$\mathcal{O}mni\mathcal{H}allu$: Unified Hallucination Detection for Cross-Modal Comprehension and Generation in Multimodal Large Language Models

Anonymous ACL submission

Abstract

While recent Multimodal Large Language Models (MLLMs) have made exciting strides in various tasks and scenarios, they suffer from a significant issue of hallucinations, where generated outputs contradict or misrepresent input semantics. Existing research often focuses on either comprehension or generation tasks within specific modalities, which restricts the generalizability of hallucination studies in MLLMs. To bridge this gap, we introduce **OmniHallu**, a unified hallucination detection and evaluation framework for cross-modal comprehension and generation in MLLMs. We present a unified benchmark, **OmniHallu-Bench**, for evaluating both comprehension and generation tasks across modalities, covering text-to-image (T2I), text-to-video (T2V), text-to-audio (T2A), as well as image-to-text (I2T), video-to-text (V2T), and audio-to-text (A2T) processes. Additionally, we propose a novel multi-agent hallucination detection architecture that automatically decomposes and verifies claims, facilitating structured hallucination assessment. Extensive evaluations and analysis demonstrate the effectiveness of our methods, establishing a robust foundation for hallucination detection in MLLMs. This work contributes toward building more reliable and interpretable multimodal AI systems. We will release our source code and data in the camera-ready version.

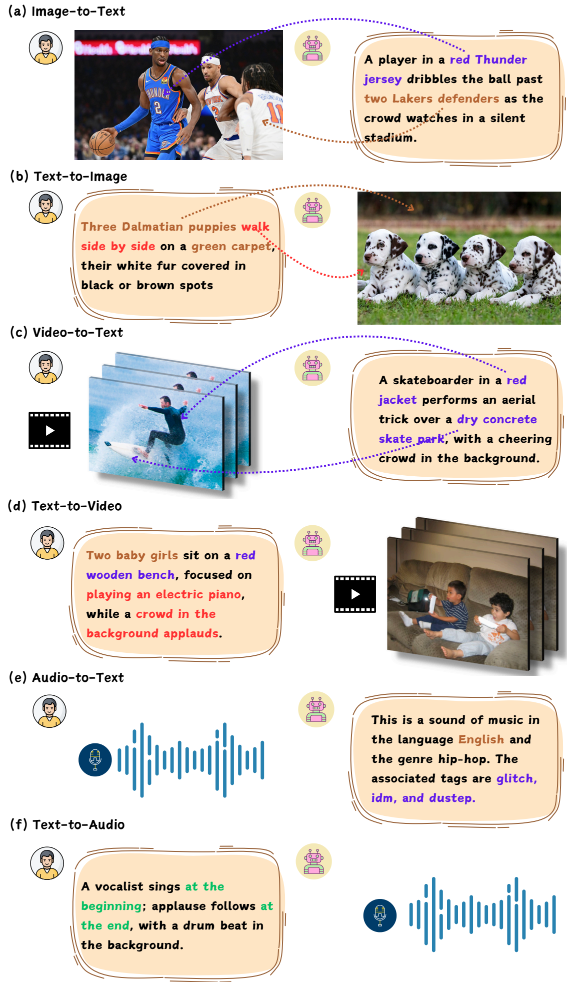

Figure 1: MLLMs can produce hallucinations in both comprehension and generation processes across modalities, encompassing different types such as object, attribute, relation, and event hallucinations.

