任务，其中 $\mathcal{X}$ 和 $\mathcal{Y}$ 分别表示输入和输出集。根据之前关于 CP 框架的研究，我们首先建立一个包含 n 样本的校准集合，用 $\{(X_i, Y_i)\}_{i=1}^{n}$ 表示。接下来，对于 K 类分类任务，我们收集 M 样本作为测试数据，用 $\{(X_i, Y_i)\}_{i=N+1}^{M+N}$ 表示。随后，我们定义一个 LVLM 分类器模型 $\hat{f} : \mathcal{X} \to \mathcal{Y}$ 。我们将第 i 个样本的正确类别（真实值）表示为 y_i^* 。

对于每个数据点，在没有任何系统提示处理的情况下，样本直接输入到 LVLM 分类器 $\hat{f}$ 中，并进行 P 次随机采样。对于第 i 个数据样本获得的结果表示为 $\hat{y}_k^{(i)}$ ；例如，当 $i=1$ 和 $K=5$ 时，随机采样结果为

$$\{\underbrace{\hat{y}_1^{(1)}, \hat{y}_2^{(1)}, \hat{y}_2^{(1)}, \hat{y}_3^{(1)}, \ldots, \hat{y}_5^{(1)}, \hat{y}_5^{(1)}}_{P}\}.$$

。

在 CP 框架中，一个关键的组成部分是非一致性得分 $\mathcal{S}: \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ ，它提供了一种启发式的方法来衡量分类器的预测与给定输入的符合程度。对于分类任务，分类器的输出分布用作

$$\mathcal{S}(x, y) = 1 - \hat{f}(y \mid x) \quad \text{for } y = 1, \ldots, |\mathcal{Y}|, \tag{1}$$

，我们称 $S_i = \mathcal{S}(X_i, Y_i)$ 为 i -个校准示例的非一致性得分。

2.2 方法

对于一个新的、未见过的测试样本 x_{test} ，生成 $\hat{C}_\alpha(X_{test})$ 的步骤如下：

1. 计算校准数据 $(S_1, \ldots, S_N)$ 的非一致性分数，其中 $S_i = \mathcal{S}(X_i, Y_i)$ ；
2. 将 $\tau = Q_{1-\alpha}(\{S_i\}_{i=1}^n)$ 定义为分数的经验分布的保序 α 分位数，其中 $\alpha \in (0,1)$ 是我们设定的显著性水平。较小的 α 值对应于较低的允许错误率。在计算 τ 时，将集合 $\{S_i\}_{i=1}^N$ 按降序排列，并选择与第 $\lceil (1-\alpha)(n+1)/n \rceil$ 个次序统计量对应的分位数；
3. 最后，我们使用之前定义的预测集：

$$\hat{C}_\alpha(x_{\text{test}}) = \{y \in \mathcal{Y} : \mathcal{S}(x_{\text{test}}, y) \leq \tau\} \tag{2}$$

步骤 1 和 2 通常被称为校准，而步骤 3 被称为预测。直观地说，预测集包括与校准集中的样本同样或更好符合的所有预测。

2.3 理论保证

保形预测（CP）的覆盖保证来源于其两个基本理论属性：与分布无关的有效性和边际覆盖率。正如 Vovk 等人（2005）所展示的，由前一小节定义的保形预测器生成的预测集满足以下覆盖保证：

$$\mathbb{P}\left[Y_{\text{test}} \in \hat{C}_\alpha(X_{\text{test}})\right] \geq 1 - \alpha \tag{3}$$

前提是数据满足可交换性。可交换性要求数据的联合概率分布在置换下保持不变。形式上，一个序列 $(Z_1, \ldots, Z_n)$ 是可交换的，当且仅当对于任何 $\{1, \ldots, n\}$ 的置换 π ：

$$(Z_1, \ldots, Z_n) \triangleq (Z_{\pi(1)}, \ldots, Z_{\pi(n)})$$

其中 $\triangleq$ 表示分布相等。可交换性是比独立同分布（i.i.d.）更弱的条件。虽然 i.i.d. 变量必然是可交换的，但可交换变量不需要是独立的——它们只需是同分布的。这产生了 Romano 的上界：

$$\mathbb{P}\{Y_{n+1} \in \hat{C}(X_{n+1})\} \leq 1 - \alpha + \frac{1}{n+1}. \tag{4}$$

