Uncertainty-o: One Model-agnostic Framework for Unveiling Uncertainty in Large Multimodal Models

Ruiyang Zhang[1], Junhao Zhang[2], Hao Fei[3], Zhedong Zheng[1],

[1]FIC, University of Macau, [2]CSIRO Data61, [3]University of Oxford,

Correspondence: zhedongzheng@um.edu.mo

Abstract

Large Multimodal Models (LMMs), characterized by diverse modalities, architectures, and capabilities, are rapidly emerging. In this paper, we aim to address three key questions: (1) Can we unveil the uncertainty of LMMs in a model-agnostic manner, regardless of the modalities, architectures, and capabilities involved? (2) How to reveal the aleatoric uncertainty in multimodal prompts, which stems from inherent observational noise and leads to confusion in LMMs. (3) LMM inherent epistemic uncertainty ultimately reflects in multimodal responses, how can we mine and quantify it from these multimodal responses? To answer these questions, we propose Uncertainty-o: (1) Uncertainty-o is one model-agnostic framework for unveiling uncertainty in LMMs, independent of their modalities, architectures, and capabilities. (2) We empirically explore multimodal prompt perturbation for revealing prompt aleatoric uncertainty, providing our insights and findings. (3) We propose multimodal semantic uncertainty, which quantifies LMM inherent epistemic uncertainty from multimodal responses. Experiments with 18 benchmarks across various modalities and 10 LMMs (open- and closed-source) validate the efficacy of Uncertainty-o in reliably estimating LMM uncertainty, thereby facilitating various downstream tasks, including hallucination detection, hallucination mitigation, and uncertainty-aware Chain-of-Thought (CoT). The code is released at https://github.com/Ruiyang-061X/Uncertainty-o.

1 Introduction

Large Multi-Modal Models (LMMs) (Caffagni et al., 2024; Carolan et al., 2024) significantly expand the boundaries of human-made intelligence. **(1) Rich modalities are continuously integrated.** Beyond the initial success of the 'Text-in-Text-out' paradigm (Achiam et al., 2023), LMMs now incorporate images (Liu et al., 2023b; Rombach et al., 2022), videos (Lin et al., 2023; Luo et al., 2023), audio (Radford et al., 2023; Kreuk et al., 2022), point clouds (Xu et al., 2024; Nichol et al., 2022), IMU (Girdhar et al., 2023), fMRI (Han et al., 2024), and more (Wu et al., 2024b). **(2) Diverse model architectures are rapidly evolving.** The foundational designs of Large Language Models (LLMs) (Achiam et al., 2023) and Diffusion Models (DMs) (Rombach et al., 2022) have been extensively adapted (Liu et al., 2023b) and integrated (Wu et al., 2024c) to create more advanced and complex structures. **(3) Distinct model capabilities have significantly improved.** LMMs can now comprehend intricate multimodal contexts

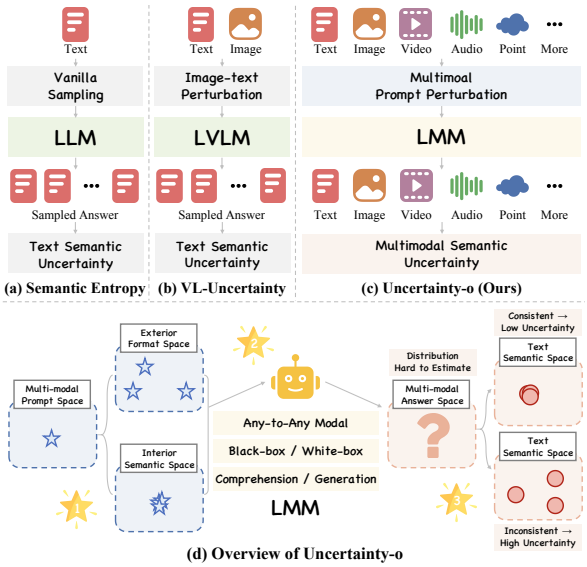


Figure 1: **Comparison with Previous Works and Overview of Uncertainty-o.** (a,b) Existing studies focus on single modality or two modalities (Zhang et al., 2024c); (c) In contrast, our Uncertainty-o captures uncertainty in *large multimodal models* in a model-agnostic manner. It achieves reliable uncertainty estimation via *multimodal prompt perturbation*. (d) In particular, we harness *multimodal semantic uncertainty*, which maps multimodal answer space into a unified text semantic space. We then estimate the uncertainty according to the text consistency.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 2270–2285
October 24-29, 2026 ©2026 Association for Computational Linguistics

and instructions (Lyu et al., 2024), while simultaneously generating diverse and high-fidelity content (Wu et al., 2024b).

However, large models (Liu et al., 2023b; Achiam et al., 2023) still follow a vanilla 'Prompt-In-Response-Out' pipeline, which remains raw and primitive compared to human thinking process (Amayuelas et al., 2023; Pi et al., 2024). When asked a question or given a task, humans not only provide an answer or take action, but also experience internal feelings: confidence or uncertainty, ease or difficulty (Peterson et al., 2019; Collins et al., 2023). These feelings guide us in predicting outcomes and adjusting our plans accordingly (Callaway et al., 2022). To bridge this gap and explore the psychology of LMMs (Ke et al., 2024), we focus on one critical property: LMM uncertainty (Aichberger et al., 2024).

In this paper, we aim to answer 3 key questions: **(1) Can we unveil LMM uncertainty in a model-agnostic manner?** Tailoring specific design for different LMMs can lead to a complex and inflexible framework, limiting adaptability to revolving LMMs (Caffagni et al., 2024). **(2) How to reveal the aleatoric uncertainty in multimodal prompts?** Notably, ambiguity and observational noise in prompt can significantly lead to confusion of LMM. Moreover, interaction and relationship between various introduced modalities exacerbate this phenomenon (Lu, 2023). **(3) How to mine uncertainty from multimodal responses where LMM manifests its epistemic uncertainty?** The cues are hidden in these multimodal responses (Rombach et al., 2022; Kreuk et al., 2022; Nichol et al., 2022; Wu et al., 2024b), *e.g.*, generated images (Rombach et al., 2022) or point clouds (Lee and Benes, 2024), requiring effective methods to extract and quantify LMM epistemic uncertainty.

