CrossCheckGPT: Universal Hallucination Ranking for Multimodal Foundation Models

Guangzhi Sun[1*] **Potsawee Manakul**[1,2*] **Adian Liusie**[1] **Kunat Pipatanakul**[2]
Chao Zhang[3] **Phil Woodland**[1] **Mark Gales**[1]

[1]University of Cambridge [2]SCB 10X, SCBX Group [3]Tsinghua University

gs534@cam.ac.uk, potsawee@scb10x.com, al826@cam.ac.uk

Abstract

Multimodal foundation models are prone to hallucination, generating outputs that either contradict the input or are not grounded by factual information. Given the diversity in architectures, training data and instruction tuning techniques, there can be large variations in systems' susceptibility to hallucinations. To assess system hallucination robustness, hallucination ranking approaches have been developed for specific tasks such as image captioning, question answering, summarization, or biography generation. However, these approaches typically compare model outputs to gold-standard references or labels, limiting hallucination benchmarking for new domains. This work proposes CrossCheckGPT, a reference-free universal hallucination ranking for multimodal foundation models. The core idea of CrossCheckGPT is that the distribution of hallucination content is different among different systems, hence cross-system consistency can provide meaningful and accurate hallucination assessment scores. CrossCheckGPT can be applied to any model or task, provided that the information consistency between outputs can be measured through an appropriate distance metric. Focusing on multimodal large language models that generate text, we explore two information consistency measures: CrossCheck-explicit and CrossCheck-implicit. We showcase the applicability of our method for hallucination ranking across various modalities, namely the text, image, and audio-visual domains. Further, we propose the first audio-visual hallucination benchmark, AVHalluBench, and illustrate the effectiveness of CrossCheckGPT, achieving correlations of 98% and 89% with human judgements on MHaluBench and AVHalluBench, respectively.

1 Introduction

In generative foundation models, 'hallucination' refers to instances where generated outputs, while seemingly credible, are inconsistent with the provided context or contradict established facts [25, 49, 45]. Hallucination impacts many generative applications and can lead to misinformation [53, 34]. Given the differences in architectures, data, and alignment techniques across models, it is crucial to quantify the susceptibility of a system to hallucination, allowing the assessment of hallucination risks and choosing systems with higher factual consistency.

Current hallucination benchmarks rank systems for individual tasks including question-answering [23, 14, 19, 10, 44], summarization [30, 28], biography generation [29], instruction following [31], image captioning [37], and visual question-answering [20, 50]. Many benchmarks use proxy measures, such as answering questions designed to trigger hallucinations, but are task-specific and

[*]Equal contribution

Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models at NeurIPS 2024.

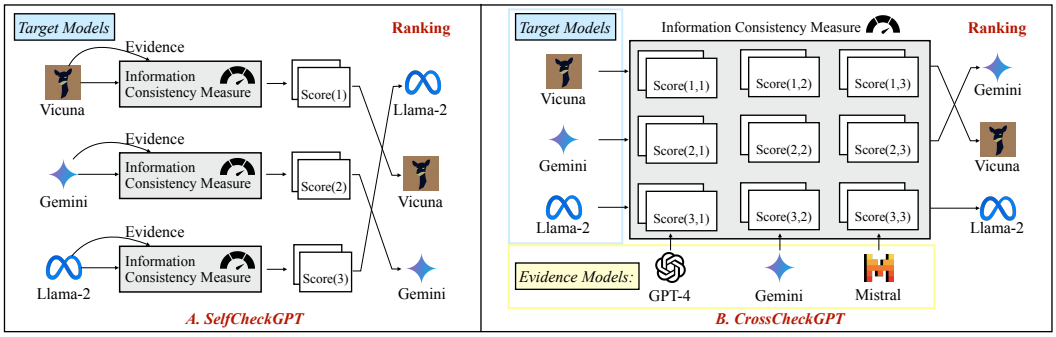


Figure 1: SelfCheckGPT (Left) and CrossCheckGPT (Right) for hallucination rankings. The approach can rank a set of MLLMs on any task without reference, enabling hallucination benchmarks for various generative tasks.

rely on gold-standard labels, limiting generalizability. On the other hand, hallucination detection approaches such as SelfCheckGPT [29] and UniHD [4] directly examine generated responses against self-evidence, without needing gold-standard answers. These methods, though, simply aim to identify when a model hallucinates, and scores are not directly comparable across different models.

In this paper, CrossCheckGPT, which is a universal hallucination ranking approach, is proposed to benchmark multimodal foundation models. The core idea of CrossCheckGPT is that the distribution of hallucinated content is different among different systems, while factual content is likely to be consistent across models. An illustration of the approach and a comparison with SelfCheckGPT is shown in Fig. 1. Instead of checking for self-consistency, as in SelfCheckGPT, CrossCheckGPT checks *cross-consistency* by comparing against evidence generated from a set of independent models. This produces more accurate and directly comparable hallucination scores, as well as yielding more robust rankings. CrossCheckGPT can be applied to any foundation model and task as long as a suitable information consistency measure is used.

CrossCheckGPT is validated on WikiBio [29] and MHaluBench [4] as text-to-text and image-to-text description tasks, and our experiments show that CrossCheckGPT achieves a notable 98% Spearman's Rank Correlation (SRC) on MHaluBench against human ranking compared to -10% SRC using SelfCheckGPT and 33% using UniHD. In addition, a comprehensive audio-visual hallucination benchmark dataset (AVHalluBench) is proposed, covering a diverse range of styles, domains and elements such as visual text, speech and music. The AVHalluBench is used to rank recent audio and video LLMs such as Gemini 1.5 Pro, conducting the first study on audio-visual hallucination benchmarking. The key contributions of this paper are summarized as follows:

- We propose CrossCheckGPT, a reference-free hallucination ranking approach that can be applied universally across text-generation tasks for systems of different modalities.
- We conduct comprehensive experiments over a range of tasks and modalities, demonstrating the effectiveness of CrossCheckGPT as a hallucination benchmarking approach for ranking text, image or audio-visual systems. Experimental results illustrate that CrossCheckGPT consistently outperforms alternate approaches, such as SelfCheckGPT [29] and UniHD [4].
- We analyze hallucination within video understanding and curate AVHalluBench, which to the best of our knowledge, is the first publicly released audio-visual hallucination benchmark.

2 Related Work

LLM Hallucination Benchmarking: Hallucination benchmarks typically rely on proxy tasks to probe the likelihood of LLM making factual errors. For example, question-answering (QA) based benchmarks, such as TriviaQA [14], TruthfulQA [23], HaluEval-QA [19], MemoTrap [31] and FEWL [51] design questions specifically to probe truthfulness and factual accuracy and rank systems by their accuracy. Other methods, such as FaithDial [10], XSum [35] and CNN-DM [39] measure hallucination in dialogue responses or summarization. However, these benchmarks require references (e.g., ground-truth answers or gold-standard references) to compare to model-generated outputs. On the other hand, SelfCheckGPT [29] can be used to rank systems on hallucination levels by measuring

systems' self-consistency scores on equivalent tasks. However, SelfCheckGPT was designed as a hallucination detection method and may not be calibrated across systems.

Multimodal LLM Hallucination Benchmarking: Multimodal hallucination has been mainly explored in the image-to-text domain for visual LLMs. One stream of methods, including CHAIR [37], LURE [57] and MHaluBench [4], directly evaluate the generated text descriptions of images using gold-standard annotations or external toolkits. Another stream of methods, such as POPE [20] and HallusionBench [13], curate a set of questions with short answers trying to capture various aspects of hallucination. Meanwhile, AMBER [50] combines both generation and question answering in one single benchmark. Unlike these methods, CrossCheckGPT does not rely on gold-standard reference or dedicated question sets, and can be universally applied to any input modalities.

3 CrossCheckGPT

CrossCheckGPT assigns a score to an MLLM (denoted as the *target* model) by assessing how much the responses of the MLLM are supported by evidence generated from a set of MLLMs (denoted as *evidence models*). The CrossCheckGPT scores can then be used to rank the MLLMs. As illustrated in Fig. 2, we explore two information consistency measures, CrossCheck-explicit and CrossCheck-implicit, which measure the hallucination of generated responses either through the explicit generation of evidence passages or implicit prompting, respectively. CrossCheckGPT is reference-free and can be generally applied to MLLMs of any input modality and output response type.

