CrossCheckGPT: Universal Hallucination Ranking for Multimodal Foundation Models

Anonymous ACL submission

Abstract

Given the diversity in architectures, training data and instruction tuning techniques, multimodal foundation models can exhibit large variations in susceptibility to hallucinations. Existing hallucination ranking approaches require gold-standard references or labels for specific tasks, which limits their applicability for hallucination benchmarking in 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 hallucinated content is less consistent across systems than factual content, 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 demonstrate the applicability of our method for hallucination ranking across various input 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.

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 (Liu et al., 2024; Wang et al., 2023a; Tonmoy et al., 2024). Hallucination impacts many generative applications and can lead to misinformation (Yang et al., 2023; Nahar et al., 2024). 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 QA (Lin et al., 2022; Han et al., 2019; Li et al., 2023a; Dziri et al., 2022; Thorne et al., 2018), summarization (Maynez et al., 2020; Manakul et al., 2023a), biography generation (Manakul et al., 2023b), instruction following (McKenzie et al., 2023), image captioning (Rohrbach et al., 2018), and visual QA (Li et al., 2023d; Wang et al., 2023b). Many benchmarks use proxy measures, such as answering questions designed to trigger hallucinations, but are task-specific and rely on gold-standard labels, limiting generalizability. On the other hand, hallucination detection approaches such as SelfCheckGPT (Manakul et al., 2023b) and UniHD (Chen et al., 2024a) 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.

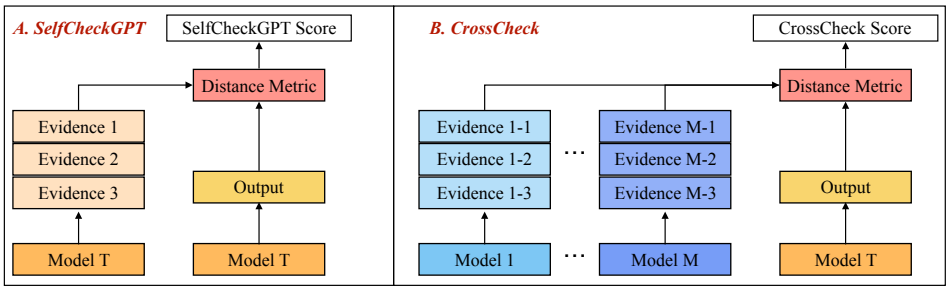

Figure 1: SelfCheckGPT (Left) and CrossCheckGPT (Right) for hallucination rankings. The approach can rank a set of MLLMs on any task without reference, enabling hallucination benchmarks for various generative tasks.

The effectiveness of CrossCheckGPT is demonstrated as a *universal* evaluation framework for any Multimodal Large Language Model (MLLM) that generates text outputs. It is applicable irrespective of the input modality. Two information consistency measures are investigated: CrossCheck-explicit, which generates multiple text samples from each evidence system; and CrossCheck-implicit, which prompts the evidence model to determine whether it agrees with the assessed outputs.

Additionally, AVHalluBench, a comprehensive audio-visual hallucination benchmark, is developed covering various styles, domains, and elements like visual text, speech, and music. As the first study on audio-visual hallucination benchmarking, AVHalluBench is used to rank recent audio and video LLMs such as Gemini 1.5 Pro. The key contributions of this work 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 input 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 (Manakul et al., 2023b) and UniHD (Chen et al., 2024a).
- We analyze hallucination in 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 often use proxy tasks to assess the likelihood of LLM factual errors. QA-based benchmarks, such as TriviaQA (Han et al., 2019), TruthfulQA (Lin et al., 2022), HaluEval-QA (Li et al., 2023a), MemoTrap (McKenzie et al., 2023) and FEWL (Wei et al., 2024) design questions to probe the truthfulness and factual accuracy and rank systems by accuracy. Other methods, such as FaithDial (Dziri et al., 2022), XSum (Narayan et al., 2018) and CNN-DM (See et al., 2017) measure hallucination in dialogue or summarization. These benchmarks require references (i.e. ground-truth answers) to compare to model-generated outputs. In contrast, SelfCheckGPT (Manakul et al., 2023b) ranks systems by measuring self-consistency but was designed for hallucination detection and may lack cross-system calibration.

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 (Rohrbach et al., 2018), LURE (Zhou et al., 2024) and MHaluBench (Chen et al., 2024a), directly evaluate the generated text descriptions of images using gold-standard annotations or external toolkits. Another stream of methods, such as POPE (Li et al., 2023d) and HallusionBench (Guan et al., 2024), curate a set of questions with short answers trying to capture various aspects of hallucination. Meanwhile, AMBER (Wang et al., 2023b) combines both generation and question answering in one single benchmark. Unlike these methods, CrossCheckGPT does not rely on a 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 its responses are supported by evidence generated from other MLLMs (denoted as *evidence models*).

These scores are used to rank MLLMs. As shown in Fig. 2, we explore two information consistency measures: CrossCheck-explicit and CrossCheck-implicit described in §3.1. CrossCheckGPT is reference-free and can be generally applied to MLLMs of any input modality and response type.

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.

$$\mathcal{S}_{\text{self}}(\hat{M}) = \frac{1}{|\mathcal{Q}|}\frac{1}{I}\sum_{Q\in|\mathcal{Q}|}\sum_{i=1}^{I}\mathcal{S}^{\text{self}}_{r_i,Q}(\hat{M}) \quad (1)$$

$$\text{where} \quad \mathcal{S}^{\text{self}}_{r_i,Q}(\hat{M}) = \frac{1}{\hat{N}}\sum_{n=1}^{\hat{N}} x^{(n)}_{r_i,Q}(\hat{M})$$

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

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 $\hat{M}$ is computed using Eqn. 2,

$$\mathcal{C}_{\text{explicit}}(\hat{M}) = \frac{1}{|\mathcal{Q}|}\frac{1}{I}\sum_{Q\in|\mathcal{Q}|}\sum_{i=1}^{I}\mathcal{C}^{\text{explicit}}_{r_i,Q}(\hat{M}) \quad (2)$$

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

Here, $\mathcal{M}$ denotes the set of evidence models for CrossCheck-explicit. Self-consistency can be taken into account by including the target model $\hat{M}$ into the set of 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 reliability levels, a factor η_j is assigned to the passages generated from 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 CrossCheck-implicit score is computed using Eqn. 3,

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

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

where $y_{r_i,Q}(M_j)$ denotes the hallucination score of sentence r_i using CrossCheck-implicit. Unlike CrossCheck-explicit (which computes $x_{r_i,Q}(M_j)$), $y_{r_i,Q}(M_j)$ is obtained by prompting the evidence model M_j to check for factual errors in r_i based on input Q. The LLM judge then uses r_i and analysis from model M_j to predict $y_{r_i,Q}(M_j)$, whether the response is hallucinatory. If hallucinatory, $y_{r_i,Q}(M_j) = 1$, otherwise $y_{r_i,Q}(M_j) = 0$. Concurrent work, PoLL (Verga et al., 2024) employs multiple models to evaluate texts, similar to CrossCheck-implicit. However, this work focuses on multimodal inputs and hallucination benchmarking.