In response to those substantial questions, we propose Uncertainty-o (see Fig. 1): **(1) Uncertainty-o, in a model-agnostic manner, unveils uncertainty in LMMs, regardless of the modalities involved, inherent architecture, and capability focus.** Notably, Uncertainty-o is also scalable and can incorporate emerged modalities, architectures, and capabilities. **(2) We empirically explore multimodal prompt perturbation for revealing prompt aleatoric uncertainty.** By causing variance at the prompt end, we can observe fluctuations in LMM responses. Higher degree of fluctuation indicates higher level of aleatoric uncertainty, illustrating challenge and complexity of prompt. We provide a comprehensive empirical analysis of perturbation for 5 commonly used modalities, including text, image, video, audio, and point cloud. Finally, we offer our findings and insights, which are modality-agnostic and can serve as a cookbook when designing perturbation for other potential modalities. **(3) We propose multimodal semantic uncertainty, which can mine LMM inherent epistemic uncertainty from multimodal responses.** Through mapping multimodal responses from LMM to text semantic space, the observation of answer semantic distribution becomes addressable. The answer distribution entropy in the text semantic space serve as an effective indicator of epistemic uncertainty. Furthermore, uncertainties of each modality are merged to reveal overall LMM uncertainty.

Through extensive experiments with 18 multimodal benchmarks (including 5 modalities) and 10 LMMs (both open- and closed-source), we validate effectiveness of Uncertainty-o in accurately capturing uncertainty of LMMs. Reliable uncertainty estimation (see Fig. 2) facilitates various downstream tasks, including hallucination detection, hallucination mitigation, and uncertainty-aware CoT. In summary, our contributions are as follows:

- **Unified Framework to Unveil LMM Uncertainty.** We propose Uncertainty-o, which, in a model-agnostic manner, reveals one critical property in LMM psychology: uncertainty. Uncertainty-o can seamlessly handle LMMs with various modalities, complex model architectures, and distinct capacity emphases.
- **Cookbook for Multimodal Prompt Perturbation.** We propose multimodal prompt perturbation and empirically explore how to perturb the prompt to maximize its contribution to revealing prompt *aleatoric uncertainty*.
- **Multimodal Uncertainty in Semantic Space.** We propose multimodal semantic uncertainty, explicitly enabling the mining of inherent *epistemic uncertainty* from multimodal generated content via semantic space mapping.

2 Uncertainty-o

2.1 Multimodal Prompt Perturbation

Definition 3.1 (Response Semantic Distance). *The LMM M generate responses for prompts x_i*

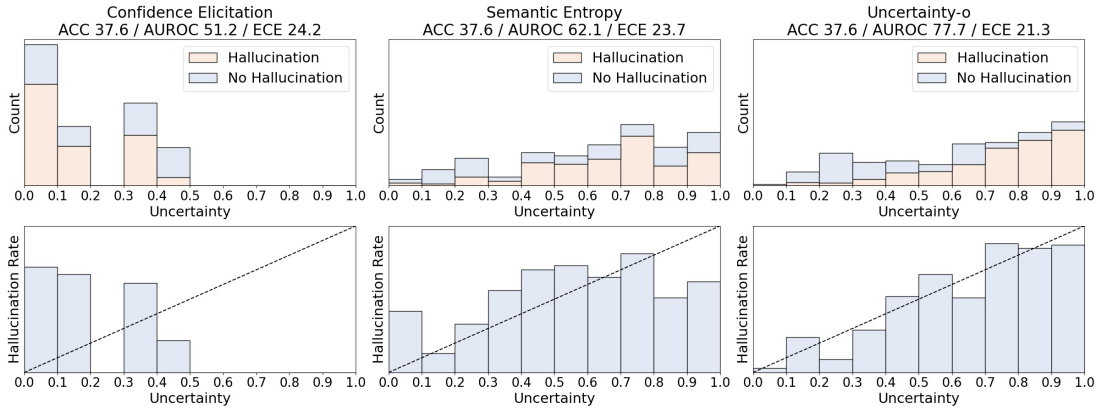


Figure 2: **Reliability of Our Estimated Uncertainty.** From comparison with previous methods, we observe that: (1) Compared to Confidence Elicitation (Xiong et al., 2023), Uncertainty-o effectively avoids the over-confidence problem, which inaccurately assigns low uncertainty (or high confidence) to hallucinatory responses. (2) Compared to Semantic Entropy (Farquhar et al., 2024), Uncertainty-o provides more reliable uncertainty estimation, *e.g.*, the uncertainty is more closely aligned with the error rate in each bin. Results from OneLLM (Han et al., 2024) on ClothoV2 (Drossos et al., 2020).

and x_j (perturbed from original prompt x). For perturbed prompts, the predictions from LMM is $y_i = M(x_i)$, and $y_j = M(x_j)$, respectively. We focus on the LMM's Epistemic Uncertainty, which is the uncertainty in the model parameters, Var$(\theta|x)$. Define the prediction difference $D(x)$ as:

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

where $\|\cdot\|$ denotes ℓ_2 norm.

Proposition 3.2. *Response semantic distance $D(x)$ from prompt perturbation is proportional to the square root of the LMM uncertainty:*

$$D(x) \propto \sqrt{\text{Var}(\theta|x)}. \tag{2}$$

Therefore, the prediction semantic distance $D(x)$ can serve as a measure of the LMM uncertainty under given context. See Appendix.D for detailed proof.

Based on *Proposition 3.2*, we propose semantic-equivalent perturbation for multimodal prompts to induce variations in LMM responses, enabling the capture of LMM uncertainty (see Fig. 3). This approach generates semantic-equivalent variations of original prompts across multiple modalities and feeds them into the LMM. The variance in the sampled responses could effectively illustrate LMM uncertainty. Specifically, we propose multimodal prompt perturbation across five modalities, including text, image, audio, video, and point cloud. Notably, we adopt a progressive perturbation strategy, where each prompt is gradually perturbed from low to high degrees, introducing increasing levels of challenge to the LMM. Perturbations across different modalities are applied simultaneously to further enhance uncertainty estimation.

In practice, different perturbation techniques are applied depending on the modality. For text prompts, both LLM-based rephrasing and rule-based methods, such as word swapping, are used. For image prompts, spatial transformations like rotation and attribute distortions such as blurring or brightness adjustments are applied. In the case of audio prompts, key elements like volume, pitch, and timbre are perturbed, with additional temporal adjustments such as temporal shifting. For video prompts, both spatial and temporal perturbations are applied. Temporal perturbations include techniques like temporal cropping and frame dropping, while spatial perturbations involve applying image perturbation techniques to each video frame. For point cloud prompts, 3D characteristics are perturbed with techniques like random sampling and point cloud jittering.