1 Introduction

In recent years, MLLMs (Huang et al., 2023b; Weng et al., 2024; Li et al., 2024b; Chen et al., 2025) have made remarkable progress across various tasks, spanning natural language processing, computer vision, audio processing, and multimodal learning. These advancements have enabled MLLMs to surpass traditional models in multiple domains, bringing them closer to achieving human-level intelligence (Wang et al., 2024a; Fei et al., 2024; Luo et al., 2024). However, a critical challenge that remains is hallucination (Bai et al., 2024b; Huang et al., 2024), where generated outputs deviate from or contradict factual information or user instructions. Existing works (Manakul et al., 2023; Li et al., 2023b; Wang et al., 2024c) often focus on hallucination detection within a single modality or specific tasks, limiting their applicability to general multimodal settings. A unified hallucination detection framework that generalizes across cross-modal comprehension and generation

Benchmark	Function	Granularity	#Instances	Task	#Modalities	Rationale
QAGS (Wang et al., 2020a)	Check	Summary	474	T2T	1	✘
HaluEval (Li et al., 2023a)	Detection	Response	30,000	T2T	1	✘
POPE (Li et al., 2023b)	Evaluation	Response	500	I2T	2	✘
AMBER (Wang et al., 2024b)	Evaluation	Response	1,004	I2T	2	✘
FactVC (Liu and Wan, 2023)	Evaluation	Response	1,800	V2T	2	✘
AHLALM (Nishimura et al., 2024)	Evaluation	Response	1,000	A2T	2	✘
SoraDetector (Chu et al., 2024)	Evaluation	Response	50	T2V	2	✘
MHaluBench (Chen et al., 2024a)	Detection	Res.,Seg.,Cla.	420	T2I, I2T	2	✔
OmniHallu-Bench	Detection	Res.,Seg.,Cla.	5,000	T2I, T2V, T2A, I2T, V2T, A2T	4	✔

Table 1: Comparison of existing benchmarks for fact-checking, hallucination evaluation, and detection.

tasks remains an open challenge.

From a modality perspective, hallucinations in MLLMs extend beyond text-based tasks to image, video, and audio modalities. The cross-modal nature of MLLMs often exacerbates hallucinations due to increased reasoning complexity. For example, an MLLM may miscount objects in an image, fail to capture causal relationships in a video, or misidentify sound sources in an audio clip. These issues arise from multiple factors, including inadequate multimodal feature extraction, contextual ambiguity, and misalignment between multimodal representations and language understanding. **From a task perspective**, hallucinations manifest in both comprehension and generation settings. I2T, V2T, and A2T tasks evaluate a model's ability to interpret and extract meaningful information from multimodal inputs, while T2I, T2V, and T2A tasks assess its capacity to generate coherent and semantically accurate outputs conditioned on textual prompts. Despite these distinctions, hallucinations in different tasks and modalities share common underlying causes, including insufficient perception and reasoning capabilities, which lead to errors such as object, attribute, relation, and event hallucinations. Given these shared characteristics, establishing a unified hallucination detection framework is both practical and necessary.

To address these challenges, we introduce **OmniHallu**, a unified hallucination detection framework for cross-modal comprehension and generation in MLLMs. Our approach enables a standardized detection of hallucinations across common multimodal tasks, covering T2I, T2V, T2A, as well as I2T, V2T, and A2T processes. Additionally, we propose a novel multi-agent hallucination detection architecture that systematically decomposes and verifies claims, ensuring a structured and interpretable rationale for hallucination assessment. Beyond merely detecting hallucinations, our framework provides fine-grained analysis and insights into their causes.

To benchmark our framework, we introduce **OmniHallu-Bench**, a large dataset covering both comprehension and generation tasks across all four modalities. We develop a hybrid dataset construction pipeline that integrates high-quality samples from existing datasets with state-of-the-art model-generated outputs, all collected and generated data undergo rigorous human verification to ensure quality and reliability. Our approach builds upon a multi-agent architecture, which has demonstrated flexibility, modularity, and robustness in complex reasoning tasks. We integrate Large Language Models (LLMs) such as GPT-4o (Hurst et al., 2024) with domain-specific expert models to systematically detect hallucinations in MLLM outputs. The LLM serves as a central controller, orchestrating expert models for specific verification tasks. The outputs from these expert models are further processed by a reasoning model (e.g., OpenAI-o1 (Jaech et al., 2024), DeepSeek-R1 (Guo et al., 2025)), which consolidates verification results and generates explainable rationales for hallucination assessment.