值得注意的是，随着校准集大小 n 的增加，覆盖概率严格收敛到 $1-\alpha$ 。至关重要的是，所述的符合预测过程展现了模型无关性和分布自由有效性——除了可交换性外，它对数据分布没有任何假设。在后续任务中，我们应该使用此公式来评估结果是否满足我们的保证：

$$\mathbb{P}\left[Y_{\text{test}} \notin \mathcal{C}_\alpha(X_{\text{test}})\right] \leq \alpha \tag{5}$$

当经验错误率低于 α 水平时，我们可以自信地得出结论，预测集满足指定的覆盖保证。

3 评估

3.1 实验设置

基准测试。我们的实验采用多项选择基准测试。对于多项选择数据集，我们使用两个基准测试：MMMU 和 ScienceQA。具体来说，MMMU 包含 11.5K 来自大学水平的多模态问题

Benchmarks	LVLMs	Split_Ratio								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
ScienceQA	Qwen2-VL-7B-Instruct	0.1895	0.1949	0.1943	0.1955	0.1951	0.1947	0.1923	0.1921	0.1908
	Qwen2-VL-2B-Instruct	0.1827	0.1879	0.1908	0.1851	0.1959	0.1888	0.1927	0.1950	0.1827
	InternVL2-8B	0.1847	0.1901	0.1984	0.1942	0.1936	0.1937	0.1955	0.1959	0.1975
	InternVL2-1B	0.1874	0.1869	0.1932	0.1895	0.1905	0.1899	0.1901	0.1873	0.1937
	llava-v1.6-vicuna-13b-hf	0.1810	0.1887	0.1923	0.1865	0.1912	0.1909	0.1930	0.1833	0.1902
	llava-v1.6-mistral-7b-hf	0.1825	0.1847	0.1871	0.1841	0.1922	0.1923	0.1869	0.1884	0.1880
	llava-1.5-13b-hf	0.1889	0.1905	0.1844	0.1952	0.1959	0.1918	0.1917	0.1909	0.1960
	llava-1.5-7b-hf	0.1902	0.1867	0.1892	0.1948	0.1925	0.1921	0.1879	0.1900	0.1876
Average		0.1859	0.1888	0.1912	0.1906	0.1934	0.1918	0.1913	0.1904	0.1908
MMMU	Qwen2-VL-7B-Instruct	0.1579	0.1750	0.1858	0.1828	0.1829	0.1795	0.1879	0.1880	0.1970
	Qwen2-VL-2B-Instruct	0.1640	0.1651	0.1632	0.1758	0.1789	0.1848	0.1821	0.1877	0.1924
	InternVL2-8B	0.1740	0.1799	0.1800	0.1824	0.1843	0.1745	0.1854	0.1823	0.1783
	InternVL2-1B	0.1423	0.1538	0.1637	0.1706	0.1673	0.1605	0.1699	0.1619	0.1510
	llava-v1.6-vicuna-13b-hf	0.1463	0.1577	0.1614	0.1673	0.1626	0.1579	0.1576	0.1580	0.1531
	llava-v1.6-mistral-7b-hf	0.1617	0.1713	0.1780	0.1821	0.1770	0.1857	0.1788	0.1775	0.1853
	llava-1.5-13b-hf	0.1578	0.1747	0.1765	0.1790	0.1835	0.1844	0.1888	0.1900	0.1903
	llava-1.5-7b-hf	0.1731	0.1719	0.1758	0.1777	0.1735	0.1878	0.1774	0.1782	0.1669
Average		0.1596	0.1687	0.1731	0.1772	0.1763	0.1769	0.1785	0.1779	0.1768

Table 1: 在固定的 $\alpha = 0.2$ 下，我们对两个基准和八个 LVLMs 的不同分割比例的错误率 α 进行了比较。首先，无论是在较小的校准集还是在较大的校准集中，α 值均低于我们预设的 $\alpha = 0.2$ 。此外，通过观察每个分割比例的平均 α 值，我们发现尽管仅使用少量样本，测试集结果都相当不错。这表明我们的 SCP 方法可以保证测试集的结果。

考试、小测验和教科书，涵盖六个核心学科：艺术 & 设计、商业、科学、健康 & 医学、人文 & 社会科学以及技术 & 工程。这些问题涉及 30 个主题和 183 个子领域，包含 30 种高度异质的图像。MMMU 还提供了一个完整的测试集，其中包括 150 个开发样本和 900 个验证样本。对于 ScienceQA，这些问题来源于由 IXL Learning 管理的开放资源，该资源是一个由 K-12 领域专家策划的在线教育平台。该数据集包含符合加州公用核心内容标准的问题，包括 21,208 个样本，分为训练（12,726），验证（4,241）和测试（4,241）集。