2.2 Multimodal Semantic Uncertainty

We quantify the variance of these sampled answers to capture the inherent uncertainty of the LMM. Instead of simply calculating variance based on format differences, we focus on variations in answer semantics. To estimate the variance of sampled answer semantics, the most critical procedure lies in semantic checking between two generated answers. Given this challenge, especially for higher-level modalities, *e.g.*, video, we propose utilizing an off-the-shelf LMM captioner to extract answer semantics and summarize the key points into text captions. Consequently, the challenging multimodal seman-

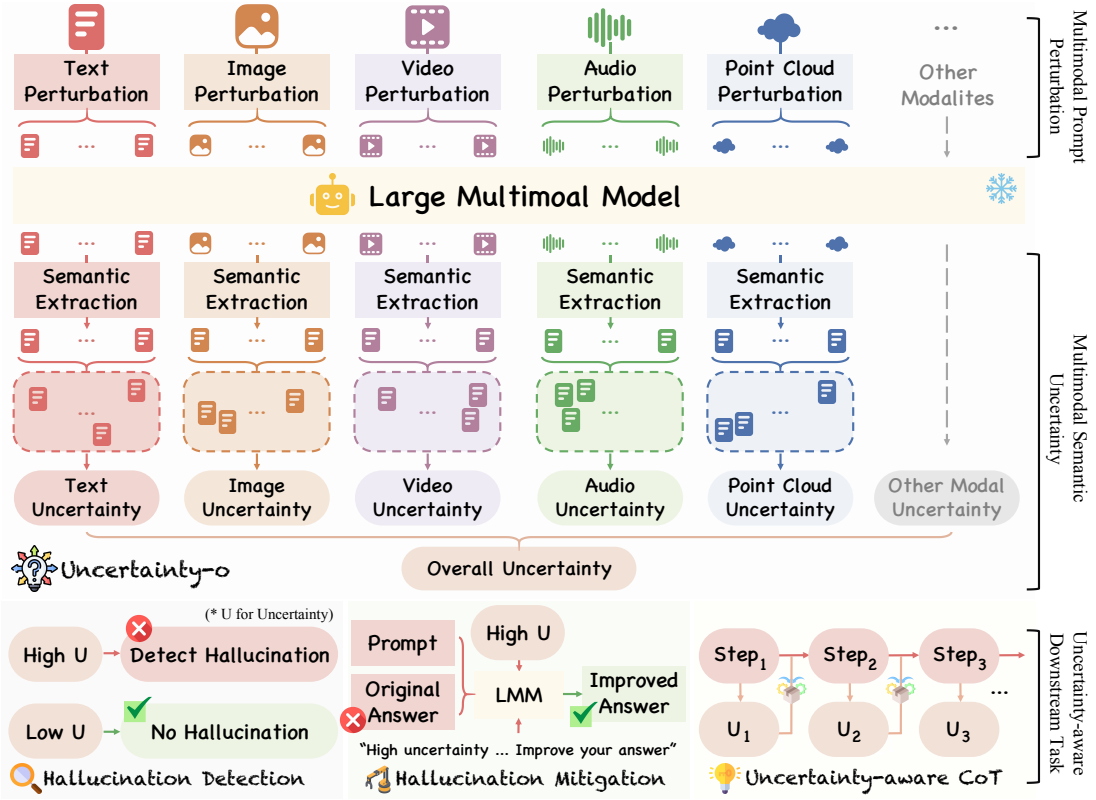


Figure 3: **Pipeline of Our Uncertainty-o.** Given a multimodal prompt and large multimodal models, we perform multimodal prompt perturbation to generate diverse responses. Due to the inherent epistemic uncertainty of these models under perturbation, varied responses are typically obtained. To quantify this uncertainty, we apply semantic clustering on the collected responses and compute their entropy. Specifically, responses are grouped into semantically similar clusters, and the entropy across these clusters is calculated as the final uncertainty measure. Higher entropy indicates greater variability in responses, suggesting lower confidence, while lower entropy reflects higher consistency and thus higher confidence. Finally, we could leverage the uncertainty for hallucination detection, hallucination mitigation and facilitate the chain of thoughts (CoT).

tic checking is converted to the more manageable text semantic checking (see Fig. 3).

For those text captions, we directly prompt the LLM to iteratively check pairs of text answers. For texts with similar semantics, we cluster them into one group. After obtaining the semantic clusters of sampled answers, we analyze the discrete distribution of answer semantics and calculate the distribution entropy. Moreover, we normalize the calculated entropy compared to the maximum entropy decided by the sampling time, and obtain estimated LMM uncertainty, with higher values indicating greater LMM uncertainty:

$$u_m = -\sum_{i=1}^{n} p_i \log(p_i) \quad \text{where} \quad p_i = \frac{c_i}{C}, \quad (3)$$

where c_i is answer count for i-th group, C is the total answer count, u_m is the estimated uncertainty for certain modal m. Finally, uncertainties from all answer modalities is averaged to obtain final LMM uncertainty u.

2.3 Uncertainty-Aware Downstream Task

Hallucination Detection. We treat answers with low uncertainty as non-hallucinatory and those with high uncertainty as hallucinatory. In practice, we first infer LMM using the original prompt to obtain an initial answer. This answer is then compared with the ground truth to determine whether it contains hallucinations. Next, we apply the proposed multimodal prompt perturbation and multimodal semantic uncertainty to estimate LMM uncertainty for its initial answer. Finally, we evaluate the alignment between the estimated uncertainty and actual hallucination using 3 different metrics: AUROC, AURAC, and ECE.

Hallucination Mitigation. We formulate hallucination mitigation as an uncertainty-aware, two-stage revision process. In the first stage, we prompt the LMM to generate an initial answer, which is then assigned an uncertainty score estimated by Uncertainty-o. After constructing an answer pool with corresponding uncertainties, we select the