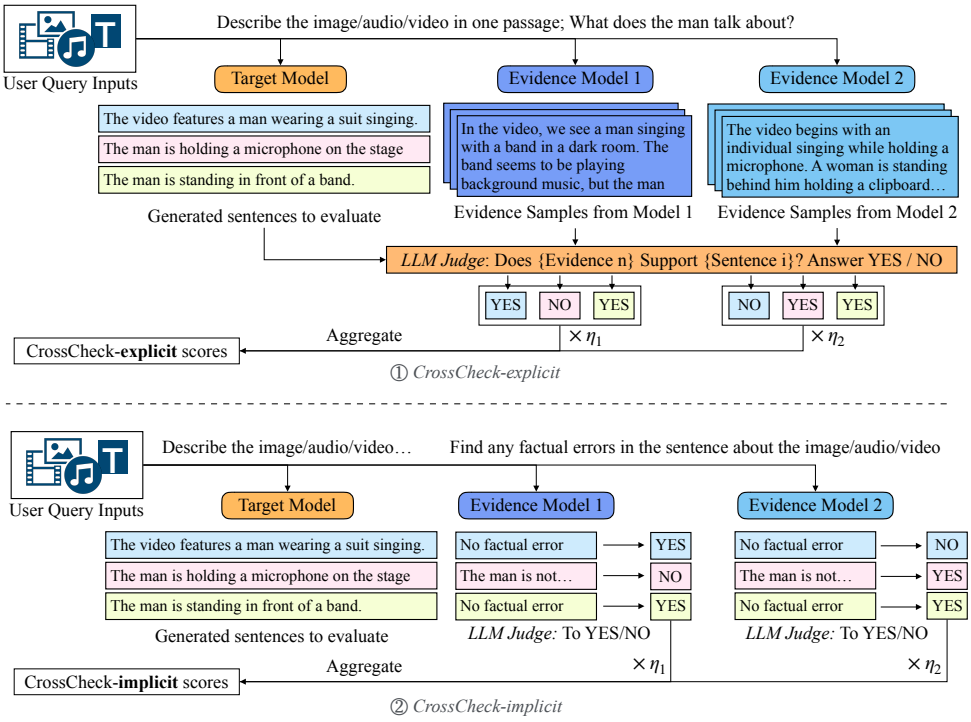


Figure 2: Illustration of the CrossCheckGPT approach with two evidence models as an example. Two information consistency measures are shown. ① CrossCheck-explicit where N passages are stochastically generated by sampling from each evidence model and ② CrossCheck-implicit where evidence models are directly used to determine whether there are any factual errors in each sentence (without sampling). The LLM judge uses the sentence and the analysis from the evidence model to produce the Yes/No binary decision.

3.1 Information Consistency Measures

CrossCheck-explicit stochastically generates a set of evidence passages from each evidence model and computes the average distance between each evidence passage and the target response. Let $R = [r_1, \ldots, r_i, \ldots, r_I]$ denote the response of the target model $\hat{M}$, where r_i is the i-th sentence

of the response, to a given query Q, which can be of any modality. We first re-formulate the SelfCheckGPT score for sentence r_i of the target model in Eqn. (1) below,

$$\mathcal{S}_{\text{selfcheck}}(\hat{M}) = \frac{1}{|\mathcal{Q}|}\frac{1}{I}\sum_{Q\in|\mathcal{Q}|}\sum_{i=1}^{I}\mathcal{S}_{r_i,Q}^{\text{selfcheck}}(\hat{M}) \qquad \text{where} \quad \mathcal{S}_{r_i,Q}^{\text{selfcheck}}(\hat{M}) = \frac{1}{\hat{N}}\sum_{n=1}^{\hat{N}} x_{r_i,Q}^{(n)}(\hat{M}) \quad (1)$$

where $\mathcal{Q}$ is the set of queries in a test set, $\hat{N}$ is the number of stochastically generated passages by the model $\hat{M}$, and $x_{r_i,Q}^{(n)}(\hat{M})$ denotes the hallucination score of whether sentence r_i is supported by evidence n from $\hat{M}$. The hallucination score, estimated by prompting an LLM judge with the sentence and each evidence, takes a value in $\{0, 1\}$, where 0 denotes *supported* and 1 denotes *hallucinatory*.

CrossCheck-explicit, in contrast to SelfCheckGPT, uses the evidence from $|\mathcal{M}|$ evidence models and measures the distance of the response against those from all other systems. The overall CrossCheck-explicit score $\mathcal{C}_{\text{explicit}}(\hat{M})$ for a specific target model $\hat{M}$ can be computed using Eqn. (2),

$$\mathcal{C}_{\text{explicit}}(\hat{M}) = \frac{1}{|\mathcal{Q}|}\frac{1}{I}\sum_{Q\in|\mathcal{Q}|}\sum_{i=1}^{I}\mathcal{C}_{r_i,Q}^{\text{explicit}}(\hat{M}) \quad \text{where} \quad \mathcal{C}_{r_i,Q}^{\text{explicit}}(\hat{M}) = \frac{\sum_{j=1}^{|\mathcal{M}|}\eta_j\sum_{n=1}^{N_j}x_{r_i,Q}^{(n)}(M_j)}{\sum_{j=1}^{|\mathcal{M}|}\eta_j N_j} \quad (2)$$

where $\mathcal{M}$ denotes the set of evidence models used for CrossCheck-explicit. Note that self-consistency can be taken into account by including the target model $\hat{M}$ into the evidence models, $\hat{M}\in\mathcal{M}$. Each evidence model M_j stochastically generates N_j passages to check the response against, and since systems may have different levels of reliability, a factor η_j can be assigned to the passages generated from model M_j.

CrossCheck-implicit is an alternative consistency measure, where instead of explicitly generating passages for the same query, the evidence models are prompted to spot any factual errors in each sentence. The overall implicit CrossCheck-implicit score is computed using Eqn. (3),

$$\mathcal{C}_{\text{implicit}}(\hat{M}) = \frac{1}{|\mathcal{Q}|}\frac{1}{I}\sum_{Q\in|\mathcal{Q}|}\sum_{i=1}^{I}\mathcal{C}_{r_i,Q}^{\text{implicit}}(\hat{M}) \qquad \text{where} \quad \mathcal{C}_{r_i,Q}^{\text{implicit}}(\hat{M}) = \sum_{j=1}^{|\mathcal{M}|}\eta_j\, y_{r_i,Q}(M_j) \quad (3)$$

where $y_{r_i,Q}(M_j)$ denotes the hallucination score of sentence r_i computed using CrossCheck-implicit. In contrast to CrossCheck-explicit (which computes $x_{r_i,Q}(M_j)$), $y_{r_i,Q}(M_j)$ is computed by first prompting the evidence model M_j to analyze whether r_i contains any factual errors given the input Q. The LLM judge then takes the input r_i and analysis from model M_j and predicts $y_{r_i,Q}(M_j)$, whether the response is hallucinatory. If factual errors are found in r_i, $y_{r_i,Q}(M_j) = 1$, and otherwise $y_{r_i,Q}(M_j) = 0$. We note that concurrent work, PoLL [48], applies a group of models as judges to evaluate texts and can be viewed as similar to CrossCheck-implicit. This work focuses on multimodal inputs and hallucination benchmarking.

3.2 Confidence-based Weighting for Evidence Models

While all evidence models are advanced MLLMs, the quality of their evidence may vary depending on their propensity to hallucinate. Therefore, a weighting mechanism is proposed where the scores are weighted by model uncertainty reflected by SelfCheckGPT scores, as shown below:

$$\eta_j = \frac{e^{-\mathcal{S}_{\text{selfcheck}}(M_j)/T}}{\sum_{k=1}^{|\mathcal{M}|}e^{-\mathcal{S}_{\text{selfcheck}}(M_k)/T}}, \quad (4)$$

where T is the calibration temperature that determines the sharpness of the weight distribution, which is set to a constant for each benchmark. A higher SelfCheckGPT score indicates that the model tends to generate inconsistent information and is more uncertain. In addition, this weighting mechanism ensures that outlier systems will not be undermined by the evidence from weaker models.[1]

[1]Note that a weight distribution can also be associated with each specific query by using the average SelfCheckGPT score of each evidence model.

4 CrossCheckGPT for Hallucination with Multimodal Inputs

CrossCheckGPT is designed to be general and applicable to models of any input modality, provided that the outputs are of a consistent form (i.e. text) and a suitable information consistency measure is used. This general design of CrossCheckGPT enables it to also be applied to rank multi-modal systems (i.e. systems which use two or more input modalities).

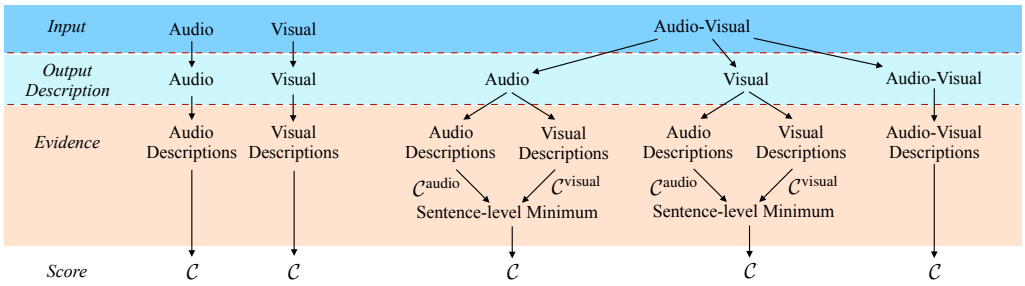


Figure 3: CrossCheckGPT score computation for AVHalluBench with audio, visual and audio-visual inputs.