基础 LVLM。在这个实验中，我们评估来自四个不同模型组的八个 LVLM 模型。具体来说，我们使用 LLaVA-1.5、LLaVA-NeXT、Qwen2-VL 和 InternVL2 对前述的基准进行了推理。LLaVA-1.5 通过一个两层 MLP 连接器将 CLIP 视觉编码器与大型语言模型（例如，Vicuna）对齐，采用两阶段训练策略（预训练和指令微调），并在视觉问答和 OCR 任务中表现出强大的性能。LLaVA-NeXT 通过引入动态高分辨率处理（AnyRes）扩展 LLaVA-1.5，通过网格划分结合全局和局部特征提升视觉推理能力，还扩展到视频理解。Qwen2-VL 采用动态分辨率适应，通过灵活的高分辨率图像拆分保留细粒度细节。InternVL2 通过扩展视觉编码器（例如，InternViT-6B），应用动态高分辨率处理和像素重组以减少视觉令牌，利用三阶段渐进式对齐策略增强通用视听语言能力。

实施细节。我们通过基准、基本 LVLM 和 SCP 方法，实现了对 VQA 预测集的边际覆盖率和统计保证。详细设置如下：(1) 答案生成初始化。我们将所有 LVLM 的温度设置为 1.0 以进行采样，从而增加答案的多样性，每个问题生成 36 个响应。(2) 无提示推理。不使用提示；模型在原始问题上执行推理。结果被输入 Qwen2.5-3B-Instruct 进行双向区分。如果生成的集合中不存在正确答案，则舍弃样本。在获得采样答案后，一个较小的 LLM 检查答案之间的语义蕴含，将它们转换为固定长度的仅选项集。应用语义聚类来计算 $\hat{f}(y \mid x)$ 的聚类频率分布。(3) 生成不符合度分数。首先，我们按分割比例将结果拆分为校准集和测试集。我们使用第三节 B 部分中的方法生成不符合度分数，并在 100 轮中平均结果。设计了两种方法：一种具有固定的分割比例，另一种具有固定的 α ，通过测量经验误差率来评估边际覆盖率和统计保证。

3.2 经验误差率

阶段 1：固定分割比例（0.5）分析。我们使用固定分割比例为 0.5，将两个数据集按照 1:1 划分为校准集和测试集。通过比较在不同 α 下的经验错误率，我们验证了 SCP 方法严格满足用户指定错误率下的公式（5）中的覆盖保证。