Image Comprehension

Method	LLaVABench			MMVet			CoCoCap		
	AUROC↑	AURAC↑	ECE↓	AUROC↑	AURAC↑	ECE↓	AUROC↑	AURAC↑	ECE↓
Confidence Elicitation	54.0	78.1	11.3	52.3	55.8	13.1	51.2	71.0	23.1
GAVIE	45.3	75.2	23.8	55.4	45.2	9.4	57.7	79.6	28.9
Semantic Entropy	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			MSVDQA			NextQA		
	AUROC↑	AURAC↑	ECE↓	AUROC↑	AURAC↑	ECE↓	AUROC↑	AURAC↑	ECE↓
Confidence Elicitation	53.5	32.6	12.2	51.0	32.8	16.1	51.5	65.9	10.9
GAVIE	50.1	33.9	18.2	39.6	30.1	20.5	47.4	60.8	11.4
Semantic Entropy	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			ClothoAQA			AudioCaps		
	AUROC↑	AURAC↑	ECE↓	AUROC↑	AURAC↑	ECE↓	AUROC↑	AURAC↑	ECE↓
Confidence Elicitation	51.2	34.1	24.2	49.2	43.5	29.0	50.9	30.2	18.6
GAVIE	51.1	31.9	30.6	53.1	46.7	23.9	47.7	34.8	19.7
Semantic Entropy	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			ShapeNet			Objaverse		
	AUROC↑	AURAC↑	ECE↓	AUROC↑	AURAC↑	ECE↓	AUROC↑	AURAC↑	ECE↓
Confidence Elicitation	47.1	78.0	36.1	46.4	72.1	39.7	49.5	28.0	23.1
GAVIE	46.0	84.5	43.5	51.1	75.4	37.5	49.8	35.1	29.4
Semantic Entropy	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: **Comprehension Hallucination Detection Results.** Uncertainty-o consistently outperforms strong baselines by a clear margin in LMM comprehension hallucination detection. For LMMs, We utilize InternVL (Chen et al., 2024a), VideoLLaMA (Zhang et al., 2023), OneLLM (Han et al., 2024), PointLLM (Xu et al., 2024) for image, video, audio, point cloud comprehension. The best results are in **bold**, and the second-best results are underlined.

Image Comprehension

Method	LLaVABench			MMVet			CoCoCap		
	AUROC↑	AURAC↑	ECE↓	AUROC↑	AURAC↑	ECE↓	AUROC↑	AURAC↑	ECE↓
Confidence Elicitation	51.8	72.6	15.4	49.5	67.8	11.7	56.2	70.4	9.2
GAVIE	53.7	75.3	13.9	49.0	65.6	10.4	58.7	71.8	10.4
Semantic Entropy	56.8	82.7	13.6	52.8	64.9	11.1	59.2	73.3	8.9
Uncertainty-o (ours)	**61.3**	**87.4**	**12.8**	**56.1**	**75.8**	**8.6**	**61.4**	**76.8**	**7.8**

Video Comprehension

Method	MSRVTTQA			MSVDQA			NextQA		
	AUROC↑	AURAC↑	ECE↓	AUROC↑	AURAC↑	ECE↓	AUROC↑	AURAC↑	ECE↓
Confidence Elicitation	54.7	39.9	11.5	48.6	29.7	15.4	57.2	63.8	13.7
GAVIE	55.9	40.1	15.6	54.3	36.4	12.8	60.3	65.3	10.4
Semantic Entropy	56.8	41.5	12.8	51.8	35.2	13.7	63.9	67.3	8.9
Uncertainty-o (ours)	**60.1**	**47.5**	**8.9**	**57.3**	**39.2**	**9.4**	**72.8**	**70.3**	**4.9**

Table 2: **Hallucination Detection for Closed-Source LMMs.** We leverage GPT4o (Hurst et al., 2024) and QwenVLMax (Bai et al., 2023) for image and video comprehension respectively.

top-K answers with the highest uncertainty scores, as these are the most likely to contain hallucinations. In the second stage, we refine these high-uncertainty answers by prompting the LMM with the original context, initial answer, and uncertainty score. We explicitly tell the LMM that its initial answer has a high uncertainty score and instruct it to revise the response due to a potential hallucination. The revised answer is then used as the LMM final output:

$$y_{\text{final}} = M(x', y_{\text{initial}}, u), \quad y_{\text{initial}} \in \mathbf{Y}, \quad (4)$$

where $\mathbf{Y}$ contains selected answers with the highest uncertainty scores. x', y_{initial}, and u refer to the updated prompt, initial answer, and estimated uncertainty, respectively. y_{final} is the revised answer. This iterative process helps the LMM refine its responses and improve accuracy.

Method	*Medical Diagnosis* AUROC↑	AURAC↑	ECE↓
Confidence Elicitation	42.1	30.9	24.5
GAVIE	44.2	33.6	31.5
Semantic Entropy	52.4	34.5	24.1
Uncertainty-o (ours)	**59.0**	**40.6**	**15.3**

Method	*Embodied Robot* AUROC↑	AURAC↑	ECE↓
Confidence Elicitation	40.1	49.6	18.9
GAVIE	45.2	53.6	23.5
Semantic Entropy	56.2	55.9	21.1
Uncertainty-o (ours)	**58.2**	**66.5**	**12.2**

Table 3: **Hallucination Detection for Safety-Critic Tasks.** We harness MIMICCXR (Johnson et al., 2019) for medical image diagnosis and OpenEQA (Majumdar et al., 2024) for video embodied QA.

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

Table 4: **Generation Hallucination Detection Results.** We utilize StableDiffusion (Rombach et al., 2022), VideoFusion (Luo et al., 2023), AnyGPT (Zhan et al., 2024), RGB2point (Lee and Benes, 2024) for image, video, audio, point cloud generation.

Uncertainty-Aware Chain-of-Thought. We leverage Uncertainty-o to enhance the vanilla CoT process. Self-reflection is crucial for LMMs during CoT, and our estimated uncertainty serves as a guiding signal to trigger self-reflection. Specifically, at each reasoning step, in addition to generating responses, we also estimate answer uncertainty. In the subsequent step, we incorporate the previous answer and its uncertainty into the context, explicitly prompting LMM to reflect on its reasoning process when it finds uncertainty is high, as:

$$y_t, u_t = M(C_t), \quad C_{t+1} = C_t \cup \{y_t, u_t\}, \quad (5)$$

where y_t and u_t is answer and uncertainty, given the context C_t at step t. This uncertainty-aware procedure facilitates a more detailed and cautious reasoning process.

Video Comprehension		
VideoLLaMA	MSRVTTQA Acc.	MSVDQA Acc.
Direct Inference	62.9	70.5
+ Uncertainty-o	**67.2**	**72.1**
Audio Comprehension		
OneLLM	ClothoAQA Acc.	ClothoV2 Acc.
Direct Inference	57.1	45.5
+ Uncertainty-o	**59.2**	**51.9**

Table 5: **Hallucination Mitigation Results.** Our uncertainty-guided revision effectively mitigates answer hallucination.

Image Comprehension		
InternVL	MMVet Acc.	LLaVABench Acc.
Vanilla CoT	23.8	61.8
Uncertainty-Aware CoT	**27.1**	**63.1**
Point Cloud Comprehension		
PointLLM	ModelNet Acc.	ShapeNet Acc.
Vanilla CoT	48.3	52.6
Uncertainty-Aware CoT	**51.8**	**54.2**

Table 6: **Uncertainty-Aware Chain-of-Thought Results.** Our estimated uncertainty enriches the reasoning context and facilitates more thorough thinking process, leading to improved answer accuracy.