3.2 Confidence Weights for Evidence Models

The quality of evidence may depend on the propensity of a model to hallucinate. Thus, we propose a mechanism to weight the scores based on model uncertainty reflected by SelfCheckGPT:

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

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. This weighting mechanism also ensures that outlier systems will not be undermined by the evidence from weaker models.[1]

4 CrossCheckGPT for Hallucination with Multimodal Inputs

CrossCheckGPT is designed for any input modality, provided that the outputs are of a consistent form

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

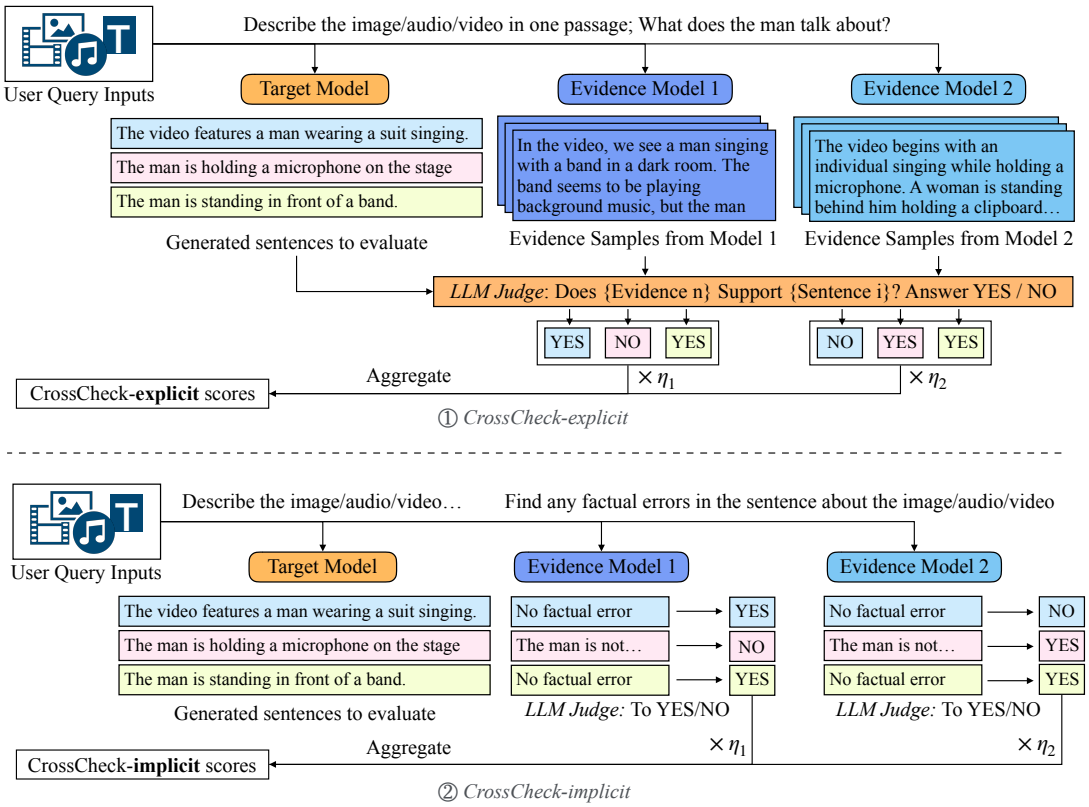

Figure 2: Illustration of the CrossCheckGPT approach with two evidence models as an example. Two information consistency measures are shown. ① CrossCheck-explicit where N passages are stochastically generated by sampling from each evidence model and ② CrossCheck-implicit where evidence models are directly 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.

(i.e. text) and a suitable information consistency measure is used. This design of CrossCheckGPT allows it to rank multi-modal systems, i.e. systems which use two or more input modalities.

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 of 175 videos from six video understanding datasets covering various styles and elements, with statistics in Table 15 in Appendix F. To evaluate benchmarking methods, AVHalluBench includes carefully written hallucination-free descriptions for audio and visual content. 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 assess inter-annotator agreement, each description is split into atomic facts (Min et al., 2023) and verified against the other annotators' descriptions as either: *Supporting*, such that the fact is supported by the other annotator, *Contradicting*, such that the fact contradicts the information provided by

[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 by Gemini-1.5/GPT-4v, prompted to describe all elements in the frame sequence. While not hallucination-free, they have high coverage and subjective details. Annotators used them as guidance but were instructed to write only objective details.

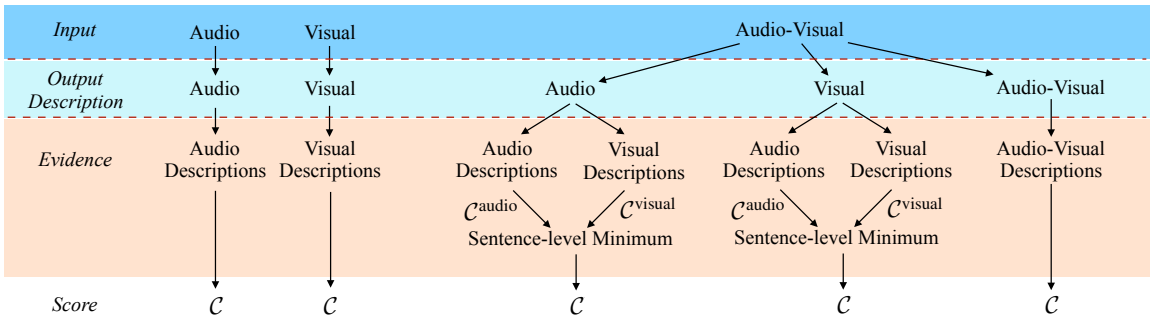

Figure 3: CrossCheckGPT score computation for AVHalluBench with audio, visual and audio-visual inputs.

the other annotator, or *Neutral* such that the facts neither support nor contradict one another. Both the decomposition and verification processes were automated using GPT-4. Of 39 videos annotated by multiple annotators, there were 471 audio-related facts and 913 visual-related facts, with agreement rates counted as Supporting/Neutral/Contradicting of 64.6%/24.6%/10.8% and 62.0%/29.0%/9.0%.