As shown in Fig. 3, CrossCheckGPT is used to evaluate models across three categories: *audio*, *visual* (e.g., image/silent video), and *audio-visual*. For audio-visual, we conduct the first study on hallucination evaluation using videos with paired audio. Due to limited systems that process audio-visual inputs, multi-modal models are prompted to split outputs into visual and auditory descriptions, evaluating them separately. Visual descriptions are used to check visual-only inputs and audio descriptions to check audio-only inputs. Information may require both modalities to check hallucination in audio-visual settings, e.g., demonstrating a skateboard trick. In this scenario, $\mathcal{C} = \min\left(\mathcal{C}^{\text{audio}}, \mathcal{C}^{\text{visual}}\right)$ is the score where $\mathcal{C}^{\text{audio}}$ and $\mathcal{C}^{\text{visual}}$ are the audio and visual descriptions.[2]

AVHalluBench: To benchmark hallucinations in audio-visual LLMs, we curate AVHalluBench, a dataset containing 175 videos selected from six video understanding datasets covering various styles and elements, with statistics shown in Table 15 in the Appendix. To verify the effectiveness of CrossCheckGPT (and future benchmarking methods), AVHalluBench includes a carefully written set of hallucination-free descriptions for audio and visual contents. After watching each video with audio, the annotators were instructed to write *one* description focusing on the audio content and *one* description focusing on the visual content of the video, *separately*.[3] To analyze the inter-annotator agreement, we split each description into atomic facts [32] and verify each fact against the descriptions written by the other annotators, categorized as either: *Supporting*, such that the fact is supported by the other annotator, *Contradicting*, such that the fact contradicts the information provided by the other annotator, or *Neutral* such that the facts neither support nor contradict one another. Both decomposition and verification processes are performed automatically using GPT-4. Of the 39 videos annotated by multiple annotators, there were 471 audio-related facts and 913 visual-related facts, and the agreement between annotators (as counted by Supporting/Neutral/Contradicting) was 64.6%/24.6%/10.8% and 62.0%/29.0%/9.0%, respectively. AVHalluBench is available at https://huggingface.co/datasets/scb10x/avhallubench.

5 Experiments

We conduct experiments to validate CrossCheckGPT on MLLMs with three input modalities, including text (§5.1), image (§5.2), and audio-visual (§5.3). During inference, we use a temperature of 1.0, a beam size of 1 and a top-p of 0.9 are used for all models. *SelfCheckGPT* [29] is applied as a hallucination ranking baseline for all modalities since it is reference-free and not task-specific.

[2]For simplicity, $\hat{M}$, r_i, and Q are dropped here, and the scores can be either implicit or explicit. Initial findings showed that CrossCheck-implicit produces different audio and visual score ranges, averaging 0.2 and 0.5, respectively. Thus, only CrossCheck-explicit is used for audio-visual inputs.

[3]To maximize coverage, initial descriptions were generated using Gemini 1.5 Pro and GPT-4v, prompted to describe all the elements present in the sequence of frames. Note that although these descriptions are *not* hallucination-free, they have a high level of coverage and subjective details. The annotators were provided with these descriptions in addition to the videos while being instructed to write only objective details of the videos.

5.1 Text-to-text Experiments

Experimental Setup: The main text-to-text experiments are performed using the subset of WikiBio data used in [29], which contains 238 biographical passages from Wikipedia. We select 10 open-source LLMs (listed in Appendix Table 7) as target models, 8 of which are used as evidence models. Four models are Llama-2-7B based [46] (e.g. Vicuna-v1.5-7B [6]) and four models are Mistral-7B based [16]. Each evidence model generates 20 stochastic passages. For the LLM judge in CrossCheck-explicit (used to determine whether sentences support one another), Mistral-7B [16] is used as it achieves the best results among all considered open-source LLMs (shown in Appendix Table 10).

To evaluate the *general* benchmarking ability of ranking methods, 10 benchmark metrics from the hallucinations leaderboard [15] (shown in Table 8) are selected to provide the overall hallucination ranking of the systems. These metrics are either based on human annotation or gold-standard references, where the overall rankings are obtained by averaging the rankings from each metric.

We report the *system-level* correlation between the hallucination ranking methods and the overall ranking measured by Spearman's Rank Correlation coefficient (SRC), denoted as System(ρ). In addition, as WikiBio contains reference texts, the references can be used as evidence texts, which can be considered an idealized fact-checking method. This method is referred to as RefCheck, and CrossCheckGPT and SelfCheckGPT scores also are compared against RefCheck at *document-level* using Pearson's Correlation Coefficient (PCC), denoted as Document(r). Furthermore, to investigate the effectiveness of CrossCheckGPT when the target LLM is much more powerful than those evidence models, we include GPT-4 in addition to the 10 target LLMs.

Hallucination Ranking Results: Existing hallucination metrics such as HaluEval-QA accuracy do not correlate well with the overall ranking at the system level. Some metrics have negative correlations while the highest (TruthfulQA MC2) is 57.14% (shown in Table 1, with further pairwise correlations provided in Appendix Table 13). This is likely because each existing metric is typically designed to measure only one aspect related to hallucinations, e.g., probing through question-answering.

Metrics	System(ρ) (%)	Document (r) (%) w/o GPT4	Document (r) (%) with GPT4
TruthfulQA MC2 [23]	57.14	-	-
SelfCheckGPT [29]	66.46	74.06	76.08
CrossCheck-implicit	56.71	18.33	17.29
CrossCheck-explicit	<u>77.44</u>	**82.28**	<u>77.23</u>
CrossCheck-implicit weighted	56.81	20.21	19.16
CrossCheck-explicit weighted	**82.32**	<u>81.78</u>	**82.18**

Table 1: General hallucination evaluation where the task for SelfCheckGPT/CrossCheckGPT is open-ended biography generation on WikiBio. System-level correlation, System(ρ), is measured against the overall ranking of the leaderboard, and document-level correlation, Document(r), is measured against RefCheck. "With GPT-4" refers to including GPT-4 as a target model. Additional metrics are presented in Table 11 in the Appendix.

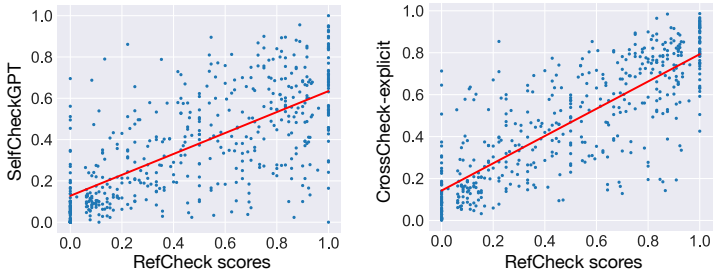


Figure 4: Scatter plot of document-level scores for SelfCheckGPT and CrossCheck-explicit against RefCheck for text-to-text experiments.

Subset	Values
Succ. Rate	90%
P-value	4×10^{-6}

Table 2: Success rate of CrossCheck outperforming SelfCheck for independent subsets of WikiBio documents. The P-value is measured by the one-tailed sign test with H_0 = CrossCheck not better than SelfCheck.

CrossCheck-explicit correlates with the overall ranking better than all other methods, with CrossCheck-explicit weighted by model uncertainty achieving the highest correlation, highlighting its effective *general* hallucination ranking ability. In addition, the document-level correlation plots are

shown in Fig. 4, and the sign test on independent subsets in Table 2 shows the statistical significance ($p = 4 \times 10^{-}6$) of CrossCheckGPT being better than SelfCheckGPT for ranking at the system-level.

5.2 Image-to-text Experiments

We validate CrossCheckGPT for the hallucination ranking of visual LLMs on image-to-text tasks. The experiments are performed on MHaluBench [4], an image-captioning hallucination dataset. Nine visual LLMs are selected as target models, all of which are used to generate evidence passages (see Appendix Table 7 for the list of models). Each evidence model generates ten image descriptions per image. The overall ranking is obtained by averaging the rankings from CHAIR [37] and POPE (MSCOCO subset) [20].[4] In addition to SelfCheckGPT, UniHD[4] is used as a stronger baseline.

For evaluation, we take a subset of 30 image descriptions generated by each target model (a total of 270 passages with 3237 facts) and annotate each description with a binary label of either *hallucinatory* or *factual*. The Cohen's κ between the two annotators is 0.632, indicating substantial agreement. The models are ranked by the average percentage of factual errors produced by each target model, and hallucination ranking performance is measured at the *system-level* using SRC, denoted System(ρ) and at the *image-level* using PCC, denoted as Image(r).