3 Experiment

3.1 Comparison with State-of-the-Arts

Comprehension Hallucination Detection. We compare Uncertainty-o with previous state-of-the-art methods in image, video, audio, and point cloud comprehension hallucination detection tasks (see Tab. 1). Uncertainty-o yields significant and consistent performance improvements over strong baselines. This validates the general efficacy of the proposed multimodal prompt perturbation in better eliciting and capturing uncertainty, thereby facilitating reliable hallucination detection. Notably, Uncertainty-o surpasses baselines by 15.6% (AUROC) on ClothoV2, 9.5% (AURAC) on MMVet, 7.9% (ECE) on ModelNet, and 4.1% (AURAC)

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: **Ablation of Perturbation Combination.** We observe that combining perturbations for multimodal prompts yield better results. Single modal prompt perturbation could also yield performance improvements. Results from comprehension hallucination detection on MSRVTTQA (Xu et al., 2017), AudioCaps (Kim et al., 2019), and Objaverse (Deitke et al., 2023). Row with gray shading is our default setting.

on NextQA. We also observe consistent improvement for closed-source LMMs, which validate the robustness of our Uncertainty-o (see Tab. 2). We further present a comparison of detection efficacy in the Medical Diagnosis and Embodied Robot domains (see Tab. 3). We observe that Uncertainty-o effectively handles complexity in real-world applications and outperforms baselines by clear margins.

Generation Hallucination Detection. We also present results on generation hallucination detection tasks across image, video, audio, and point cloud modalities (see Tab. 4). We adapt GAVIE (Liu et al., 2023a) here and extend text-only semantic entropy in (Farquhar et al., 2024) with our multimodal semantic uncertainty. We observe that Uncertainty-o consistently shows clear improvements over these baselines. For example, our Uncertainty-o achieves 23.2% higher AUROC, 10% higher AURAC in the point cloud generation task. These results validate that the interaction between multimodal prompt perturbation and multimodal semantic uncertainty can effectively achieve more accurate hallucination detection.

Hallucination Mitigation. We report comprehension hallucination mitigation results on four video and audio benchmarks (see Tab. 5). With our uncertainty-guided revision process, we observe consistent accuracy improvement.

Uncertainty-Aware CoT. We show the comparison between Uncertainty-Aware CoT and vanilla CoT (see Tab. 6). Incorporating uncertainty into the reasoning context improves answer accuracy across different LLMs.

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: **Pairing Order for Perturbed Prompts.** Pairing prompts with similar degrees of perturbation from different modalities yields the best results. Results from comprehension hallucination detection on MSRVTTQA. Our default setting is in gray.

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: **Ablation of Sampling Time.** Moderate sampling time enables the most effective uncertainty estimation. We report MSRVTTQA comprehension hallucination detection results. Our default setting is highlighted with gray shading.

3.2 Ablation Studies and Further Discussions

Empirical Study on Multimodal Prompt Perturbation. We present an empirical comparison of 24 different prompt perturbations for video, audio, and point cloud (see Fig. 4). For video perturbation, we find that perturbations that preserve the original video semantics, *e.g.*, adjust video speed, achieve superior results. For audio perturbation and point cloud perturbation, we observe similar pattern.

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

Perturbation Combination. We report the ablation of perturbation combinations in Tab. 7. Concurrently perturbing multimodal prompts yields the best results.

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

Pairing Order. We present an ablation of pairing order in Tab. 8. For multimodal prompts perturbed to various degrees, pairing them with a similar degree in a progressive manner yields the best results. This approach creates a prompt set with challenge levels from low to high, effectively eliciting LMM uncertainty.

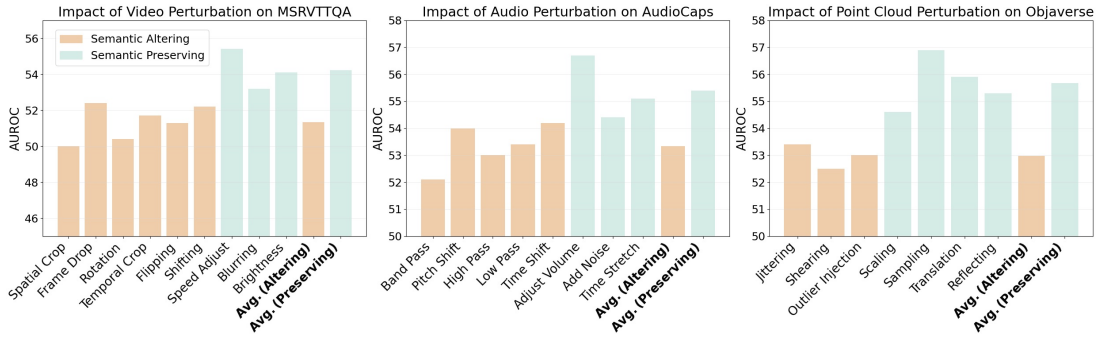


Figure 4: **Empirical Comparison of Different Prompt Perturbations.** On average, semantic-preserving perturbations are more effective for eliciting LMM uncertainty than semantic-altering ones. Results from comprehension hallucination detection for video, audio, point.

LMM Captioner	AUROC↑	AURAC↑	ECE↓
OneLLM-7B	59.5	74.5	8.3
Qwen3-VL-30B-A3B-Thinking	59.8	74.6	8.1
GPT-5	60	74.9	7.8
Gemini-3-Flash	59.7	74.6	7.9

Table 10: **Sensitivity Analysis of LMM Captioner.** We observe that the hallucination detection performance of Uncertainty-o shows low sensitivity to the adopted captioner model. When replacing our adopted OneLLM-7B with stronger proprietary models, the performance of Uncertainty-o improves only marginally.

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

Sampling Time. We illustrate the ablation of sampling time during the uncertainty estimation process (see Tab. 9).

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

Sensitivity Analysis of LMM Captioner. We conduct a sensitivity analysis of the captioner utilized to extract the semantics of multimodal responses (see Tab. 10). Specifically, we perform analysis on the image generation hallucination detection task and conduct experiments on Flickr with StableDiffusion.

Qualitative Results. We present successful hallucination detection cases for both comprehension and generation scenarios in Fig. 5.