5 Experiments

The central hypothesis for CrossCheckGPT is first examined manually (§5.1). Then, CrossCheckGPT is validated with three input modalities: text (§5.2), image (§5.3), and audio-visual (§5.4). Last, two variants of CrossCheckGPT are compared (§5.5) and ablation studies provided (§5.6). The inference hyperparameters are temperature = 1.0, beam size = 1, top-p = 0.9. Models and prompts are listed in Appendix A and B. *SelfCheckGPT* is applied as a hallucination ranking baseline for all modalities as it is reference-free and not task-specific.

5.1 Hypothesis Justification

As CrossCheckGPT is based on this hypothesis,

The distribution of hallucinated content is less consistent across systems than factual content.

we first examine it by counting the occurrences of the same fact that appears in the responses of other models where these facts are based on the atomic facts in MHaluBench. Then, using the annotations from Section 5.3, the average occurrence rate of true facts and hallucinations are computed respectively, as presented in Table 1. Evidently, true facts are consistently more frequent than hallucinations by a significant margin. While some hallucinations are similar across some models, they occur at a much lower rate compared to true facts. Further analysis via examples and pairwise CrossCheckGPT scores are provided in Appendix L.

	True Facts	Hallucinations
%Occur. Rate ($\mu \pm \sigma$)	$\mathbf{79}_{\pm 19}$	$40_{\pm 26}$

Table 1: Occurrence rate of the true facts and hallucinations. If a fact appears 40 times in 80 evidence passages from all evidence models, the occurrence rate is 50% for that fact. Means and standard deviations are computed across all annotated facts in MHaluBench.

5.2 Text-to-text Experiments

First, the main text-to-text experiments are performed using the subset of WikiBio data used in (Manakul et al., 2023b), which contains 238 biographical passages from Wikipedia. We select 10 open-source LLMs as target models, 8 of which are used as evidence models. Four models are Llama-2-7B based (Touvron et al., 2023) (e.g. Vicuna-v1.5-7B (Chiang et al., 2023)) and four models are Mistral-7B based (Jiang et al., 2023). Each evidence model generates 20 stochastic passages. For the LLM judge in CrossCheck-explicit, Mistral-7B (Jiang et al., 2023) is used as it achieves the best results among all considered open-source LLMs (shown in Table 11 in Appendix C). The recently proposed method, PRD (Li et al., 2023b) which adopts multiple LLMs for pairwise comparison ranking, is also used as a baseline in addition to SelfCheckGPT.

To evaluate the *general* benchmarking ability of ranking methods, 10 benchmark metrics from the hallucinations leaderboard (Hong et al., 2024) (in Table 9) are selected to provide the overall hallucination ranking of the systems. These metrics are based on human annotation or gold-standard references, where the overall rankings are obtained by averaging the rankings from each metric.

The *system-level* correlation is computed between the hallucination ranking methods and the overall ranking using Spearman's Rank Correlation coefficient (SRC), denoted as System(ρ). Since WikiBio contains reference texts, these can be used

as evidence for an idealized fact-checking method, termed RefCheck. CrossCheckGPT and SelfCheckGPT scores are compared against RefCheck at the *document-level* using Pearson's Correlation Coefficient (PCC), denoted as Document(r). Furthermore, to assess the effectiveness of CrossCheckGPT with a more powerful target LLM, GPT-4 is included alongside 10 target LLMs.

Hallucination Ranking Results: As shown in Table 2, existing hallucination metrics such as HaluEval-QA accuracy show weak correlations with the overall system ranking, with some metrics even negatively correlated and the highest (TruthfulQA MC2) at 57.14%. This is likely because most metrics target only a single aspect of hallucinations, e.g., probing through question-answering. CrossCheck-explicit shows the strongest correlation with overall rankings compared to other methods, with the uncertainty-weighted version achieving the highest correlation, highlighting its effective general hallucination ranking ability.

Metrics	System(ρ)	Document(r) w/o GPT4	Document(r) w/ GPT4
Text-to-text Baselines			
TruthfulQA MC2	57.14	-	-
SelfCheckGPT	66.46	74.06	76.08
PRD	72.56	–	–
CrossCheckGPT			
Implicit	56.71	18.33	17.29
Explicit	77.44	**82.28**	77.23
Implicit weighted	56.81	20.21	19.16
Explicit weighted	**82.32**	81.78	**82.18**

Table 2: General hallucination evaluation where the task for SelfCheckGPT/CrossCheckGPT is open-ended biography generation on WikiBio. System(ρ), is measured against the overall ranking of the leaderboard, and Document(r) is measured against RefCheck. "With GPT-4" refers to including GPT-4 as a target model. Additional metrics are presented in Table 12 in Appendix D, and the scatter plots of CrossCheckGPT and SelfCheckGPT are provided in Fig. 4.

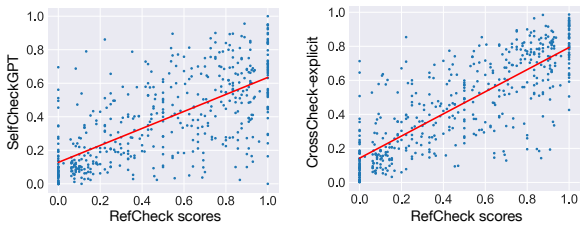

Figure 4: Scatter plot of document-level scores for the SelfCheckGPT baseline versus CrossCheck-explicit against RefCheck for text-to-text experiments.