We conduct extensive evaluations of our multi-agent hallucination detection architecture on OmniHallu-Bench. The results demonstrate strong performance in hallucination detection and verification, validating the effectiveness of our approach and establishing a reliable baseline for future research on hallucinations in MLLMs. In summary, our work makes the following key contributions:

- We introduce **OmniHallu**, a unified hallucination detection framework for cross-modal comprehension and generation in MLLMs.
- We present **OmniHallu-Bench**, a high-quality benchmark for detecting hallucinations in both comprehension and generation tasks across modalities.
- We propose a **multi-agent hallucination de-**

tection architecture that leverages LLMs for planning, task-specific models for verification, and reasoning models for inference, enabling an automated and systematic hallucination detection framework.

2 Related Work

Hallucinations in Comprehension Tasks. Text, as a structured and explicit modality, provides well-defined semantics that enable efficient encoding into a learned representation space, facilitating comprehension and reasoning. However, the complexities of encoding and learning multimodal information pose significant challenges for non-text modalities. Large Vision-Language Models (LVLMs) often misinterpreting or fabricating objects, attributes, and spatial relationships in images. Studies such as (Zhang et al., 2023; Tjio et al., 2021) reveal that LVLMs still exhibit considerable hallucinations in fundamental tasks like object recognition and attribute alignment, limiting their reliability in real-world applications. Understanding video content presents an even greater challenge. Large Video Models (LVMs) often misinterpret temporal and spatial relationships in video sequences, leading to incorrect scene comprehension. These models may also struggle with distinguishing visually similar but semantically distinct frames, resulting in misattributed actions and events (Iashin and Rahtu, 2020; Suin and Rajagopalan, 2020). Additionally, hallucinations in causal reasoning, such as incorrect cause-effect predictions in video narratives, remain a persistent challenge. Similarly, Large Audio Models (LAMs) have gained prominence in speech recognition, music analysis, and audio synthesis. However, they remain prone to hallucinations, including misinterpretation of background sounds, inaccuracies in audio summaries, and difficulty capturing fine-grained audio features like pitch and timbre (Shen et al., 2023), leading to errors in comprehension and transcription tasks.

Hallucinations in Generation Tasks. Hallucinations are not limited to comprehension tasks but are equally pervasive in generation tasks across modalities. In text-to-image generation, research such as (Liu et al., 2024b; Dai et al., 2023) indicates that LVLMs frequently fail to align with user prompts, leading to errors in object positioning, attribute consistency, and logical coherence. Fine-grained inconsistencies, such as incorrect depictions of textures or unrealistic object interactions, remain persistent issues. In text-to-video generation, the complexity increases as models must generate temporally coherent frames that maintain consistency over time. Studies like (Chu et al., 2024; Rawte et al., 2024) show that hallucinations in video generation are particularly pronounced, as models often struggle with motion continuity, scene composition, and maintaining contextual relevance across multiple frames. Similarly, text-to-audio generation suffers from hallucination-related issues. Recent studies (Han et al., 2021; Shen et al., 2023; Ye et al., 2021) have demonstrated that LAMs can introduce non-existent sound effects, distort speech patterns, or fail to maintain consistent tonal qualities. These hallucinations are particularly problematic in applications like automatic music composition or speech synthesis, where accuracy in timing and acoustic properties is crucial.

Detection of Hallucinations. Given the widespread hallucination issues in MLLMs, extensive research has been conducted on their detection. However, most existing methods focus on specific modalities or hallucination types, limiting their generalizability across multimodal tasks. For image-based hallucinations, early studies primarily addressed object hallucination, where models generate descriptions containing non-existent or incorrect objects. Beyond object-level hallucinations, (Liu et al., 2024b) introduced IVL-Hallu, which categorizes hallucinations into attribute, object, multimodal conflicting, and counter-common-sense types. For video-based hallucination detection, research has focused on ensuring factual consistency in comprehension and generation. (Liu and Wan, 2023) introduced FactVC, a factuality metric improving hallucination assessment in video captions. For audio-based hallucination detection, models often over-rely on visual modality during pre-training, leading to errors in generated descriptions. (Nishimura et al., 2024) categorized audio hallucinations into three types, emphasizing the challenge of visually-induced hallucinations in audio models.