Metrics	System(ρ) (%)			Image(r) (%)
	Overall	CHAIR	Human	Human
UniHD [4]	42.02	36.98	33.33	36.70
SelfCheckGPT [29]	43.70	23.10	-10.00	20.93
CrossCheck-implicit	**50.42**	**64.71**	**98.33**	48.72
CrossCheck-explicit	42.86	43.70	75.00	35.16
CrossCheck-implicit weighted	**50.42**	**64.71**	**98.33**	**52.83**
CrossCheck-explicit weighted	47.06	46.22	73.33	36.98

Table 3: System-level correlation measured by System(ρ) and Image-level correlation measured by Image(r) for various hallucination evaluation methods on the MHaluBench dataset. System-level correlation is measured against the overall ranking, rankings from CHAIR scores and human annotation.

Hallucination Ranking Results: Similar to before, Table 3 presents the system-level and image-level correlations against overall rankings and rankings derived from human annotations. Both variants of CrossCheckGPT outperform SelfCheckGPT and UniHD, with CrossCheck-implicit weighted performing best out of all methods, achieving a 98.33% correlation with the rankings from human annotations. Equivalent statistical significance analysis and scatter plots are shown in Table 14 and Fig. 7 in the Appendix F, respectively.

5.3 Video-to-text Experiments

Next, we apply CrossCheckGPT to AVHalluBench to investigate hallucination ranking in audio-visual LLMs. We consider 7 models that can handle video inputs and 6 models that can handle audio inputs. Three models, FAVOR [41], Video-LLaMA [55], and Gemini 1.5 Pro [43], are in the intersection of the two sets, and can handle audio-visual inputs. When ranking hallucinations for visual description, we consider audio-visual LLMs with *visual-only* inputs and *audio-visual* inputs as separate systems, and hence, there are $7+3=10$ target models for ranking. We conduct a similar ranking scheme for audio descriptions, where there are $6+3=9$ target models. All the target models are also used as evidence models in CrossCheck-explicit,[5] and each model generates ten evidence passages. When using audio-visual LLMs as evidence models, audio-visual inputs are given to obtain the visual or audio descriptions as evidence. As only 5 target models can handle speech inputs, we further make a dedicated ranking only for these models with prompts explicitly asking for speech description.

Hallucination Ranking Results: First, system-level and video-level correlations are shown in Table 4, measured by System(ρ) and Video(r). CrossCheck-explicit correlates with RefCheck best, with an 89.09% System(ρ) for the visual description. Similar to the text-to-text results, we observe that

[4]CHAIR and POPE are the two popular representative metrics for free-form text generation and binary classification hallucination benchmarks respectively [50].

[5]Gemini 1.5 Pro is not used for CrossCheck-implicit due to the number of request limitations.

Metrics	Visual Description (%)		Audio Description (%)	
	System(ρ)	Video(r)	System(ρ)	Video(r) (w. speech)
SelfCheckGPT	86.67	65.77	60.00	51.13 (44.55)
CrossCheck-implicit weighted	54.29	30.73	40.00	2.15 (16.20)
CrossCheck-explicit weighted	**89.09**	**78.58**	**71.67**	**68.10** (**47.60**)

Table 4: System-level and video-level correlations of SelfCheckGPT and CrossCheckGPT against RefCheck using manual descriptions in AVHalluBench. Weighted version of CrossCheckGPT is used with $C = 0.1$. Ranking correlations for systems that handle speech are in brackets.

CrossCheck-explicit performs better than CrossCheck-implicit. For both text-to-text and video-to-text experiments, this is likely due to the high diversity in the evidence passages as indicated by high raw SelfCheckGPT scores, which we discuss further in Section 5.4.

Impact of Audio-Visual Inputs: As supporting information from another modality is expected to reduce hallucination, this section investigates whether audio-visual inputs reduce the raw hallucination scores compared to the scores when a single modality is used. Table 5 presents the average raw hallucination scores (rather than correlations), for three MLLMs that can take audio-visual inputs.

Model	Input modality	Visual Description (%)		Audio Description (%)	
		$\mathcal{S}_{\text{selfcheck}} \downarrow$	$\mathcal{C}_{\text{explicit}} \downarrow$	$\mathcal{S}_{\text{selfcheck}} \downarrow$	$\mathcal{C}_{\text{explicit}} \downarrow$
FAVOR [41]	Visual	60.67	53.85	—	—
	Audio	—	—	49.62	66.69
	Audio-Visual	56.42	49.60	33.25	35.20
Video-LLaMA [55]	Visual	41.14	52.02	—	—
	Audio	—	—	56.42	68.05
	Audio-Visual	47.73	49.13	70.23	41.25
Gemini 1.5 Pro [43]	Visual	19.87	31.74	—	—
	Audio	—	—	**25.82**	34.66
	Audio-Visual	**12.77**	**23.27**	48.51	**28.79**

Table 5: SelfCheckGPT scores ($\mathcal{S}_{\text{selfcheck}}$) and weighted CrossCheck-explicit scores ($\mathcal{C}_{\text{explicit}}$) on AVHalluBench for audio-visual LLMs. Calibration temperature $T = 0.1$ is used here.

When considering the CrossCheckGPT scores, we observe that having audio-visual inputs reduces hallucination rates, as measured by the raw CrossCheckGPT scores, as expected. While Gemini 1.5 Pro achieved the best scores, it can be more susceptible to hallucination when silent videos are used as inputs as it often fabricates its audio descriptions. Moreover, except for Gemini 1.5 Pro, when audio-visual inputs are used the reduction in hallucination scores is larger for audio description tasks than for visual description tasks. This likely occurs as for audio description tasks, visual information often provides useful information on the source of the sound, which can significantly reduce the uncertainty of the sound. For visual description tasks, while particular audio cues (especially from speech) can provide useful information, misleading or unrelated sounds may cause additional hallucinations. For example, in Fig 10 where there is a self-playing piano, audio inputs can mislead a model to believe that the piano is played by an individual. Further examples are presented in Appendix H with the raw hallucination scores for audio and visual-only inputs shown in Tables 16 and 17 in Appendix.

5.4 CrossCheck-explicit vs. CrossCheck-implicit

While CrossCheck-implicit is more sample-efficient than CrossCheck-explicit and only requires generating the error analysis once, the performance of CrossCheck-implicit can be highly dependent on the task. For the text-to-text and video-to-text experiments, CrossCheck-implicit performs worse than CrossCheck-explicit, as opposed to the findings in the image-to-text experiments. We hypothesize that for challenging and open-ended tasks, CrossCheck-explicit is preferred as it can better cover the output space by disentangling the evidence generation and verification tasks, yielding more calibrated uncertainty measures. However, in other circumstances, CrossCheck-implicit may help the model focus on specific aspects of the input and yield more accurate rankings. For challenging

and open-ended tasks with diverse outputs, the raw SelfCheckGPT scores are expected to be high and therefore can be used as a proxy to determine which consistency measure to select. For example, the average SelfCheckGPT score across models is 40.63% for text-to-text, which is much higher than 17.16% for image-to-text. We recommend using CrossCheck-explicit when the SelfCheckGPT scores are high, and CrossCheck-implicit when they are sufficiently low, which is demonstrated to be a reasonable rule, illustrated by the results in Appendix Table 18.

5.5 Ablation Studies

Self-Bias: LLMs are known to have self-preferential bias [2, 56] and may prefer outputs from similar models. Therefore LLMs using the same base model may provide inflated CrossCheckGPT scores. The results in Table 6 show that self-bias is an issue, and for example, when only using Llama-2-based evidence models, the outputs from Vicuna get a lower hallucination score whereas when only using Mistral-based evidence models, Mistral has the lowest hallucination score, resulting in contradictory conclusions. This bias can be mitigated by adopting a wide range of evidence models, which is adopted in CrossCheckGPT scores, hence achieving more reliable evaluation with strong correlations.

Evidence Models	System(ρ)	Document(r)	Vicuna $\mathcal{C}_{\text{explicit}}$	Mistral $\mathcal{C}_{\text{explicit}}$
Llama-2-based models only	55.49%	81.10%	42.94%	45.68%
Mistral-based models only	81.71%	81.06%	44.98%	41.81%
All models	82.32%	82.28%	44.82%	44.93%

Table 6: The mitigation of self-bias in CrossCheckGPT scores and its influence measured by document-level correlations and CrossCheck-explicit scores of Vicuna and Mistral on WikiBio. There are 4 Llama-2-based models and 4 Mistral-based models in the set of evidence models.