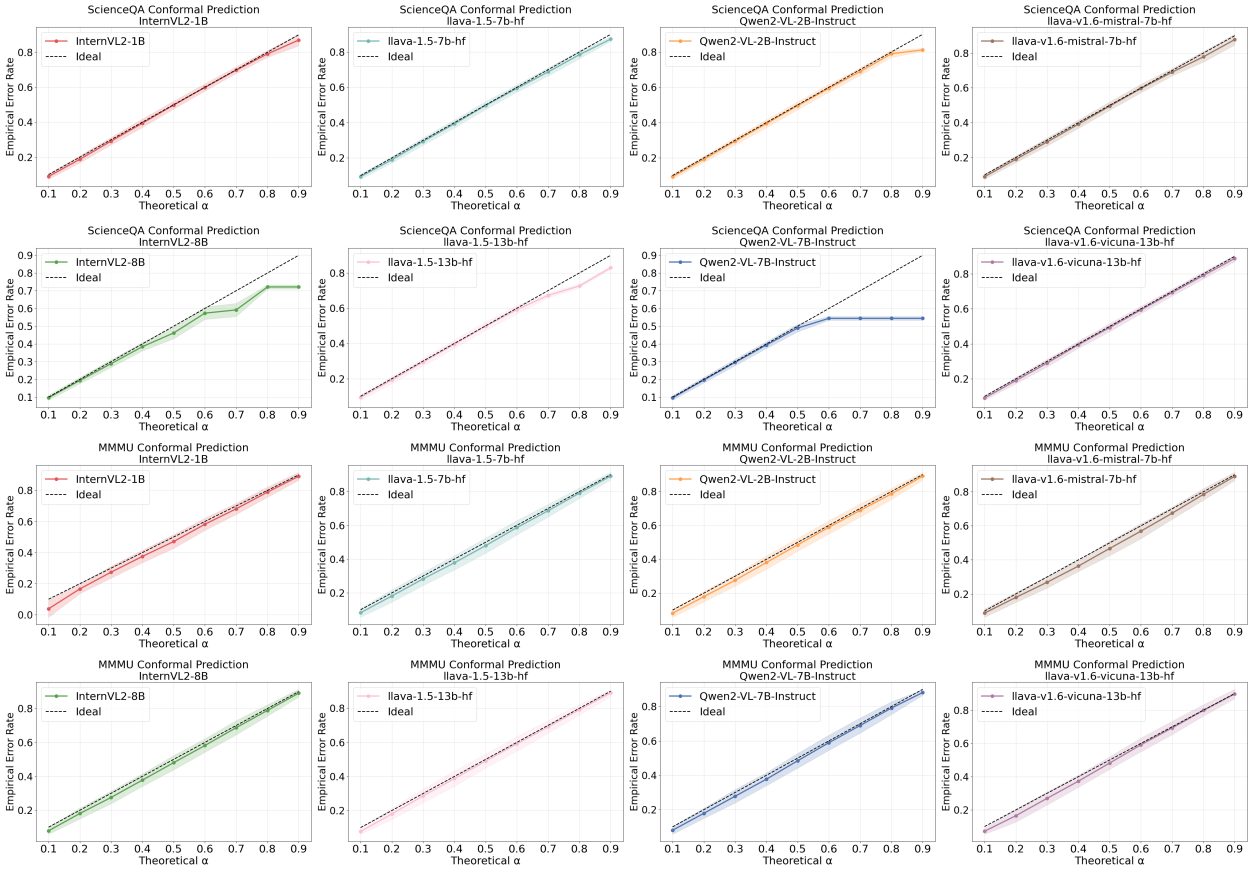

Figure 1: ScienceQA 和 MMMU 基准测试中的经验误差率。我们计算经验误差率的平均值，并在图中以实线表示。横轴表示 α 值，纵轴对应于经验误差率。图中还包括一条虚线，它遵循 $y = x$ ，表示理想情况下经验误差率应该始终低于该线。同时，我们计算经过 100 次试验的经验误差率的标准差，并将其表示为实线周围的透明带。

首先，使用 ScienceQA 数据集（图 1 上面板），实线代表平均经验错误率，透明带表示 100 次采样试验中的方差。随着 α 的增加（即，较高的允许错误阈值），测试错误率系统性上升，但始终严格低于 α ，从而满足方程（5）。在 LVLMs 中存在显著的变化：对于 InternVL2-8B，经验错误率表现出随着 α 增加而波动性增长，而 Qwen2-VL-7B-Instruct 在超过 $\alpha = 0.6$ 时显示出饱和的错误率。我们假设这种饱和现象源于接近基准真实值的近乎最优模型输出，这使得错误增加的空间变得极小。这一现象将在后续章节通过预测集大小进一步分析。

其次，MMMU 结果（图 1 下部面板）显示出比 ScienceQA 更优越的对齐，尽管数据集规模较小，但平均经验误差率较低且稳定性增强。然而，MMMU 表现出较大的标准偏差，可能是由于其更高的区分难度。至关重要的是，这两个数据集的经验误差率都严格低于 α ，重申了 SCP 方法的统计保证。

阶段 2：固定 α 并变动分割比例。在固定 α 的情况下，我们评估了基准测试和 8 个 LVLM 的分割比例下的经验误差率（表 1）。结果显示边缘覆盖率的实现，与前面章节中 MMMU 和 ScienceQA 的对比分析一致。

非一致性分数与预测集动态。根据方程（1）定义的非一致性分数的计算，建立了用户指定的误差容限 α 与预测集细粒度之间的直接关系。较低的 α 值对应于更高的允许误差率，这需要更广泛的预测集以涵盖更广泛的候选选项。此设计确保即便在放宽误差约束的情况下，模型的预测仍然在统计上有效。相反，较高的 α 值施加了更严格的误差控制，迫使模型生成紧凑的预测集，以排除模糊或低置信度的选项。α 与集大小之间的相互作用反映了一个基本的权衡：严格的误差阈值降低了预测的不确定性，但可能排除潜在有效的候选项，而宽松的阈值则以更高的经验误差率来保持包容性。