3 Preliminaries

Unified Formulation of Multimodal Hallucinations. Let $\mathcal{T}$, $\mathcal{I}$, $\mathcal{V}$, and $\mathcal{A}$ denote the sets of textual, image, video, and audio data, respectively. We consider a MLLM as a function

$$f_\theta : \mathbf{X} \rightarrow \mathbf{Y},$$

where $\mathbf{X} \in \{\mathcal{T}, \mathcal{I}, \mathcal{V}, \mathcal{A}\}$ is the input, $\mathbf{Y} \in \{\mathcal{T}, \mathcal{I}, \mathcal{V}, \mathcal{A}\}$ is the output, and θ denotes the parameters of the MLLM. We focus on two broad categories of tasks: comprehension tasks, where the model processes non-textual input $\mathbf{x} \in \{\mathcal{I}, \mathcal{V}, \mathcal{A}\}$ and produces a textual output $\hat{y} \in \mathcal{T}$, and generation tasks, where the model takes a textual input $\mathbf{x} \in \mathcal{T}$, such as a prompt or instructions, and generates an output in another modality $\hat{y} \in \{\mathcal{I}, \mathcal{V}, \mathcal{A}\}$.

Definition of Multimodal Hallucination. We say that an MLLM's output $\hat{y}$ is hallucinated if it introduces or claims content that contradicts the input $\mathbf{x}$. Formally, let $\mathcal{G}$ be the set of *ground-truth* elements derived from the input $\mathbf{x}$. A generated output $\hat{y}$ is considered hallucinated if

$$\text{Hallucinate}(\hat{y} \mid \mathbf{x}) = \begin{cases} 1 & \text{if } \exists\, \phi(\hat{y}) \notin \mathcal{G} \\ 0 & \text{otherwise,} \end{cases}$$

where $\phi(\hat{y})$ denotes any semantic claim extractable from $\hat{y}$, and $\text{Hallucinate}(\cdot)$ is an indicator function.

Unified Hallucination Types across Modalities. Although hallucinations manifest differently across text, image, video, and audio modalities, they can be categorized into four key types. Object hallucinations occur when non-existent entities are introduced, such as describing a "car" in an image where none exists. Attribute hallucinations involve misrepresentations of properties like color, size, or timbre, such as calling a blue hat "red" or misidentifying a female voice as "male." Relation hallucinations arise when the relationships between entities are incorrectly stated, for example, describing "a dog chasing a cat" when the roles are reversed or the interaction never occurred. Event hallucinations misrepresent event-level details, such as describing a person as "falling" in a video when they are actually sitting down, or claiming a ball was thrown before it was even picked up. These hallucination types are prevalent across different modalities and pose distinct challenges for MLLMs in ensuring factual consistency.

Unified Detection of Hallucinations. To systematically detect hallucinations in MLLMs, we adopt a multi-agent framework that integrates claim decomposition, expert verification, and reasoning-based assessment. Given a model-generated output, our method first decomposes it into atomic claims, ensuring that each claim is a discrete, verifiable statement. These claims are then processed by a set of expert agents specialized in different modalities, leveraging state-of-the-art models for cross-modal consistency checking. For comprehension tasks, these agents assess whether each claim aligns with the given input, while for generation tasks, verification is performed by comparing claims against the fundamental concepts inferred from the textual prompt. Finally, a reasoning agent consolidates the individual verifications to derive a robust hallucination classification.