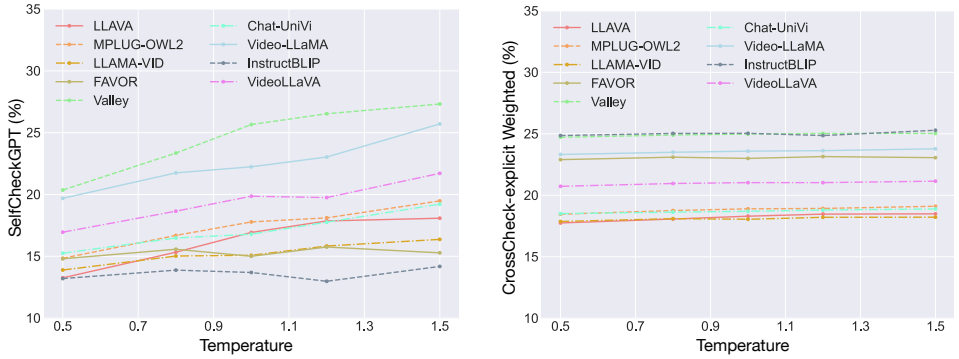


Figure 5: Variation of SelfCheckGPT scores (Left) and the weighted CrossCheck-explicit scores (Right) against the varying temperature during description generation.

Robustness to Manipulation: To investigate whether a ranking method can be easily manipulated, we examine the influence of the generation temperature (which can be selected for any model). The results in Fig. 5 show that by increasing the temperature of the target model from 0.5 to 1.5, SelfCheckGPT scores increase by as much as 35%, drastically influencing the rankings. In contrast, CrossCheckGPT provides more stable rankings for all generation temperatures. Results are demonstrated for MHaluBench, but similar trends are observed for WikiBio as well.

6 Conclusions

This paper proposes CrossCheckGPT, a universal hallucination ranking method for multimodal large language models. We evaluated two variants of CrossCheckGPT on text-to-text, image-to-text and video-to-text tasks, demonstrating that it consistently outperforms all baseline methods, achieving 98% and 89% system-level correlation against humans on MHaluBench and AVHalluBench respectively. We also introduce AVHalluBench, the first resource to study audio-visual hallucination issues in video understanding.

Acknowledgments

This work is supported by Cambridge University Press & Assessment (CUP&A), a department of The Chancellor, Masters, and Scholars of the University of Cambridge.

References

[1] E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah, Étienne Goffinet, D. Hesslow, J. Launay, Q. Malartic, D. Mazzotta, B. Noune, B. Pannier, and G. Penedo. The falcon series of open language models. *arXiv:2311.16867*, 2023.

[2] J. D. Brown. Evaluations of self and others: Self-enhancement biases in social judgments. *Social cognition*, 4(4):353–376, 1986.

[3] S. Chen, X. He, L. Guo, X. Zhu, W. Wang, J. Tang, and J. Liu. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv:2304.08345*, 2023.

[4] X. Chen, C. Wang, Y. Xue, N. Zhang, X. Yang, Q. Li, Y. Shen, L. Liang, J. Gu, and H. Chen. Unified hallucination detection for multimodal large language models. *arXiv:2402.03190*, 2024.

[5] Z. Chen, H. Liu, W. Yu, G. Sun, H. Liu, J. Wu, C. Zhang, Y. Wang, and Y. Wang. M^3av: A multimodal, multigenre, and multipurpose audio-visual academic lecture dataset. *arXiv:2403.14168*, 2024.

[6] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing. Vicuna: An opensource chatbot impressing gpt-4 with 90% chatgpt quality., 2023.

[7] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.

[8] W. Dai, J. Li, D. Li, A. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=vvoWPYqZJA.

[9] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=r1l73iRqKm.

[10] N. Dziri, E. Kamalloo, S. Milton, O. Zaiane, M. Yu, E. Ponti, and S. Reddy. Faithdial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics*, 10: 1473–1490, 2022.

[11] S. Feng, V. Balachandran, Y. Bai, and Y. Tsvetkov. FactKB: Generalizable factuality evaluation using language models enhanced with factual knowledge. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 933–952, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.59. URL https://aclanthology.org/2023.emnlp-main.59.

[12] Y. Gong, H. Luo, A. H. Liu, L. Karlinsky, and J. R. Glass. Listen, think, and understand. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=nBZBPXdJlC.

[13] T. Guan, F. Liu, X. Wu, R. Xian, Z. Li, X. Liu, X. Wang, L. Chen, F. Huang, Y. Yacoob, D. Manocha, and T. Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, 2024.

[14] M. Han, M. Kang, H. Jung, and S. J. Hwang. Episodic memory reader: Learning what to remember for question answering from streaming data. In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4407–4417, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1434. URL https://aclanthology.org/P19-1434.

[15] G. Hong, A. P. Gema, R. Saxena, X. Du, P. Nie, Y. Zhao, L. Perez-Beltrachini, M. Ryabinin, X. He, and P. Minervini. The hallucinations leaderboard–an open effort to measure hallucinations in large language models. *arXiv preprint arXiv:2404.05904*, 2024.

[16] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7b. *arXiv:2310.06825*, 2023.

[17] P. Jin, R. Takanobu, C. Zhang, X. Cao, and L. Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *CVPR*, 2024.

[18] G. li, Y. Wei, Y. Tian, C. Xu, J.-R. Wen, and D. Hu. Learning to answer questions in dynamic audio-visual scenarios. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[19] J. Li, X. Cheng, X. Zhao, J.-Y. Nie, and J.-R. Wen. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.397. URL https://aclanthology.org/2023.emnlp-main.397.

[20] Y. Li, Y. Du, K. Zhou, J. Wang, X. Zhao, and J.-R. Wen. Evaluating object hallucination in large vision-language models. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20.

[21] Y. Li, C. Wang, and J. Jia. Llama-vid: An image is worth 2 tokens in large language models. *arXiv:2311.17043*, 2023.

[22] B. Lin, B. Zhu, Y. Ye, M. Ning, P. Jin, and L. Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv:2311.10122*, 2023.

[23] S. Lin, J. Hilton, and O. Evans. TruthfulQA: Measuring how models mimic human falsehoods. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229.

[24] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=w0H2xGHlkw.

[25] H. Liu, W. Xue, Y. Chen, D. Chen, X. Zhao, K. Wang, L. Hou, R. Li, and W. Peng. A survey on hallucination in large vision-language models. *arXiv:2402.00253*, 2024.

[26] R. Luo, Z. Zhao, M. Yang, J. Dong, M. Qiu, P. Lu, T. Wang, and Z. Wei. Valley: Video assistant with large language model enhanced ability. *arXiv:2306.07207*, 2023.

[27] D. Mahan, R. Carlow, L. Castricato, N. Cooper, and C. Laforte. Stable beluga models. URL https://huggingface.co/stabilityai/StableBeluga2.

[28] P. Manakul, A. Liusie, and M. Gales. MQAG: Multiple-choice question answering and generation for assessing information consistency in summarization. In J. C. Park, Y. Arase, B. Hu, W. Lu, D. Wijaya, A. Purwarianti, and A. A. Krisnadhi, editors, *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 39–53, Nusa Dua, Bali, Nov. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.ijcnlp-main.4. URL https://aclanthology.org/2023.ijcnlp-main.4.

[29] P. Manakul, A. Liusie, and M. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557.

[30] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.173. URL https://aclanthology.org/2020.acl-main.173.

[31] I. R. McKenzie, A. Lyzhov, M. Pieler, A. Parrish, A. Mueller, A. Prabhu, E. McLean, A. Kirtland, A. Ross, A. Liu, A. Gritsevskiy, D. Wurgaft, D. Kauffman, G. Recchia, J. Liu, J. Cavanagh, M. Weiss, S. Huang, T. F. Droid, T. Tseng, T. Korbak, X. Shen, Y. Zhang, Z. Zhou, N. Kim, S. R. Bowman, and E. Perez. Inverse scaling: When bigger isn't better. *TMLR*, 2023.

[32] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741.

[33] S. Mukherjee, A. Mitra, G. Jawahar, S. Agarwal, H. Palangi, and A. Awadallah. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv:2306.02707*, 2023.

[34] M. Nahar, H. Seo, E.-J. Lee, A. Xiong, and D. Lee. Fakes of varying shades: How warning affects human perception and engagement regarding llm hallucinations. *arXiv:2404.03745*, 2024.

[35] S. Narayan, S. B. Cohen, and M. Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1206. URL https://aclanthology.org/D18-1206.

[36] OpenAI. GPT-4 technical report. *arXiv:2303.08774*, 2023.

[37] A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, and K. Saenko. Object hallucination in image captioning. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1437. URL https://aclanthology.org/D18-1437.

[38] R. Sanabria, O. Caglayan, S. Palaskar, D. Elliott, L. Barrault, L. Specia, and F. Metze. How2: A large-scale dataset for multimodal language understanding. In *Proc. ViGIL*, 2018.

[39] A. See, P. J. Liu, and C. D. Manning. Get to the point: Summarization with pointer-generator networks. In R. Barzilay and M.-Y. Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1099. URL https://aclanthology.org/P17-1099.

[40] X. Shen, D. Li, J. Zhou, Z. Qin, B. He, X. Han, A. Li, Y. Dai, L. Kong, M. Wang, Y. Qiao, and Y. Zhong. Favdbench: Fine-grained audible video description. In *Proc. CVPR*, 2023.