5.3 Image-to-text Experiments

Second, CrossCheckGPT is validated for the hallucination ranking of visual LLMs on image-to-text tasks. The experiments are conducted using MHaluBench (Chen et al., 2024a), an image-captioning hallucination dataset. Nine visual LLMs are selected as target models, all of which are used to generate evidence passages (see Table 8 for the list of models). Each evidence model generates 10 image descriptions per image. The overall ranking is obtained by averaging the rankings from CHAIR (Rohrbach et al., 2018) and POPE (MSCOCO subset) (Li et al., 2023d).[4] In addition to the general baseline like SelfCheckGPT, UniHD (Chen et al., 2024a) is used as a stronger baseline.

For evaluation, we take a subset of 30 image descriptions from MHaluBench generated by each target model (from a total of 270 passages with 3237 facts) and each description is annotated 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).

Hallucination Ranking Results: Table 3 presents the system-level and image-level correlations against overall rankings and rankings derived from human annotations. Both variants of CrossCheckGPT outperform the baselines, with CrossCheck-implicit weighted performing the best out of all methods, achieving a 98.33% correlation with the rankings from human annotations.

5.4 Video-to-text Experiments

Third, CrossCheckGPT is applied 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 (Sun et al., 2023), Video-LLaMA (Zhang et al., 2023), and Gemini 1.5 Pro (Team, 2024), 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 sepa-

[4]CHAIR and POPE are the two popular representative metrics for free-form text generation and binary classification hallucination benchmarks respectively (Wang et al., 2023b).

Metrics	System(ρ)			Image(r)
	Overall	CHAIR	Human	Human
Image-to-text Baselines				
UniHD	42.02	36.98	33.33	36.70
SelfCheckGPT	43.70	23.10	-10.00	20.93
CrossCheckGPT				
Implicit	**50.42**	**64.71**	**98.33**	48.72
Explicit	42.86	43.70	75.00	35.16
Implicit weighted	**50.42**	**64.71**	**98.33**	**52.83**
Explicit weighted	47.06	46.22	73.33	36.98

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

Metrics	Visual Desc.		Audio Desc.	
	Sys(ρ)	Vid(r)	Sys(ρ)	Vid(r) (w/speech)
Video-to-text Baseline				
SelfCheckGPT	86.67	65.77	60.00	51.13 (44.55)
CrossCheckGPT				
Implicit weighted	54.29	30.73	40.00	2.15 (16.20)
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 CrossCheckGPT uses $C = 0.1$. Ranking correlations for systems that handle speech are in brackets.

Model	Input	Visual Desc.		Audio Desc.	
		$\mathcal{S}_{\text{self}} \downarrow$	$\mathcal{C}_{\text{exp}} \downarrow$	$\mathcal{S}_{\text{self}} \downarrow$	$\mathcal{C}_{\text{exp}} \downarrow$
FAVOR	V	60.67	53.85	—	—
	A	—	—	49.62	66.69
	A+V	56.42	49.60	33.25	35.20
Video-LLaMA	V	41.14	52.02	—	—
	A	—	—	56.42	68.05
	A+V	47.73	49.13	70.23	41.25
Gemini 1.5 Pro	V	19.87	31.74	—	—
	A	—	—	**25.82**	34.66
	A+V	**12.77**	**23.27**	48.51	**28.79**

Table 5: SelfCheckGPT scores ($\mathcal{S}_{\text{self}}$) and weighted CrossCheck-explicit scores ($\mathcal{C}_{\text{exp}}$) with $T = 0.1$ on AVHalluBench for audio-visual LLMs. Input modality: A = Audio, V = Visual, A+V = Audio-Visual.

rate 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 serve as evidence models in CrossCheck-explicit, except Gemini 1.5 Pro due to the number of request limitations, 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. Since only five target models handle speech, we provide a dedicated ranking for these models, using prompts that explicitly request speech descriptions.

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 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 5.5.

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 using a single modality. Table 5 presents the average raw hallucination scores (rather than correlations), for three MLLMs that can take audio-visual inputs.

As expected, examining the CrossCheckGPT scores shows that audio-visual inputs reduce hallucination rates. Although Gemini 1.5 Pro achieved the best scores, it is more prone to hallucination with silent videos, often fabricating audio descriptions. 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 is likely because, 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 can cause hallucinations. For example, in Fig 11 of Appendix K where there is a self-playing piano, audio 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.

5.5 Crosscheck Explicit versus Implicit

Implicit is more sample-efficient than *Explicit*, requiring only one inference, but its performance is more sensitive to the task. In text-to-text and

video-to-text experiments, Implicit underperforms Explicit, unlike in image-to-text tasks. We hypothesize that for complex, open-ended tasks, Explicit is preferable as it separates evidence generation from verification, providing more calibrated uncertainty measures. However, in other circumstances, Implicit helps the model focus on specific aspects of the input and yield more accurate rankings. For tasks with diverse outputs, the raw SelfCheckGPT scores are expected to be high, and they can be used to select a CrossCheckGPT variant. For example, the raw SelfCheckGPT score across models is 40.63% for text-to-text, which is considerably higher than 17.16% for image-to-text. We recommend using *Explicit* when the SelfCheckGPT scores are high, and *Implicit* when they are sufficiently low. This is further illustrated in Table 18 in Appendix I.

5.6 Ablation Studies

Self-Bias: LLMs are known to have self-preferential bias (Zheng et al., 2023) and may prefer outputs from similar models. Thus, LLMs of the same/similar base model may provide inflated CrossCheckGPT scores. Table 6 shows that self-bias is present, 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. Nevertheless, the bias can be mitigated by adopting a wide range of evidence models, which is adopted in CrossCheckGPT scores, hence achieving a reliable evaluation with strong correlations.

Model Family	Sys(ρ)	Doc(r)	$\mathcal{C}_{\text{exp}}$(Vicuna)	$\mathcal{C}_{\text{exp}}$(Mistral)
Llama-2	55.49	81.10	42.94	45.68
Mistral	81.71	81.06	44.98	41.81
All	82.32	82.28	44.82	44.93

Table 6: 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.

Robustness to Manipulation: The way generation temperature (which can be set for any model) affects hallucination scores is examined. Fig. 5 shows that when increasing the temperature from 0.5 to 1.5, CrossCheckGPT maintains stable rankings across all temperatures. In contrast, SelfCheckGPT has the raw score increase by up to 35% (shown in Fig. 8 in Appendix J), which can significantly alter rankings.