4 OmniHallu-Bench: A Comprehensive Hallucination Detection Benchmark

To systematically evaluate multimodal hallucinations, we construct a benchmark dataset covering image, video, and audio captioning, as well as text-to-image, text-to-video, and text-to-audio generation tasks. Our dataset consists of 5,000 samples, ensuring a balanced distribution across different modalities and hallucination types. Specifically, captioning tasks account for 60% of the dataset, while generation tasks constitute the remaining 40%. The proportion of image, video, and audio samples is maintained at 5:3:2, ensuring comprehensive coverage of all modalities. We categorize hallucinations into four distinct types: object, attribute, relation, and event hallucinations, with respective proportions of 35%, 25%, 15%, and 25%. This distribution reflects the common hallucination patterns observed in MLLMs and enables a fine-grained evaluation of their capabilities. Our dataset integrates high-quality samples from established datasets alongside current leading model-generated outputs.

Image-to-Text Comprehension. For image captioning, we draw samples from COCO Caption (Chen et al., 2024b), Nocaps (Agrawal et al., 2019), and Flickr30k (Plummer et al., 2016). These datasets contain human-annotated captions, offering high-quality references for evaluating hallucinations. We also leverage InternVL2.5-78B (Chen et al., 2024b), Qwen2.5-VL-72B (Yang et al., 2024a), GPT-4o, and Gemini-1.5-Pro (Team et al., 2024) to generate outputs, all of which exhibit strong captioning abilities yet remain susceptible to hallucinations.

Video-to-Text Comprehension. For video captioning, we sample data from MSVD (Chen et al., 2022), MSRVTT (Xu et al., 2016), and VATEX (Wang et al., 2020b), which provide ground-truth textual descriptions of diverse video content.

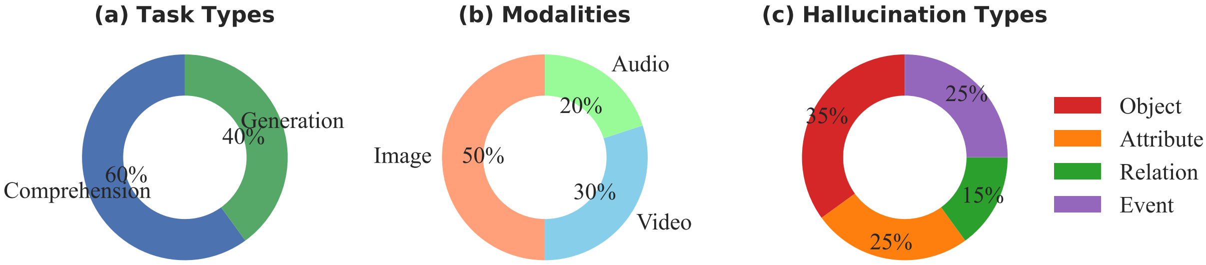

Figure 2: Main statistics of our OmniHallu-Bench dataset.

We also use InternVL2.5-78B, Qwen2.5-VL-72B, VideoLLaMA3 (Zhang et al., 2025), and LLaVA-OneVision (Li et al., 2024a) to generate outputs, which are representative of current leading LVMs but still exhibit a notable presence of hallucinations.

Audio-to-Text Comprehension. For audio captioning, we collect samples from AudioCaps (Kim et al., 2019), ClothoV2 (Drossos et al., 2019), and AudioSetCaps (Bai et al., 2024a), which provide high-quality human-written descriptions of diverse soundscapes. We further include generative outputs from Qwen2-Audio-7B-Instruct, GAMA (Ghosh et al., 2024), Pengi (Deshmukh et al., 2024), and SALMONN (Sun et al., 2024), capturing hallucinations related to semantic misinterpretation in LAMs.

Text-to-Image Generation. For text-to-image tasks, we source initial prompts from T2I-CompBench++ (Huang et al., 2025) and HRS-Bench (Bakr et al., 2023), two prominent benchmarks designed for evaluating text-to-image synthesis quality. These prompts are augmented using ChatGPT to introduce various hallucination types, enhancing the complexity of generated content. The refined prompts are then used to generate images via DALL-E 3, Stable Diffusion 3.5 Large, and Midjourney v6.