[41] G. Sun, W. Yu, C. Tang, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang. Fine-grained audio-visual joint representations for multimodal large language models. *arXiv:2310.05863*, 2023.

[42] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. MA, and C. Zhang. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=14rn7HpKVk.

[43] G. Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv:2403.05530*, 2024.

[44] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074.

[45] S. M. T. I. Tonmoy, S. M. M. Zaman, V. Jain, A. Rani, V. Rawte, A. Chadha, and A. Das. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv:2401.01313*, 2024.

[46] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams,

J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*, 2023.

[47] L. Tunstall, E. Beeching, N. Lambert, N. Rajani, K. Rasul, Y. Belkada, S. Huang, L. von Werra, C. Fourrier, N. Habib, N. Sarrazin, O. Sanseviero, A. M. Rush, and T. Wolf. Zephyr: Direct distillation of lm alignment. *arXiv:2310.16944*, 2023.

[48] P. Verga, S. Hofstatter, S. Althammer, Y. Su, A. Piktus, A. Arkhangorodsky, M. Xu, N. White, and P. Lewis. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*, 2024.

[49] C. Wang, X. Liu, Y. Yue, X. Tang, T. Zhang, C. Jiayang, Y. Yao, W. Gao, X. Hu, Z. Qi, Y. Wang, L. Yang, J. Wang, X. Xie, Z. Zhang, and Y. Zhang. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv:2310.07521*, 2023.

[50] J. Wang, Y. Wang, G. Xu, J. Zhang, Y. Gu, H. Jia, M. Yan, J. Zhang, and J. Sang. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

[51] J. Wei, Y. Yao, J.-F. Ton, H. Guo, A. Estornell, and Y. Liu. Measuring and reducing llm hallucination without gold-standard answers via expertise-weighting. *arXiv:2402.10412*, 2024.

[52] J. Xiao, X. Shang, A. Yao, and T.-S. Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *Proc. CVPR*, 2021.

[53] X. Yang, L. Pan, X. Zhao, H. Chen, L. Petzold, W. Y. Wang, and W. Cheng. A survey on detection of llms-generated content. *arXiv:2310.15654*, 2023.

[54] Q. Ye, H. Xu, J. Ye, M. Yan, A. Hu, H. Liu, Q. Qian, J. Zhang, F. Huang, and J. Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv:2311.04257*, 2023.

[55] H. Zhang, X. Li, and L. Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In Y. Feng and E. Lefever, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-demo.49. URL https://aclanthology.org/2023.emnlp-demo.49.

[56] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

[57] Y. Zhou, C. Cui, J. Yoon, L. Zhang, Z. Deng, C. Finn, M. Bansal, and H. Yao. Analyzing and mitigating object hallucination in large vision-language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=oZDJKTlOUe.

[58] B. Zhu, E. Frick, T. Wu, H. Zhu, and J. Jiao. Starling-7b: Improving llm helpfulness & harmlessness with rlaif, November 2023.

A Experimental Setup Details

We list the models involved in this paper in Table 7, and text-to-text metrics in Table 8.

Target LLMs	Modality	Evidence Models (explicit)	Evidence Models (explicit)	License
Llama-2-7B [29]	Text	✓	✓	llama2
Llama-2-7B-Chat [29]	Text	✓	✓	llama2
Mistral-7B-Instruct-v0.1 [16]	Text	✗	✗	Apache-2.0
Mistral-7B-Instruct-v0.2 [16]	Text	✓	✓	Apache-2.0
Vicuna-v1.5-7B[6]	Text	✓	✓	llama2
Falcon-7B[1]	Text	✗	✗	Apache-2.0
Starling-7B-alpha[58]	Text	✓	✓	Apache-2.0
StableBeluga-7B[27]	Text	✓	✓	llama2
Zephyr-7b-beta[47]	Text	✓	✓	MIT
Mistral-7B-OpenOrca[33]	Text	✓	✓	Apache-2.0
GPT-4 [36]	Text	✗	✗	N/A
LLaVA-v1.5 [24]	Vision	✓	✓	llama2
InstructBLIP (vicuna-7B) [8]	Vision	✓	✗	BSD 3-Clause
mPLUG-Owl2 [54]	Vision	✓	✓	MIT
Valley [26]	Vision	✓	✓	Apache-2.0
Video-LLaVA [22]	Vision	✓	✓	Apache-2.0
Chat-Univi [17]	Vision	✓	✓	Apache-2.0
LLaMA-VID [21]	Vision	✓	✗	Apache-2.0
LTU [12]	Audio	✓	✓	Apache-2.0
Qwen-Audio-Chat [7]	Audio	✓	✓	Tongyi Qianwen
SALMONN [42]	Audio	✓	✓	Apache-2.0
Video-LLaMA [55]	Audio-visual	✓	✓	BSD 3-Clause
FAVOR [41]	Audio-visual	✓	✓	Apache-2.0
Gemini 1.5 Pro [43]	Audio-visual	✓	✗	N/A

Table 7: Models and reference benchmarks for validating CrossCheckGPT.

Reference Benchmarks (Metrics)	Description
TriviaQA [14] (Acc)	A realistic text-based question-answering dataset containing documents collected from Wikipedia and the web.
TruthfulQA MC1 [23] (Acc) TruthfulQA MC2 [23] (Acc)	A benchmark to measure whether a language model is truthful in generating answers to questions, spanning 38 categories.
XSum [35] (FactKB [11])	The factual accuracy of summarization models by verifying the presence of knowledge base facts in generated summaries.
CNN-DM [39] (BERTP)	The CNN-DailyMail dataset is a collection of news articles and accompanying summaries measured by BERTScore-Precision.
MemoTrap [31] (Acc)	Assessing whether LLMs fall into memorization traps which occur when LLMs memorize specific examples in training.
FaithDial [10] (Acc)	A benchmark for hallucination-free dialogues by editing hallucinated responses in Wizard of Wikipedia (WoW) [9]
HaluEval-QA [19] (Acc) HaluEval-summarization [19] (Acc) HaluEval-Dialogue [19] (Acc)	A large collection of generated and human-annotated hallucinated samples for evaluating the performance of LLMs in recognizing hallucination. It contains the QA, summarization and dialogue tasks.

Table 8: Dataset, models and reference benchmarks for validating CrossCheckGPT. Acc stands for accuracy.

B Exact Prompts

We provide the exact prompts we used in our experiments in Table 9 for various tasks.

Task	Prompt
Text-to-text generation	Generate a passage about <name>.
Image-to-text description	Describe the image in one paragraph.
Visual description for video	Describe the video in one paragraph.
Audio description for video	Describe the audio in one paragraph.
Prompt for speech content	What does the man/woman say in the video?
LLM Judgment for CrossCheck-explicit	Context: <evidence_passage>\n\nSentence: <sentence> \n\nIs the sentence supported by the context above? Answer Yes or No.\n\nAnswer:
CrossCheck-implicit factual errors	You are given the following sentence about <name/image/video> that might be inaccurate:\n<sentence>\n List possible inaccurate information in this sentence.
LLM Judgment for CrossCheck-implicit	You are given the following sentence about <name/image/video>:\n<sentence>\nThe following is an analysis of possible inaccuracies in this sentence:\n<list_of_possible_errors>\nBased on the analysis, determine if the sentence contains any inaccurate information. Answer Yes or No.\n\nAnswer:

Table 9: Exact prompt used for different tasks.

C CrossCheckGPT as a Hallucination Detection Method

CrossCheckGPT can be used as a Hallucination detection method, which performs better than the best output-probability-based method reported in SelfCheckGPT[29].

Evidence Model	Non-Factual	Non-Factual*	Factual	Document (r)
Llama 30B Max($\mathcal{H}$) [29]	80.92	37.32	37.90	35.57
Llama-2-7B-Chat	85.84	57.22	54.41	56.25
Vicuna-v1.5-7B	83.13	53.38	51.13	54.64
Mistral-7B-Instruct-v0.2	**87.21**	**59.60**	**56.72**	**63.04**

Table 10: AUC-PR and document-level correlation against human annotation for detecting hallucinations in GPT-3 using individual evidence models on non-factual and factual statements in WikiBio [29].

D Text-to-text Additional Results

We provide the version of Table 1 with all ten benchmark metrics in Table 11. Moreover, we investigate the *specific-task* hallucination ranking ability where the inputs to SelfCheckGPT and CrossCheckGPT are from a specific task (rather than text generation). We conduct task-specific experiments using the inputs from TruthfulQA MC1 and HaluEval QA containing multiple-choice and yes-no questions respectively. The results in Table 12 show high system-level correlations and moderate document-level correlations, indicating that CrossCheckGPT can operate as a task-specific metric without requiring any ground truth.