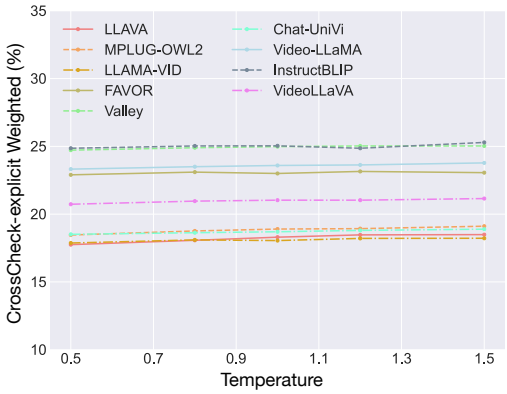

Figure 5: Variation of weighted CrossCheck-explicit scores against temperature during generation. Fig. 8 in Appendix J shows the same plot for SelfCheckGPT.

Robustness to Selection of Evidence Models: In our experiments, we selected widely used and high-performing evidence models that ranked at the top of both performance and hallucination benchmarks. As expected, increasing the number of evidence models improved performance and robustness. Table 3 shows the system correlation using 2, 4, or 6 evidence models while maintaining an equal number of Llama-based and Mistral-based models.

Number of Evidence Models	System(ρ) ($\mu \pm \sigma$)
1 (i.e. SelfCheckGPT)	66.46
Select 2 from 8	71.57 $\pm$ 4.66
Select 4 from 8	73.80 $\pm$ 3.62
Select 6 from 8	75.92 $\pm$ 3.78
All 8 models	**82.32**

Table 7: Performance vs subsets of evidence models randomly selected on the text-to-text task. The balance between llama-based and mistral-based models is kept.

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.

7 Limitations

Our investigation has limitations 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 the performance of further MLLMs is measured over more benchmark metrics. Second, 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 taking model correlation into account for the weighting mechanism. Last, there is a limited number of currently available audio-visual LLMs for evidence generation.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models. *arXiv:2311.16867*.

Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. 2023. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv:2304.08345*.

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

Zhe Chen, Heyang Liu, Wenyi Yu, Guangzhi Sun, Hongcheng Liu, Ji Wu, Chao Zhang, Yu Wang, and Yanfeng Wang. 2024b. M^3av: A multimodal, multigenre, and multipurpose audio-visual academic lecture dataset. *arXiv:2403.14168*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90% chatgpt quality.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo Ponti, and Siva Reddy. 2022. Faithdial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490.

Shangbin Feng, Vidhisha Balachandran, Yuyang Bai, and Yulia Tsvetkov. 2023. FactKB: Generalizable factuality evaluation using language models enhanced with factual knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 933–952, Singapore. Association for Computational Linguistics.

Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James R. Glass. 2024. Listen, think, and understand. In *The Twelfth International Conference on Learning Representations*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*.

Moonsu Han, Minki Kang, Hyunwoo Jung, and Sung Ju Hwang. 2019. Episodic memory reader: Learning what to remember for question answering from streaming data. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4407–4417, Florence, Italy. Association for Computational Linguistics.

Giwon Hong, Aryo Pradipta Gema, Rohit Saxena, Xiaotang Du, Ping Nie, Yu Zhao, Laura Perez-Beltrachini, Max Ryabinin, Xuanli He, and Pasquale Minervini. 2024. The hallucinations leaderboard–an open effort to measure hallucinations in large language models. *arXiv preprint arXiv:2404.05904*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv:2310.06825*.

Peng Jin, Ryuichi Takanobu, Caiwan Zhang, Xiaochun Cao, and Li Yuan. 2024. Chat-univi: Unified visual

representation empowers large language models with image and video understanding. In *CVPR*.

Guangyao li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. 2022. Learning to answer questions in dynamic audio-visual scenarios. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Ruosen Li, Teerth Patel, and Xinya Du. 2023b. Prd: Peer rank and discussion improve large language model based evaluations. *arXiv: 2307.02762*.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2023c. Llama-vid: An image is worth 2 tokens in large language models. *arXiv:2311.17043*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023d. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

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

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. A survey on hallucination in large vision-language models. *arXiv:2402.00253*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. 2023. Valley: Video assistant with large language model enhanced ability. *arXiv:2306.07207*.

Dakota Mahan, Ryan Carlow, Louis Castricato, Nathan Cooper, and Christian Laforte. Stable beluga models.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023a. MQAG: Multiple-choice question answering and generation for assessing information consistency in summarization. In *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. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023b. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Ian R. McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, Andrew Gritsevskiy, Daniel Wurgaft, Derik Kauffman, Gabriel Recchia, Jiacheng Liu, Joe Cavanagh, Max Weiss, Sicong Huang, The Floating Droid, Tom Tseng, Tomasz Korbak, Xudong Shen, Yuhui Zhang, Zhengping Zhou, Najoung Kim, Samuel R. Bowman, and Ethan Perez. 2023. Inverse scaling: When bigger isn't better. *TMLR*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv:2306.02707*.

Mahjabin Nahar, Haeseung Seo, Eun-Ju Lee, Aiping Xiong, and Dongwon Lee. 2024. Fakes of varying shades: How warning affects human perception and engagement regarding llm hallucinations. *arXiv:2404.03745*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *arXiv:2303.08774*.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A large-scale dataset for multimodal language understanding. In *Proc. ViGIL*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Xuyang Shen, Dong Li, Jinxing Zhou, Zhen Qin, Bowen He, Xiaodong Han, Aixuan Li, Yuchao Dai, Lingpeng Kong, Meng Wang, Yu Qiao, and Yiran Zhong. 2023. Favdbench: Fine-grained audible video description. In *Proc. CVPR*.

Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Fine-grained audio-visual joint representations for multimodal large language models. *arXiv:2310.05863*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv:2403.05530*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. 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. Association for Computational Linguistics.

S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv:2401.01313*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment. *arXiv:2310.16944*.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2023a. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv:2310.07521*.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023b. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Jiaheng Wei, Yuanshun Yao, Jean-Francois Ton, Hongyi Guo, Andrew Estornell, and Yang Liu. 2024. Measuring and reducing llm hallucination without gold-standard answers via expertise-weighting. *arXiv:2402.10412*.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. NExT-QA: Next phase of question-answering to explaining temporal actions. In *Proc. CVPR*.