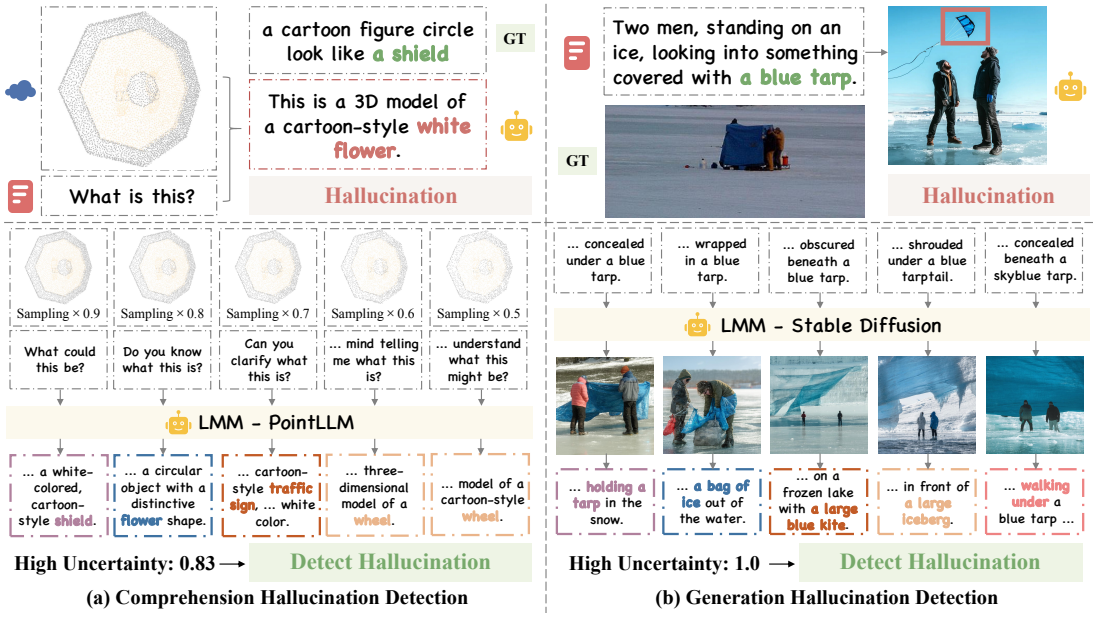
4 Related Work

Large Multimodal Model. Large Multimodal Models (LMMs) (Caffagni et al., 2024) perceive and generate content across a wide range of modalities (Caffagni et al., 2024; Wu et al., 2023; Zhang et al., 2024a,b), such as text, image, audio, video, point cloud, and more. On one hand, various works (Liu et al., 2023b) focus on enhancing multimodal comprehension ability (Fei et al., 2024), emphasizing the understanding of complexity (Wu et al., 2024a) and interaction (Zhang et al., 2024d) between prompts from various modalities. Another line of research focuses on generating multimodal content (Rombach et al., 2022), with an emphasis on the fidelity (Song et al., 2024) and diversity (Fang et al., 2024a) of multimodal responses. Moreover, recent works (Wu et al., 2024b; Zhan et al., 2024) try to build 'Any-to-Any' capability in large models, where models can simultaneously comprehend complex contexts and generate high-fidelity content following those contexts.

Uncertainty in LMM. MAP (Ji et al., 2022) proposes a probabilistic distribution encoder to model uncertainty during large model pretraining. VL-Uncertainty (Zhang et al., 2024c) leverages semantic-equivalent perturbations of visual and textual prompts to better capture MLLM uncertainty. DropoutDecoding (Fang et al., 2024b) treats token distribution discrepancies from the average token distribution as uncertainty, dropping tokens with high uncertainty. Calibration-MLLM (Chen et al., 2024b) utilizes calibration between different stages of MLLM, such as before and after visual fine-tuning, to reveal uncertainty. UAL (Wang et al., 2024) utilizes uncertainty for large model alignment and improves token convergence in feature space.

5 Conclusion

In this paper, we study LMMs reliability and explore the intrinsic uncertainty in LMMs. We propose Uncertainty-o, a model-agnostic framework that unveils uncertainty in LMMs and simultane-



(a) Comprehension Hallucination Detection (b) Generation Hallucination Detection

Figure 5: **Qualitative Results of Successful Hallucination Detection.** Uncertainty-o is proficient at accurately detecting multiple hallucinations, including comprehension (point cloud QA (Deitke et al., 2023)) and generation tasks (image generation (Plummer et al., 2015)) based on the uncertainty value.

ously processes five different modalities (visual, auditory, textual, video, and point clouds). Specifically, we first introduce multimodal prompt perturbation and empirically demonstrate that semantic-preserving perturbations are more effective for capturing LMM uncertainty. Then, we introduce multimodal semantic uncertainty based on the imposed perturbations, effectively mining uncertainty from multimodal generated responses. Extensive experiments and analysis validate superiority of Uncertainty-o in accurately estimating LMM uncertainty and facilitating various downstream tasks.

Limitations

For our experimental design, we focus on 5 commonly used modalities: text, image, video, audio, and point cloud. We do not cover the full spectrum of existing modalities, such as fMRI. In addition, due to the rapid development of foundation models, we do not include some recently proposed models, such as GPT-5.2 or Gemini-3.1-Pro. However, Uncertainty-o is designed as model-agnostic framework for estimating uncertainty in Large Multimodal Models and provides comprehensive cookbook for more effective uncertainty estimation process. As a result, it is quite convenient to include other existing modalities. Moreover, because Uncertainty-o is model-agnostic and does not depend on model internal architecture or states, it can also directly estimate uncertainty for recently proposed foundation models.

Acknowledgments

We acknowledge supports from Guangdong Basic and Applied Basic Research Foundation 2025A1515012281, Nanjing Municipal Science and Technology Bureau 202401035 and University of Macau MYRG-GRG2024-00077-FST-UMDF.

References

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

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

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

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

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

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

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

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024a. 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.

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

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

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

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

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

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

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

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

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

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

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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

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

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

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. 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.

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

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

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

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

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

Zhou Lu. 2023. A theory of multimodal learning. *NeurIPS*, 36:57244–57255.

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

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

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, and 1 others. 2024. Openeqa: Embodied question answering in the era of foundation models. In *CVPR*, pages 16488–16498.

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

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

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

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

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

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

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

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

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

Junda Wu, Zhehao Zhang, Yu Xia, Xintong Li, Zhaoyang Xia, Aaron Chang, Tong Yu, Sungchul Kim, Ryan A Rossi, Ruiyi Zhang, and 1 others. 2024a. Visual prompting in multimodal large language models: A survey. *arXiv:2409.15310*.

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

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

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

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

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

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, and 1 others. 2024. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv:2402.12226*.