Text-to-Video Generation. For text-to-video tasks, we utilize prompts from T2V-CompBench (Sun et al., 2025) and FETV (Liu et al., 2023), which contain structured test cases for evaluating video synthesis models. We employ ModelScope, Open-Sora 1.2 (Zheng et al., 2024), and CogVideoX-5B (Yang et al., 2024b) to generate corresponding videos.

Text-to-Audio Generation. For text-to-audio generation, we leverage prompts from WavText5Ks (Deshmukh et al., 2022), FSD50K (Fonseca et al., 2022), and SoundDescs (Koepke et al., 2023), which provide rich textual descriptions of various sound events. We generate corresponding audio samples using Make-an-Audio (Huang et al., 2023a), AudioGPT (Huang et al., 2023b), and AudioLCM (Liu et al., 2024a).

To ensure data integrity and high quality, all collected and generated samples undergo rigorous human verification. We employ a structured atomic claim decomposition process to break down generated outputs into verifiable atomic claims, enabling precise hallucination assessment. To enhance the accuracy and consistency of claim extraction, we adopt a two-stage method combining Chain-of-thought (CoT) (Wei et al., 2022) prompting and self-reflection verification. CoT prompting sequencely decomposes responses into atomic claims, while self-reflection ensures the extracted claims preserve the original semantics without alteration or omission. For comprehension tasks, claims are extracted from generated captions and cross-checked against the original multimodal input. For generation tasks, fundamental intent concepts are derived from user prompts and used as reference claims for evaluation. Each sample is manually reviewed by three annotators, who classify extracted claims as hallucinatory or non-hallucinatory. A response is labeled hallucinatory if any of its claims contain hallucinations. To ensure annotation consistency, we conduct strict human inspection and cross-validation.

5 Multi-Agent Framework for Hallucination Detection

To systematically detect and verify hallucinations in MLLMs, we propose a multi-agent framework that integrates atomic claim decomposition, modality-aware multi-agent execution, and reasoning-based verification. This framework enables structured and explainable hallucination assessment across diverse multimodal tasks, ensuring both robustness and interpretability.

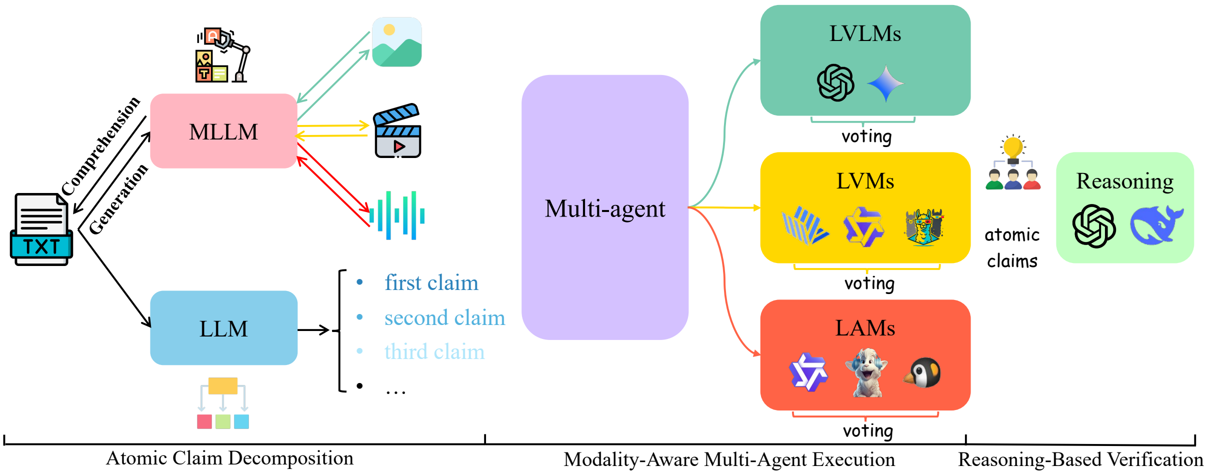

Figure 3: Main statistics of our OmniHallu-Bench dataset.