Xianjun Yang, Liangming Pan, Xuandong Zhao, Haifeng Chen, Linda Petzold, William Yang Wang, and Wei Cheng. 2023. A survey on detection of llms-generated content. *arXiv:2310.15654*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and

Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv:2311.04257*.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553, Singapore. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and mitigating object hallucination in large vision-language models. In *The Twelfth International Conference on Learning Representations*.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, and Jiantao Jiao. 2023. Starling-7b: Improving llm helpfulness & harmlessness with rlaif.

A Experimental Setup Details

We list the models involved in this paper in Table 8, and text-to-text metrics in Table 9.

Target LLMs	Modality	Evidence Models (explicit)	Evidence Models (explicit)	License
Llama-2-7B (Manakul et al., 2023b)	Text	✓	✓	llama2
Llama-2-7B-Chat (Manakul et al., 2023b)	Text	✓	✓	llama2
Mistral-7B-Instruct-v0.1 (Jiang et al., 2023)	Text	✗	✗	Apache-2.0
Mistral-7B-Instruct-v0.2 (Jiang et al., 2023)	Text	✓	✓	Apache-2.0
Vicuna-v1.5-7B(Chiang et al., 2023)	Text	✓	✓	llama2
Falcon-7B(Almazrouei et al., 2023)	Text	✗	✗	Apache-2.0
Starling-7B-alpha(Zhu et al., 2023)	Text	✓	✓	Apache-2.0
StableBeluga-7B(Mahan et al.)	Text	✓	✓	llama2
Zephyr-7b-beta(Tunstall et al., 2023)	Text	✓	✓	MIT
Mistral-7B-OpenOrca(Mukherjee et al., 2023)	Text	✓	✓	Apache-2.0
GPT-4 (OpenAI, 2023)	Text	✗	✗	N/A
LLaVA-v1.5 (Liu et al., 2023)	Vision	✓	✓	llama2
InstructBLIP (vicuna-7B) (Dai et al., 2023)	Vision	✓	✗	BSD 3-Clause
mPLUG-Owl2 (Ye et al., 2023)	Vision	✓	✓	MIT
Valley (Luo et al., 2023)	Vision	✓	✓	Apache-2.0
Video-LLaVA (Lin et al., 2023)	Vision	✓	✓	Apache-2.0
Chat-Univi (Jin et al., 2024)	Vision	✓	✓	Apache-2.0
LLaMA-VID (Li et al., 2023c)	Vision	✓	✗	Apache-2.0
LTU (Gong et al., 2024)	Audio	✓	✓	Apache-2.0
Qwen-Audio-Chat (Chu et al., 2023)	Audio	✓	✓	Tongyi Qianwen
SALMONN (Tang et al., 2024)	Audio	✓	✓	Apache-2.0
Video-LLaMA (Zhang et al., 2023)	Audio-visual	✓	✓	BSD 3-Clause
FAVOR (Sun et al., 2023)	Audio-visual	✓	✓	Apache-2.0
Gemini 1.5 Pro (Team, 2024)	Audio-visual	✓	✗	N/A

Table 8: Models and reference benchmarks for validating CrossCheckGPT.

Reference Benchmarks (Metrics)	Description
TriviaQA (Han et al., 2019) (Acc)	A realistic text-based question-answering dataset containing documents collected from Wikipedia and the web.
TruthfulQA MC1 (Lin et al., 2022) (Acc) TruthfulQA MC2 (Lin et al., 2022) (Acc)	A benchmark to measure whether a language model is truthful in generating answers to questions, spanning 38 categories.
XSum (Narayan et al., 2018) (FactKB (Feng et al., 2023))	The factual accuracy of summarization models by verifying the presence of knowledge base facts in generated summaries.
CNN-DM (See et al., 2017) (BERTP)	The CNN-DailyMail dataset is a collection of news articles and accompanying summaries measured by BERTScore-Precision.
MemoTrap (McKenzie et al., 2023) (Acc)	Assessing whether LLMs fall into memorization traps which occur when LLMs memorize specific examples in training.
FaithDial (Dziri et al., 2022) (Acc)	A benchmark for hallucination-free dialogues by editing hallucinated responses in Wizard of Wikipedia (WoW) (Dinan et al., 2019)
HaluEval-QA (Li et al., 2023a) (Acc) HaluEval-summarization (Li et al., 2023a) (Acc) HaluEval-Dialogue (Li et al., 2023a) (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 9: 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 10 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 10: 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 (Manakul et al., 2023b).

Evidence Model	Non-Factual	Non-Factual*	Factual	Document (r)
Llama 30B Max($\mathcal{H}$) (Manakul et al., 2023b)	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 11: 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 (Manakul et al., 2023b).

D Text-to-text Additional Results

We provide the version of Table 2 with all ten benchmark metrics in Table 12. 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 13 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.

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.

Metrics	System(ρ)	Document (r) w/o GPT4	Document (r) with GPT4
TriviaQA (Han et al., 2019)	23.33	-	-
TruthfulQA MC1 (Lin et al., 2022)	52.94	-	-
TruthfulQA MC2 (Lin et al., 2022)	57.14	-	-
XSum (Narayan et al., 2018)	-70.00	-	-
CNNDM (See et al., 2017)	38.33	-	-
MemoTrap (McKenzie et al., 2023)	10.88	-	-
FaithDial (Dziri et al., 2022)	-8.33	-	-
HaluEval-QA (Li et al., 2023a)	-18.33	-	-
HaluEval-Summarization (Li et al., 2023a)	48.33	-	-
HaluEval-Dialogue (Li et al., 2023a)	46.03	-	-
SelfCheckGPT (Manakul et al., 2023b)	66.46	74.06	76.08
CrossCheck-explicit	77.44	**82.28**	77.23
CrossCheck-implicit	56.71	18.33	17.29
CrossCheck-explicit weighted	**82.32**	81.78	**82.18**
CrossCheck-explicit weighted	56.81	20.21	19.16

Table 12: Full version of Table 2 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 13: 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.