固定分割比率 (0.5) 分析。在固定的校准测试分割比率为 1:1 的情况下，我们系统地评估了两个数据集和所有八个 LVLMs 的预测集大小（图 2）。对于 ScienceQA 数据集，观察到的趋势与理论预期一致：随着 α 增加，预测集大小呈单调递减趋势，这是由逐渐严格的误差控制驱动的。值得注意的是，在较低的 α 阈值

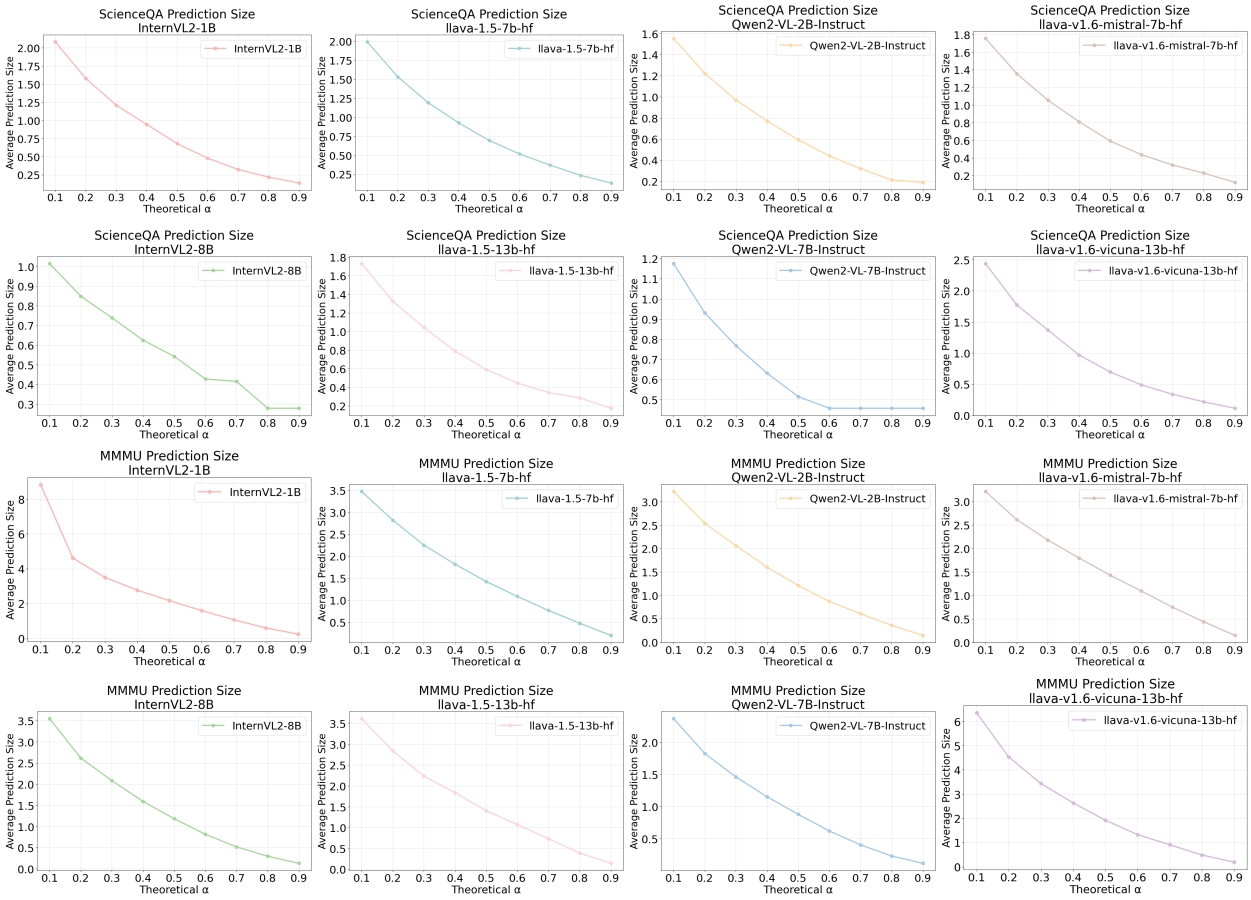

Figure 2: 在 ScienceQA 和 MMMU 基准中的预测集大小。在 ScienceQA 和 MMMU 基准中的预测集大小。我们计算预测集大小，并在图中以实线绘制。x 轴表示 α 值，而 y 轴对应预测集大小。