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

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

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

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

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

A Dicussion

Why is semantic-equivalent perturbation more effective than inequivalent perturbation? In semantic-equivalent perturbation, the semantics of all perturbed prompts remain the same. As a result, any fluctuation in answers of LMM directly reflects its uncertainty. On the other hand, semantic-inequivalent perturbation alters the original semantics, and changes in the answers are inevitable, making it difficult to accurately reflect the inherent uncertainty of the LMM.

Why is prompt perturbation superior to vanilla sampling for capturing LMM uncertainty? Multimodal prompt perturbation challenges the LMM with prompts that vary at different levels, with larger variations of answers indicating higher uncertainty. In contrast, vanilla sampling typically uses a high temperature during sampling, which tends to induce high uncertainty across all samples. For instance, the LMM could be highly confident in its answer, but the high temperature inevitably causes answer variation, leading to an artificially high uncertainty.

Can multimodal semantic uncertainty extend to more modalities? Yes. Uncertainty-o proposes a general pipeline for LMM uncertainty estimation. This pipeline is orthogonal to and can benefit from iterations of LMM. With a more general LMM, which is capable of parsing the semantic content of additional modalities into textual descriptions, Uncertainty-o can seamlessly estimate uncertainty for those answers in more modalities.

Why not directly utilize LMM to check whether two generated multimodal responses are semantically equivalent? Current LMMs still lack proficient ability for multimodal semantic checking. We have attempted two methods to utilize LMM for direct multimodal data semantic checking: (1) Combine two multimodal answers into one and prompt the LMM to check whether the answers are similar. For example, merge two point clouds into one and conduct shifting so that they do not overlap. However, current LMMs show unsatisfactory performance. (2) Directly input two answers into LMM and prompt it to check. We find that many LMMs still do not support multi-input, such as understanding two point clouds simultaneously.

B Setting

Benchmark. We conduct experiments with 18 benchmarks. For the comprehension hallucination task, the benchmarks we adopt are typically QA-type. For images, we utilize MMVet, LLaVABench, and CoCoCap. For video benchmarks, we leverage MSRVTTQA, MSVDQA, and NextQA. For the audio modality, we use ClothV2 and AudioCaps. For point cloud benchmarks, we utilize PointCap and ModelNet. For generation hallucination detection, we use caption-type benchmarks and employ captions as generation prompts. For image generation, we utilize Flickr. For video generation, we utilize MSRVTT. For audio generation, we utilize VCTK. For point cloud generation, we utilize Pix3D, aiming for image-to-point cloud generation. For the safety-critical task, we utilize MIMICCXR and OpenEQA. MIMICCXR is a benchmark of chest X-ray images annotated with medical diagnoses by radiologists. OpenEQA is an in-house video benchmark from an embodied robot's perspective, accompanied by text-based questions.

LMM. We experiment with 10 different LMMs capable of understanding and generating various modalities of data. For closed-source LMMs, we utilize GPT-4o and QwenVLMax. For open-source LMMs, InternVL is a large vision-language model that takes image-text prompts and answers with text. VideoLLaMA takes videos as prompts and generates text answers. OneLLM is proficient in processing various modal inputs, *e.g.*, audio and point clouds, and outputs text-only responses. PointLLM is specifically designed for real-world point cloud understanding. StableDiffusion serves as a text-based image generation model. VideoFusion is an efficient video generation model, taking text as a prompt. AnyGPT is capable of generating audio and visual data. RGB2point realizes point cloud generation by taking a single image.

Implementation Details. The hallucination detection procedure contains 3 key phases: (1) Initial answer obtaining. The tested LMM is first given the context to generate an initial output. We keep the variation hyper-parameter low at this stage, *e.g.*, low temperature for LLM-based large models and high guidance scale for Diffusion models. The initial answer is compared with the ground truth to decide whether it contains hallucinations. (2) Uncertainty estimation. Initial contexts are perturbed by our proposed multimodal prompt perturbation, according to the modalities they contain. All modal prompts are perturbed simultaneously to various degrees. Different levels of perturbed prompts enable the sampling process. For sampled textual answers, an off-the-shelf LLM, Qwen2.5-7B, clus-

ters them by semantics and calculates entropy as uncertainty. For other modalities' answers, *e.g.*, images, we use OneLLM-7B as a captioner to convert those answers into concise captions. Then, the uncertainty calculation is simplified to vanilla textual semantic uncertainty. (3) Hallucination detection. With hallucination labels from (1) and estimated uncertainties from (2), we utilize AUROC, AURAC, and ECE as hallucination detection metrics. These three metrics comprehensively evaluate the calibration degree between uncertainties and hallucinations. For hallucination mitigation, we follow a two-stage procedure: (1) We first prompt the LMM with initial context to obtain an initial answer and calculate its uncertainty using Uncertainty-o. (2) We then select the top 50% of answers with the highest uncertainty, as these are most likely to contain hallucinations. For each of these answers, we utilize the following prompt: 'Prompt: `$X`, Initial Answer: `$Y`, Your answer has a high uncertainty score of `$U`, which ranges from 0 to 1. Could you improve your answer and revise it to be more accurate? `$X`, `$Y`, and `$U` refer to the actual content of original prompt, inital answer and estimated uncertainty. Finally, we use the revised answer as the final output from the LMM. For uncertainty-aware CoT, we treat it as a multi-round thinking process. In the first round, we append the original prompt with 'Let's think step-by-step. Now, provide your first step of the answer:' We then estimate the LMM uncertainty with Uncertainty-o. From the second step onward, in addition to the original prompt, we append all previous answers and their corresponding uncertainty scores. At the end of the prompt, we add 'Respond with 'Finish.' when you think you have solved the question.' We check answer at every step for the presence of 'Finish.' and exit the loop once we find it.

C More Ablations

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

Table 11: **Ablation of Text Clustering.** During semantic uncertainty calculation, clustering based on semantics is more effective than clustering based merely on lexical presentation.

Text Clustering Manner. We report the ablation of text clustering methods (see Tab. 11). Clustering text based on its inherent semantics, rather than merely its lexical presentation, yields the best results.

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

Table 12: **Ablation of Cycle Consistency for Generation Hallucination Detection.** Uncertainty-o surpasses cycle consistency, as cycle consistency is constrained by the text similarity scoring capability of the LLM. We experiment on Flickr with StableDiffusion.

Method	Avg. Time	AUROC↑	AURAC↑	ECE↓
Confidence Elicitation	0.8s	52.3	55.8	13.1
GAVIE	1.7s	55.4	45.2	9.4
Semantic Entropy	7.1s	60.6	56.9	8.8
Uncertainty-o	3.8s	63.5	66.4	8.2

Table 13: **Time Cost Analysis of Comprehension Hallucination Detection.** We conduct evaluation on MM-Vet with InternVL2-8B.