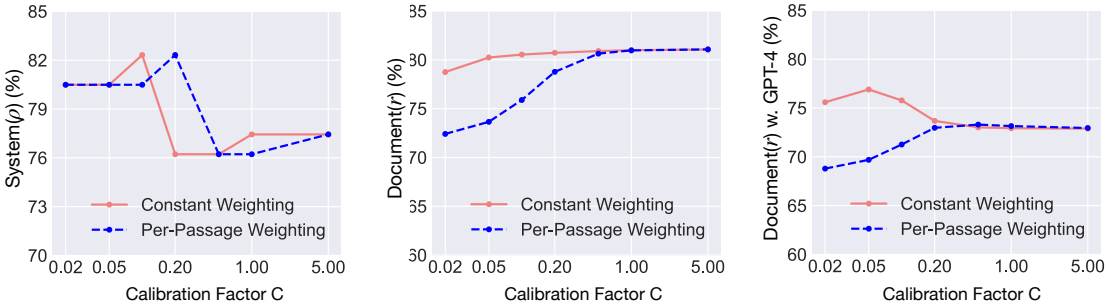

Figure 6: The variation of System(ρ) and Document(r) against calibration temperature T in Eqn. (4) for weighted CrossCheck-explicit. Constant weighting refers to applying the same weight for all documents, while per-passage weighting refers to the use of passage-specific weighting derived from SelfCheckGPT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

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

The scatter plot, similar to text-to-text ones in Fig. 4, is shown in Fig. 7 for image-to-text tasks. Additionally, we report the statistical significance of CrossCheckGPT being better than SelfCheckGPT on MHaluBench by performing the sign test at the image level shown in Table 14.

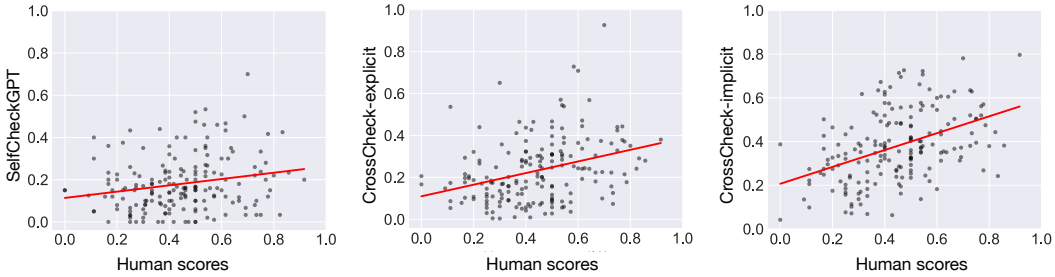

Figure 7: Scatter plot of SelfCheckGPT, CrossCheck-explicit and CrossCheck-implicit scores against human annotation for image-to-text tasks.

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.

F Statistics of AVHalluBench

We provide detailed statistics of 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 (Xiao et al., 2021)	32 (18%)	22.0	19	7	1
M3AV (Chen et al., 2024b)	27 (16%)	11.3	27	0	27
How2 (Sanabria et al., 2018)	27 (16%)	9.5	27	4	2
MUSIC-AVQA (li et al., 2022)	23 (13%)	29.0	0	23	0
VALOR32k (Chen et al., 2023)	26 (15%)	8.7	11	7	8
FAVDBench (Shen et al., 2023)	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.

G Human Annotation Details

We employed three annotators, all of whom have a university education. Initially, a pilot annotation session was conducted, followed by a discussion among the annotators to align their understanding of the instructions. The annotation process was voluntary, with no financial compensation involved. All annotators provided consent to share the data.

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 (Luo et al., 2023)	52.43	55.98	48.22
Video-LLaVA (Lin et al., 2023)	30.59	33.52	40.57
Chat-Univi (Jin et al., 2024)	29.40	32.68	41.75
LLaMA-VID (Li et al., 2023c)	38.61	39.14	**40.48**
Video-LLaMA (Zhang et al., 2023)	41.14	52.02	48.80
FAVOR (Sun et al., 2023)	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 (Gong et al., 2024)	**21.95**	-	37.44	-	18.06	-
Qwen-Audio-Chat (Chu et al., 2023)	36.57	37.08	43.66	43.41	20.21	52.20
SALMONN (Tang et al., 2024)	34.99	34.80	42.21	40.15	18.32	**48.17**
FAVOR (Sun et al., 2023)	49.62	41.51	66.69	55.41	23.26	61.01
Video-LLaMA (Zhang et al., 2023)	56.42	-	68.05	-	**17.10**	-
Gemini 1.5 Pro	25.82	**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.

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.

J Additional Results on Robustness to Manipulation

In Figure 5 the main text, we show that CrossCheckGPT is robust against generation temperature, while in Figure 8, we show that SelfCheckGPT is more sensitive to the temperature.

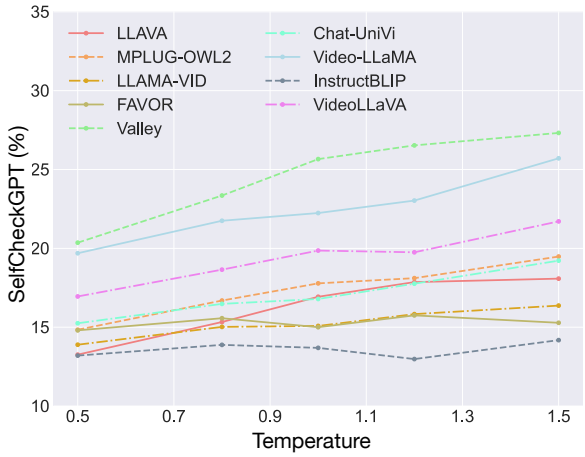

Figure 8: Variation of SelfCheckGPT scores against temperature during generation. This plot is complementary to Fig. 5 (CrossCheckGPT) in the main text.

K Case Studies for Hallucination with Audio-Visual Inputs

In addition to the piano example shown in Fig. 11 that has been mentioned in the main text, we show here two additional examples in Fig. 10 and Fig. 9 where audio-visual inputs influence the hallucination compared to using audio or visual inputs alone.

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 9: 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.

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 10: 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 11: 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”.

L Additional Analysis on Hallucination Distribution

This section further validates the fundamental hypothesis using quantitative and qualitative examples. First, the pairwise CrossCheckGPT scores among different evidence models are shown in Fig. 12. This is to demonstrate that the models generate different hallucinations, since those CrossCheckGPT scores are different, and are all different from SelfCheckGPT scores.