下，大小缩减的速度减缓，这表明在极端误差容忍度下，集合大小趋于渐近稳定。一个显著的偏差出现在 Qwen2-VL-7B-Instruct 模型中，其中预测集大小在超过 $\alpha = 0.6$ 后趋于平稳。这种饱和状态表明模型的预测与真实值几乎完全一致，留下极小的进一步减少误差的空间。此外，经验误差率（图 1）与预测集大小（图 2）之间的对称逆关系源于方程 (5) 中的边缘覆盖限制。经验误差率的微小扰动会引起预测集大小的比例调整，以确保符合规定的统计保证。

MMMU 数据集观察结果。虽然 MMMU 数据集大体反映了 ScienceQA 中观察到的趋势，但 InternVL2-1B 模型在 $\alpha = 0.1$ 上表现出异常行为，产生的预测集合不成比例地大。这一现象可以追溯到 3.2.2 节中详细描述的分位数估计过程：当计算 0.9 -分位数阈值时，大量非常接近 1 的不一致性评分会人为地增加了包含标准。因此，几乎所有候选选项都满足放宽的阈值，导致预测集合膨胀。相比之下，当不一致性评分聚集接近于 0 时，反映出模型对其预测有很高的信心，从而产生更小的预测集合。这种二分法强调了保形预测方法对不一致性评分分布的敏感性，特别是在极端 α 值的情况下。观察到的异常情况突显了在保持实用性的同时维护理论保证的过程中，稳健的评分校准和阈值选择的重要性。

我们提出了一种基于分离一致性预测的统计可靠性框架，以解决大型视觉-语言模型在视觉问答任务中产生幻觉的问题。通过采用动态门槛校准和跨模态一致性验证，我们将数据分为校准集和测试集，用不一致性得分量化输出不确定性，并从校准集分位数构建预测集。在用户指定的风险水平 α 上，我们的方法严格控制真实答案的边际覆盖。针对不同 LVLM 架构的多个多模态基准实验表明，SCP 在所有 α 值上符合理论统计保证，并且预测集大小与 α 呈反比调整，有效过滤掉低置信度输出。无需先验分布假设或模型重新训练，我们的模型无关且计算高效的框架为安全关键场景中的可靠多模态评估提供了坚实的理论和实践支持。

References

Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, 2024.

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

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Jesse C Cresswell, Yi Sui, Bhargava Kumar, and Noël Vouitsis. Conformal prediction sets improve human decision making. In *Proceedings of the 41st International Conference on Machine Learning*, pages 9439–9457, 2024.

Yusong Ke. Statistical guarantees of correctness coverage for medical multiple-choice question answering. *arXiv preprint arXiv:2503.05505*, 2025.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*, 2023.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

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

Zhiyuan Wang, Jinhao Duan, Chenxi Yuan, Qingyu Chen, Tianlong Chen, Yue Zhang, Ren Wang, Xiaoshuang Shi, and Kaidi Xu. Word-sequence entropy: Towards uncertainty estimation in free-form medical question answering applications and beyond. *Engineering Applications of Artificial Intelligence*, 139:109553, 2025a.

Margarida Campos, António Farinhas, Chrysoula Zerva, Mário AT Figueiredo, and André FT Martins. Conformal prediction for natural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 12:1497–1516, 2024.

Anastasios N Angelopoulos, Stephen Bates, et al. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023.

Zhiyuan Wang, Jinhao Duan, Lu Cheng, Yue Zhang, Qingni Wang, Xiaoshuang Shi, Kaidi Xu, Heng Tao Shen, and Xiaofeng Zhu. Conu: Conformal uncertainty in large language models with correctness coverage guarantees. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6886–6898, 2024.

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. Benchmarking llms via uncertainty quantification. *Advances in Neural Information Processing Systems*, 37: 15356–15385, 2024.

Anastasios Nikolas Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *The Twelfth International Conference on Learning Representations*, 2024.

Qingni Wang, Tiantian Geng, Zhiyuan Wang, Teng Wang, Bo Fu, and Feng Zheng. Sample then identify: A general framework for risk control and assessment in multimodal large language models. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Zhiyuan Wang, Qingni Wang, Yue Zhang, Tianlong Chen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. Sconu: Selective conformal uncertainty in large language models. *arXiv preprint arXiv:2504.14154*, 2025c.