Metrics	System(ρ)	Document (r)	
		w/o GPT4	with GPT4
TriviaQA [14]	23.33	-	-
TruthfulQA MC1 [23]	52.94	-	-
TruthfulQA MC2 [23]	57.14	-	-
XSum [35]	-70.00	-	-
CNNDM [39]	38.33	-	-
MemoTrap [31]	10.88	-	-
FaithDial [10]	-8.33	-	-
HaluEval-QA [19]	-18.33	-	-
HaluEval-Summarization [19]	48.33	-	-
HaluEval-Dialogue [19]	46.03	-	-
SelfCheckGPT [29]	66.46	74.06	76.08
CrossCheck-explicit	<u>77.44</u>	**82.28**	<u>77.23</u>
CrossCheck-implicit	56.71	18.33	17.29
CrossCheck-explicit weighted	**82.32**	<u>81.78</u>	**82.18**
CrossCheck-explicit weighted	56.81	20.21	19.16

Table 11: Full version of Table 1 including all other metrics. General hallucination evaluation where the task for SelfCheckGPT/CrossCheckGPT is open-ended text generation on WikiBio. System-level correlation, System(ρ), is measured against the overall ranking in the leaderboard, and document-level correlation, Document(r), is measured against RefCheck. With GPT-4 refers to including GPT-4 as the target LLM.

Metrics	System(ρ)		Document (r)	
	TruthfulQA MC1	HaluEval QA	TruthfulQA MC1	HaluEval QA
SelfCheckGPT	**76.19**	30.95	30.87	6.76
CrossCheckGPT	**76.19**	**88.10**	**33.68**	**22.00**

Table 12: Task-specific hallucination evaluation where the task of SelfCheckGPT/CrossCheckGPT is, in this example, either TruthfulQA MC1 or HaluEval QA. Note that rankings are performed on 8 target models that are instruction-tuned as these tasks are QA-based and require some instruction-following ability.

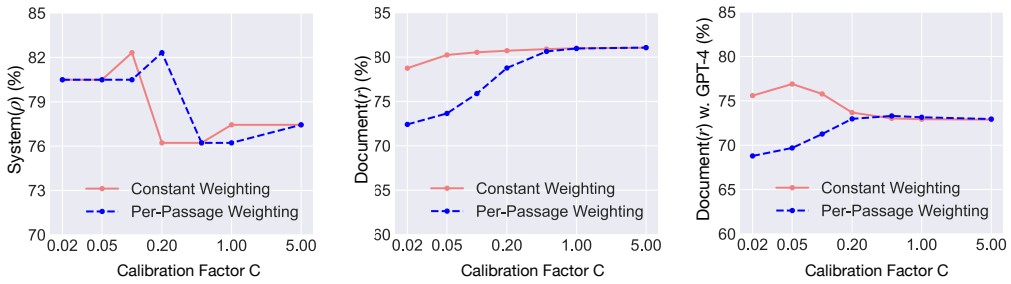


Figure 6: The variation of System(ρ) and Document(r) against calibration temperature T in Eqn. (4) for weighted CrossCheck-explicit. Constant weighting refers to applying the same weight for all documents, while per-passage weighting refers to the use of passage-specific weighting derived from SelfCheckGPT scores of each passage.

We first show the **variation of system and document-level correlation against varying calibration temperatures** for CrossCheck-explicit weighted in Fig. 6 using WikiBio data. A comparison between using per-query weights and using the same weights for the entire task is also provided. As a result, $C = 0.1$ is chosen as it achieves the best system-level correlation. Besides, the same weighting across the whole task is used at $C = 0.1$ as the large variance among weights of different queries introduces more noise in scoring and hence hinders the correlation.

E System-level Correlations between Individual Text-based Hallucination Benchmarks

We provide the system-level correlations between individual text-based hallucination benchmarks to show that they capture different aspects and do not correlate well with each other in Table 13.

	TriviaQA	TruthfulQA	Xsum	CNN-DM	MemoTrap	FaithDial	HaluQA	HaluSumm	HaluDial
TriviaQA [14]	1.00	0.20	-0.72	0.15	0.07	0.13	0.27	0.40	0.50
TruthfulQA [23]	0.20	1.00	-0.10	0.38	0.27	0.05	-0.50	0.37	0.63
Xsum [35]	-0.72	-0.10	1.00	-0.03	-0.40	0.12	-0.57	-0.63	-0.68
CNN-DM [39]	0.15	0.38	-0.03	1.00	0.28	-0.05	-0.05	0.33	0.37
MemoTrap [31]	0.07	0.27	-0.40	0.28	1.00	-0.05	-0.08	0.48	0.17
FaithDial [10]	0.13	0.05	0.12	-0.05	-0.05	1.00	-0.03	-0.22	-0.13
HaluQA [19]	0.27	-0.50	-0.57	-0.05	-0.08	-0.03	1.00	0.30	0.20
HaluSumm [19]	0.40	0.37	-0.63	0.33	0.48	-0.22	0.30	1.00	0.67
HaluDial [19]	0.50	0.63	-0.68	0.37	0.17	-0.13	0.20	0.67	1.00

Table 13: System-level correlation (ρ) between each pair of the 9 selected benchmarks metrics.

F Scatter Plots and Statistical Significance for Image-to-text

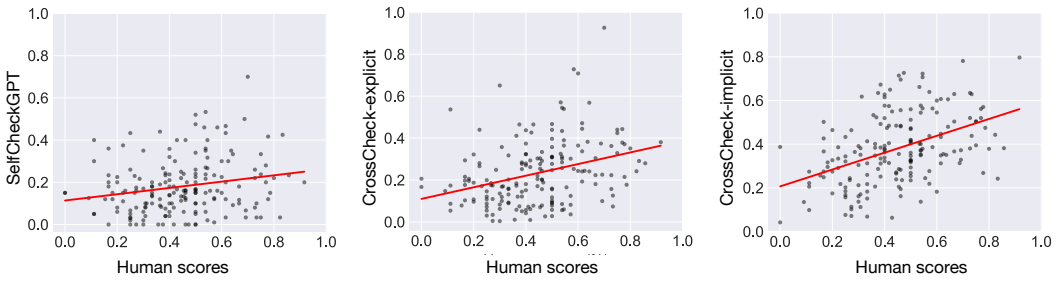


Figure 7: Scatter plot of SelfCheckGPT, CrossCheck-explicit and CrossCheck-implicit scores against human annotation for image-to-text tasks.

The scatter plot, similar to text-to-text ones in Fig. 4, is shown in Fig. 7.

Methods	Success rate (p-value)
CrossCheck-explicit	65.5% (<0.00001)
CrossCheck-implicit	84.5% (<0.00001)
CrossCheck-explicit weighted	67.0% (<0.00001)
CrossCheck-implicit weighted	**88.0%** (<0.00001)

Table 14: Success rate and statistical significance of CrossCheckGPT approaches measured via sign-test on independent subsets of images.

Additionally, we report the statistical significance of CrossCheckGPT being better than SelfCheckGPT on MHaluBench by performing the sign test at the image level.

G Statistics of AVHalluBench

We provide detailed statistics about AVHallubench in Table 15, including the number of videos, average lengths of each subset, as well as various audio and visual elements involved.

Source Dataset	Num. of Videos	Avg. Length (sec.)	w/ Speech	w/ Music	w/ Visual Text
NeXT-QA [52]	32 (18%)	22.0	19	7	1
M3AV [5]	27 (16%)	11.3	27	0	27
How2 [38]	27 (16%)	9.5	27	4	2
MUSIC-AVQA [18]	23 (13%)	29.0	0	23	0
VALOR32k [3]	26 (15%)	8.7	11	7	8
FAVDBench [40]	38 (22%)	8.0	8	15	13
Overall	175	14.2	92 (52%)	56 (32%)	51 (29%)

Table 15: Statistics of the AVHalluBench dataset with the percentage shown in brackets.

H Additional SelfCheckGPT and CrossCheckGPT Scores on AVHalluBench

We provide the detailed SelfCheckGPT and CrossCheckGPT scores on AVHalluBench for all MLLMs that handle video or audio inputs in this paper in Table 16 for video descriptions and Table 17 for audio descriptions.

Models	SelfCheckGPT	CrossCheck-explicit	CrossCheck-implicit
Valley [26]	52.43	55.98	48.22
Video-LLaVA [22]	30.59	33.52	<u>40.57</u>
Chat-Univi [17]	<u>29.40</u>	<u>32.68</u>	41.75
LLaMA-VID [21]	38.61	39.14	**40.48**
Video-LLaMA [55]	41.14	52.02	48.80
FAVOR [41]	60.67	53.85	50.49
Gemini 1.5 Pro	**19.87**	**31.74**	-

Table 16: SelfCheckGPT and CrossCheckGPT scores for 6 visual-LLMs that take video as inputs on AVHalluBench. Note that FAVOR, Video-LLaMA and Gemini 1.5 Pro are only given visual inputs. Gemini 1.5 Pro was not used for CrossCheck-implicit.