Method	Avg. Time	AUROC↑	AURAC↑	ECE↓
GAVIE	16.2s	51.6	24.1	21.7
Semantic Entropy	124.4s	54.7	30.9	23
Uncertainty-o	69.3s	61.1	38.9	15.5

Table 14: **Time Cost Analysis of Generation Hallucination Detection.** We evaluate on MSRVTT with VideoFusion

Cycle Consistency. We also experiment with cycle consistency for detecting hallucinations in image generation settings (see Tab. 12). We observe that Uncertainty-o surpasses this baseline method, which is limited by the scoring ability of the LLM for text pair similarity.

Time Cost Analysis. We conduct time-cost analysis of Uncertainty-o along with various baseline methods (see Tab. 13 and Tab. 14). Specifically, we analyze both comprehension hallucination detection and generation hallucination detection. We follow the official best setting of Semantic Entropy, which conducts 10 rounds of sampling. We observe that for comprehension hallucination detection, although Uncertainty-o incurs higher time cost than baselines such as Confidence Elicitation and GAVIE, hallucination detection performance of our method surpasses these baselines by clear margins. Meanwhile, Uncertainty-o incurs lower time cost than Semantic Entropy while achieving higher hallucination detection accuracy. Semantic Entropy relies on vanilla temperature-based sampling, which requires excessive sampling rounds to ensure accurate uncertainty estimation. In con-

Video Comprehension		
VideoLLaMA	MSRVTTQA Acc.	MSVDQA Acc.
Direct Inference	62.9	70.5
Self Refine	64.2	70.9
Uncertainty-o wo. Score	65.8	71.4
Uncertainty-o	67.2	72.1
Audio Comprehension		
OneLLM	ClothoAQA Acc.	ClothoV2 Acc.
Direct Inference	57.1	45.5
Self Refine	57.9	47.3
Uncertainty-o wo. Score	58.7	50.8
Uncertainty-o	59.2	51.9

Table 15: **Comparison with Hallucination Mitigation Baselines.** For the hallucination mitigation task, we further compare with several self-refinement-based baselines.

trast, our proposed multimodal prompt perturbation facilitates effective mining of model inherent uncertainty, alleviating the need for excessive sampling and thereby achieving lower time cost. For generation hallucination detection, Uncertainty-o incurs higher time cost than GAVIE but achieves significantly higher hallucination detection accuracy. Uncertainty-o achieves lower time cost than Semantic Entropy due to fewer sampling rounds.

Comparison with Baseline Methods for Hallucination Mitigation. We compare our method with several self-refinement-based baselines in hallucination mitigation to illustrate the effectiveness of our method (see Tab. 15). Specifically, for Self Refine, we strictly follow its official refinement pipeline to reproduce the results. For Uncertainty-o w/o Score, when Uncertainty-o estimates high uncertainty for a given question, we prompt the large multimodal model to revise its answer without providing the estimated uncertainty score. Uncertainty-o refers to our original hallucination mitigation pipeline. We observe that our hallucination mitigation pipeline yields the best performance compared with various baselines. Uncertainty-o surpasses Self Refine since our estimated uncertainty provides an effective signal for triggering the self-refinement process, whereas Self Refine relies on the model itself to decide whether refinement is needed, which can be subjective. Moreover, during self-refinement, the estimated uncertainty can also serve as an effective quantitative guide for the revision process, while Self Refine simply prompts the model to revise its answer without any external guidance. Uncertainty-o also surpasses Uncertainty-o w/o Score, illustrating that providing the quantitative uncertainty score to the model during revision helps it revise its answer more effectively.

D Detailed Proof for Proportional Theorem

D.1 Definitions and Assumptions

1. **Sampling Process:** The large model M give predictions concurrently for perturbed prompts x_i and x_j. For those input prompts, the predictions from the large model is $y_{\text{i}} = M(x_{\text{i}})$, and $y_{\text{j}} = M(x_{\text{j}})$, respectively.

2. **Uncertainty:** We focus on the large model's *Epistemic Uncertainty*, which is the uncertainty in the large model parameters. Assume the large model parameters θ are random variables with a prior distribution $P(\theta)$. We make this assumption because large models can be naturally viewed as Bayesian Neural Networks due to the existence of the temperature mechanism: although the model parameters are fixed, given the same prompt, the model responses can vary.

3. **Prediction Difference:** Define the prediction difference $D(x)$ as:

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

where $\|\cdot\|$ denotes semantic space distance.

D.2 Mathematical Derivation

Large Model's Predictive Distribution. Assume the large model's output is a probability distribution $P(y|x,\theta)$, where y is the response, x is the prompt, and θ are the model parameters.

Posterior Predictive Distribution. According to Bayes' theorem, the posterior predictive distribution of the large model can be expressed as:

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

where $P(\theta|x)$ is the posterior distribution of the large model parameters.

Epistemic Uncertainty. The uncertainty in the parameters can be measured by the variance of the posterior distribution:

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

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

Prediction Difference and Epistemic Uncertainty. To relate the prediction difference $D(x)$ to parameter uncertainty, we need to consider the predictions from the perturbed prompts. Assume the parameters during i-th and j-th prompt perturbation are θ_{i} and θ_{j}, respectively, and they have the same prior distribution, i.e., $P(\theta_{\mathrm{i}}) = P(\theta_{\mathrm{j}})$.

Predictions from Perturbed Prompts. The predictions from perturbed prompts can be expressed as:

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

Expression for Prediction Difference. Assume the difference in predictions can be approximated by a First-order Taylor Expansion:

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

where $\nabla_\theta P(y|x,\theta)$ is the gradient of $P(y|x,\theta)$ with respect to θ.

Thus, the prediction difference $D(x)$ can be expressed as:

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

Relationship Between Prediction Difference from Prompt Perturbation and Epistemic Uncertainty. To simplify the analysis, assume θ_{i} and θ_{j} are independently and identically distributed (i.i.d.). Then:

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

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

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

Assuming $k = \mathbb{E}_{\theta|x}\left[\|\nabla_\theta P(y|x,\theta)\|^2\right]$, which is positive, we get:

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

Thus, the prediction difference $D(x)$ from prompt perturbation can be expressed as:

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

Simplifying further, we obtain:

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

D.3 Final Theorem

From the above derivation, we have shown that the prediction difference $D(x)$ from prompt perturbation is proportional to the square root of the model's epistemic uncertainty $\sqrt{\mathrm{Var}(\theta|x)}$. Therefore, the prediction difference $D(x)$ can serve as a measure of the model's epistemic uncertainty for a given sample.