Atomic Claim Decomposition. Hallucination detection requires an explicit breakdown of generated content into verifiable components. To achieve this, we leverage GPT-4o's advanced natural language processing and instruction-following capabilities to decompose both comprehension tasks' captions and generation tasks' prompts into a set of atomic claims. Each sample $(y, \{c_1, \cdots, c_{n_y}\})$ consists of a piece of text y and a corresponding set of atomic claims $\{c_1, \cdots, c_{n_y}\}$, where each claim provides a semantically discrete, verifiable statement extracted from the original output. These claims are designed to comprehensively represent all information contained in y while ensuring that no additional, unverifiable content is introduced. Furthermore, each claim must be grammatically independent and comprehensible in isolation, such that any reference to entities or pronouns is fully resolved, preventing ambiguity in subsequent verification steps. This decomposition process establishes a structured foundation for hallucination detection, enabling precise comparisons between generated content and reference ground truth.

Modality-Aware Multi-Agent Execution. Different hallucination types and modalities require specialized verification methodologies. Our framework dynamically selects expert models and tools based on the specific task and hallucination category, ensuring accurate and adaptable hallucination detection. For I2T and T2I generation tasks, object hallucinations are verified using Grounding DINO 1.5 Pro (Ren et al., 2024), an advanced open-set object detection model that extracts accurate visual entity information as the ground truth reference. Attribute, relation, and event hallucinations require whole semantic understanding beyond direct object detection. To address these hallucination types, we utilize multiple LVLMs, including Qwen2.5-VL-72B, InternVL2.5-78B, and GPT-4o. These models serve as expert agents, contributing to a more robust verification process.

For V2T and T2V generation tasks, the complexity of video data, coupled with the limitations of current LVLMs, necessitates a hybrid approach. We integrate methodologies inspired by DoraemonGPT (Yang et al., 2024c), each atomic claim requiring verification is reformulated into a targeted question-answering query using GPT-4o, enabling the extraction of specific, modality-aligned insights. These extracted insights provide the foundation for assessing hallucinations in video-based tasks.

For A2T and T2A generation tasks, hallucination detection is particularly challenging due to the lack of robust expert tools for fine-grained auditory understanding. Since external knowledge sources cannot directly verify complex auditory information, we employ multiple LAMs, including Qwen2-Audio-7B-Instruct, GAMA, Pengi, and SALMONN, to analyze and interpret audio content. This ensemble approach strengthens the reliability of hallucination verification in the audio domain.

Reasoning-Based Verification. Once the verification results are obtained, they are consolidated through reasoning models that synthesize the available evidence into a final hallucination determination. The atomic claim decomposition results and multi-agent verifications are processed using OpenAI-o1 and DeepSeek-R1, which integrate multiple verification sources into a structured reasoning pipeline. This process aggregates and analyzes verification outputs, identifies inconsistencies, and derives a balanced final decision. To