Models	SelfCheck		CrossCheck-explicit		CrossCheck-implicit	
	audio	w. speech	audio	w.speech	audio	w. speech
LTU [12]	**21.95**	-	<u>37.44</u>	-	<u>18.06</u>	-
Qwen-Audio-Chat [7]	36.57	37.08	43.66	43.41	20.21	<u>52.20</u>
SALMONN [42]	34.99	<u>34.80</u>	42.21	<u>40.15</u>	18.32	**48.17**
FAVOR [41]	49.62	41.51	66.69	55.41	23.26	61.01
Video-LLaMA [55]	56.42	-	68.05	-	**17.10**	-
Gemini 1.5 Pro	<u>25.82</u>	**27.38**	**34.66**	**36.52**	-	-

Table 17: SelfCheckGPT and CrossCheckGPT scores for 6 audio-LLMs on AVHalluBench. Note that FAVOR and Video-LLaMA are only given audio inputs. Gemini 1.5 Pro was not used for CrossCheck-implicit.

I CrossCheck-explicit vs. CrossCheck-implicit

We present the average SelfCheckGPT scores on each task together with the system-level correlations in Table 18 to support our recommendations on CrossCheck-explicit and CrossCheck-implicit.

J Case Studies for Hallucination with Audio-Visual Inputs

In addition to the piano example shown in Fig. 10 that has been mentioned in the main text, we show here two additional examples in Fig. 9 and Fig. 8 where audio-visual inputs influence the hallucination compared to using audio or visual inputs alone.

Tasks	Ave. $\mathcal{S}_{\text{selfcheck}}$	System(ρ) CrossCheck-explicit	CrossCheck-implicit
Text-to-text	40.63	77.44	56.71
Image-to-text	17.16	42.86	50.42
Audio description	39.91	71.67	40.00
Visual description	42.14	89.09	54.29

Table 18: SelfCheckGPT scores and system-level correlations using CrossCheck-explicit and CrossCheck-implicit on four tasks. The system-level correlation for audio and visual descriptions is measured against RefCheck, and that for text-to-text and image-to-text tasks are measured against overall ranking.

Visual-only video description: A man in a blue and white jacket and khaki pants is standing in an entryway. He is pointing to a mural on the wall that looks like a stone wall with vines and flowers. There is a vase of flowers and a white dove on a shelf in the mural. There is a staircase with a black metal railing to the right of the mural. The man is explaining that he painted the mural to make the entryway look more inviting.

Audio-visual video description: A man in a blue and white tracksuit is standing in an entryway with a faux stone wall and a staircase. He is pointing to the wall and describing how the viewer's eye will travel up the wall to a mirror at the top of the landing. The audio consists of the man's voice describing the scene.

Hallucination Content Less hallucinated contents

Figure 8: Example of audio-visual hallucination problem from Gemini 1.5 Pro. In this example, even when no audio is provided, the model still describes what the man is talking about, and having audio inputs greatly benefits the description by reducing the hallucination in describing the man's speech.

K Limitations

Our investigation is limited in the following aspects: First, hallucination is an expansive area and, as done in other studies, this paper only covers a reasonable subset of all possible domains. However, we plan to release a live hallucination leaderboard where we plan to benchmark the performance of further MLLMs over more benchmark metrics. Secondly, while the confidence-based weighting mechanism improves the performance of CrossCheckGPT, it does not take into account the similarities of different evidence models. Correlation between models, due to having similar training data or from starting at the same checkpoints, may result in evidence models making similar mistakes. This poses a future research direction, in raking model correlation into account for the weighting mechanism. Lastly, there is limited by the number of currently available audio-visual LLMs for evidence generation.

L Broader Impact

Hallucinations in multimodal foundation models have become increasingly critical and challenging. Therefore, providing a general reference-free hallucination benchmarking approach is necessary and timely, enabling practitioners to have metrics for model trustworthiness. Therefore, CrossCheckGPT has the following positive broad impact:

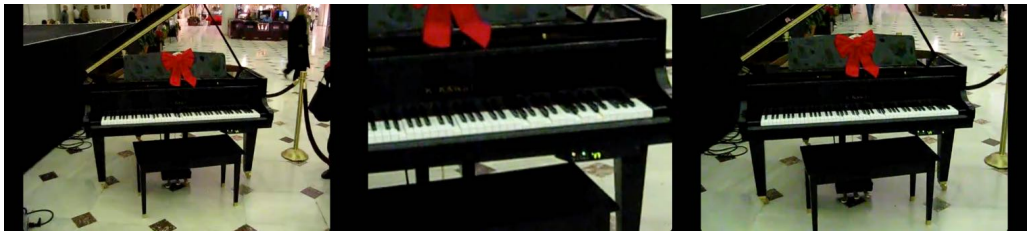
- CrossCheckGPT establishes a universal ranking system which helps identify more factual and faithful models to be selected in particular applications, reducing the dissemination of misinformation and increasing societal confidence in AI applications.

Visual-only video description: First, we see a person standing in a dark room with a flashlight on the ground. Then, a person walks into the room, followed by another person walking into the room. They both have backpacks and are wearing grey sweatshirts. Next, we see a woman wearing a white hat and a red baseball cap on her head. She is standing in the room with the two men. Overall, the video seems to be of people exploring a dark room or cave, using flashlights and navigational tools.

Audio-visual video description: In the video, a man is seen playing a first-person shooter video game. The sound of rain can be heard in the background. The man presses a button on the controller, and a robotic voice is heard. He then opens a door. A black cat can be seen sitting on a window sill in the room.

New hallucination Less hallucinated contents

Figure 9: Example of audio-visual hallucination problem from FAVOR. In this example, the audio is the man explaining what he is doing in the game. The speech description reduces the hallucination of "pressing the button" and "opening a door" in the visual description with new but random hallucinations coming out.



Visual-only video description: The video shows a grand piano placed on the stage in an auditorium. A red ribbon is tied around the leg of the piano, and the auditorium appears to be deserted. The piano has a sleek, glossy black finish and is in good condition. The video captures various shots of the piano from different angles, highlighting its size and design. The stage and the auditorium look well-maintained and spacious. The video ends with the piano and the red ribbon as the only objects.

Audio-visual video description: The video shows people standing and sitting at a train station, with a piano in the background playing music. The piano is being played by a woman wearing a hat and a woman in a green dress. The music is upbeat and there are various people walking by in the background. The video has a muted color tone, with grays and whites predominating.

New hallucination Less hallucinated contents

Figure 10: Example of audio-visual hallucination problem. In this example, the audio is the piano itself playing, which introduces additional hallucination to the visual description which describes it as "played by a woman".

- CrossCheckGPT provides a reliable ranking that would aid regulatory bodies in enforcing compliance standards for multimodal foundation models, particularly in critical areas such as healthcare, finance, and public safety.
- As a reference-free and versatile benchmarking method, CrossCheckGPT can drive developers to innovate and improve their multimodal foundation models.

However, our method by no means provides perfect hallucination scores and may inherit potential bias from the chosen evidence models. Therefore, practitioners should be independently educated and avoid overreliance on the rankings, as doing so may lead to complacency in critical thinking and reduced vigilance. From the model aspect, the approach in this paper does not give rise to any additional potential biases beyond the ones directly inherited from the pre-trained LLM checkpoints.

M Computing Resource

Our experiments are performed on a single Nvidia A100 GPU for inference. The average inference time for each target model to get the CrossCheckGPT score is 20 hours. The total amount of time to run for all models in the text-to-text leaderboard is 200 hours, in the image-to-text leaderboard is 190 hours and in the AVHalluBench is 240 hours. The total GPU hours for running the full research is 2000. There is no training process involved in the research.

N Assets and License Explanation

Links to the following licenses that apply to the models used in the paper are provided (see Table 7).

- Llama2: https://huggingface.co/meta-llama/Llama-2-7b-chat-hf/blob/main/LICENSE.txt
- Apache-2.0: https://www.apache.org/licenses/LICENSE-2.0
- MIT License: https://choosealicense.com/licenses/mit/
- BSD 3-Clause License: https://github.com/salesforce/LAVIS/blob/main/LICENSE.txt
- Tongyi Qianwen: https://github.com/QwenLM/Qwen-Audio/blob/main/LICENSE

The following licenses are applied to the datasets used in our paper:

- CC-BY-SA-3.0: Used by WikiBio hallucination data [29]. License link: https://spdx.org/licenses/CC-BY-SA-3.0.
- MIT License: Used by MHaluBench (https://huggingface.co/datasets/openkg/MHaluBench). License link see above.

The following licenses are applied to the code and Python packages we use for our experiments:

- Apache-2.0: Applies to Huggingface Transformers (https://github.com/huggingface/transformers/blob/main/LICENSE) and UniHD (https://github.com/OpenKG-ORG/EasyDetect/blob/main/LICENSE).
- MIT License: Applies to SelfCheckGPT (https://github.com/potsawee/selfcheckgpt/blob/main/LICENSE) and spaCy (https://github.com/explosion/spaCy/blob/master/LICENSE).