Example responses from 3 different evidence models on the image-to-text task are shown in Fig. 13. This is to demonstrate that different models in general have different hallucinations while the true facts are much more consistent across models.

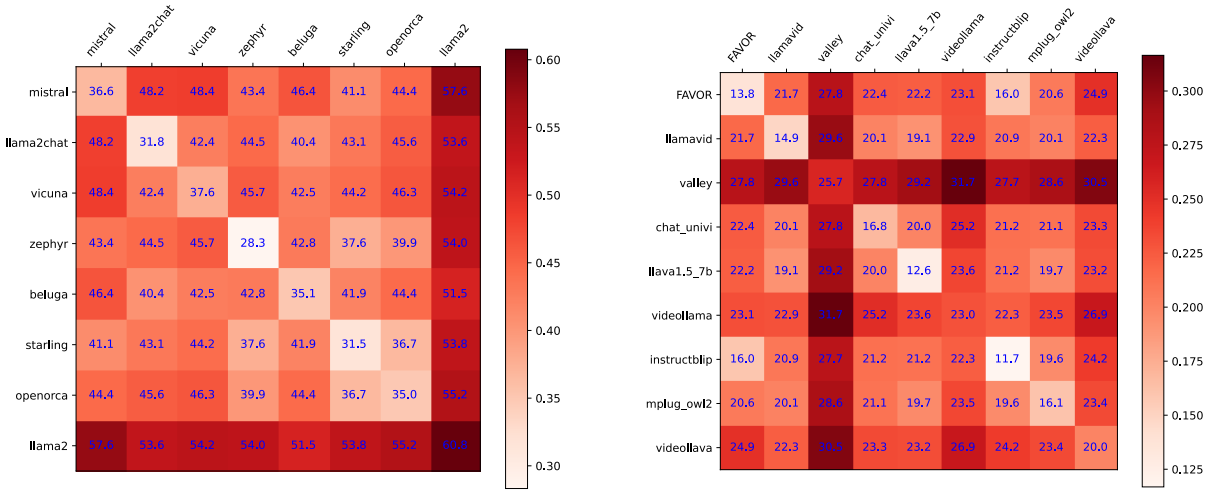

Figure 12: Heatmap of pairwise content difference measured by CrossCheck-explicit for text-only (left) and visual-LLMs (right) for all evidence models. Diagonal values are SelfCheckGPT scores for each model and off-diagonal values are CrossCheckGPT scores between each pair of models.

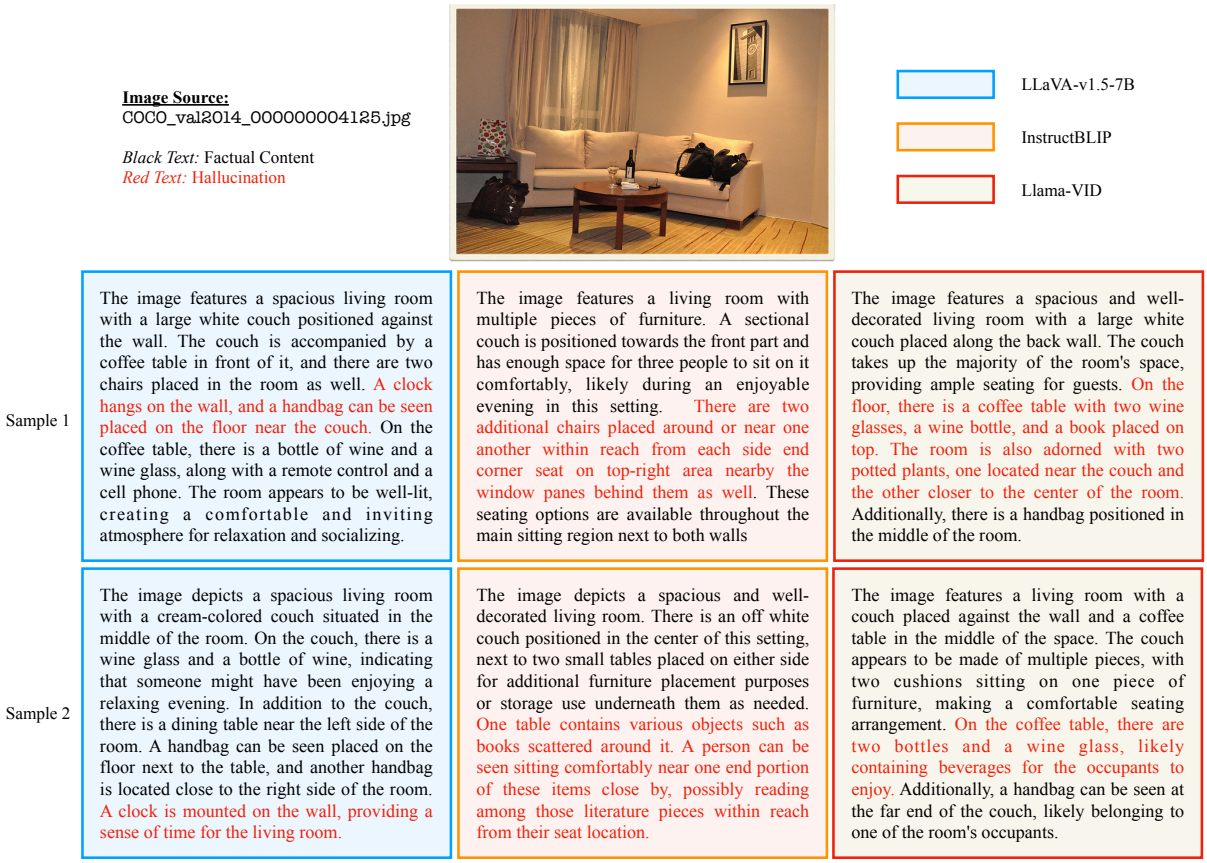

Figure 13: Sample evidence passages (2 per evidence model) drawn from LLaVA-v1.5-7B, InstructBLIP and Llama-VID to show the difference in hallucinated content. This shows that different models hallucinate about different contents and the factual content is much more consistent. For example, the hallucination of “a person" by InstructBLIP is not found in any other samples. The hallucination of “two wine bottles” from Llama-VID is not generated by any other models either. Moreover, the "two wine bottles" hallucination is found in both samples generated by the same model, showing superior robustness of CrossCheckGPT compared to SelfCheckGPT.