Method		Hallucinatory			Non-Hallucinatory			Average			
		P	R	F1	P	R	F1	Acc.	P	R	Mac.F1
Image-to-Text											
Gemini-1.5-Pro	Self-Check	82.75	64.12	72.20	66.12	82.31	73.22	72.48	73.95	73.12	72.58
	UNIHD	83.94	68.21	75.41	69.92	84.45	76.20	75.92	76.48	75.88	75.42
GPT-4o	Self-Check	79.32	73.92	76.52	74.31	80.54	77.30	76.21	76.92	76.54	76.18
	UNIHD	81.02	77.45	79.20	77.23	79.92	78.42	78.12	78.24	78.04	77.94
Ours		**84.65**	**81.34**	**82.96**	**83.15**	**82.72**	**82.93**	**82.58**	**82.46**	**82.38**	**82.12**
Text-to-Image											
Gemini-1.5-Pro	Self-Check	81.12	59.82	68.90	63.41	80.74	71.00	70.98	71.54	70.68	70.24
	UNIHD	82.64	64.52	72.48	65.92	82.32	73.01	73.84	73.12	72.95	72.70
GPT-4o	Self-Check	78.22	71.94	74.94	72.63	78.32	75.24	74.52	74.94	74.68	74.32
	UNIHD	79.88	76.54	78.18	78.65	78.80	78.72	78.64	78.72	78.54	78.42
Ours		**83.21**	**80.92**	**82.04**	**81.74**	**81.42**	**81.60**	**81.40**	**81.18**	**81.10**	**80.92**
Video-to-Text											
InternVL2.5-78B	Self-Check	74.12	60.82	66.92	65.34	80.24	72.34	69.48	69.72	69.24	69.18
Qwen2.5-VL-72B	Self-Check	76.58	65.12	70.28	68.42	81.32	74.92	72.40	72.58	72.18	71.92
Ours		**78.92**	**75.42**	**77.12**	**77.38**	**76.92**	**77.10**	**77.02**	**76.88**	**76.74**	**76.58**
Text-to-Video											
InternVL2.5-78B	Self-Check	72.94	58.12	64.78	62.72	78.34	70.32	67.42	67.68	67.24	67.02
Qwen2.5-VL-72B	Self-Check	74.75	63.42	68.42	67.10	79.42	72.24	70.12	70.38	70.04	69.92
Ours		**77.62**	**74.12**	**75.82**	**75.92**	**74.92**	**75.41**	**75.22**	**75.10**	**74.94**	**74.78**
Audio-to-Text											
GAMA	Self-Check	71.34	56.72	63.40	62.94	78.02	69.34	67.58	67.84	67.42	67.12
Qwen2-Audio-7B-Instruct	Self-Check	73.48	59.24	66.00	64.82	79.28	71.32	69.74	69.92	69.58	69.40
Ours		**76.38**	**72.12**	**74.18**	**74.64**	**73.42**	**74.02**	**73.82**	**73.64**	**73.42**	**73.18**
Text-to-Audio											
GAMA	Self-Check	70.15	55.48	62.18	61.72	77.42	68.72	66.98	67.12	66.82	66.58
Qwen2-Audio-7B-Instruct	Self-Check	72.48	58.34	65.00	63.92	78.36	70.24	68.94	69.12	68.92	68.74
Ours		**75.48**	**71.52**	**73.38**	**73.74**	**72.92**	**73.31**	**73.12**	**72.98**	**72.82**	**72.58**

Table 2: Multimodal hallucination detection results across six tasks.

enhance verification reliability, we extract the common intersection of information validated by multiple expert models. By identifying consistent verification details across models, we ensure that only the most reliable and agreed-upon information serves as the basis for hallucination detection. This approach mitigates individual model biases and enhances verification robustness. The final hallucination determination is based on these consistently validated outputs, ensuring a robust and explainable verification process. Additionally, the reasoning model generates a detailed rationale for each verification outcome, leveraging its intermediate reasoning capabilities to ensure transparency and interpretability.

6 Experiments

6.1 Settings

Baselines. We follow the baseline settings established in UNIHD (Chen et al., 2024a) and adopt Self-Check (Miao et al., 2023) based on CoT prompting as our baseline. This method evaluates the intrinsic capability of the underlying MLLM to detect hallucinations without external tools. However, since UNIHD is only applicable to image-based tasks, it cannot be directly extended to other modalities. To enable a comprehensive comparison, we select two leading MLLMs for each modality beyond image-based tasks. For video-related tasks, we compare against InternVL2.5-78B and Qwen2.5-VL-72B. For audio-related tasks, we use Qwen2-Audio-7B-Instruct and GAMA. This expanded baseline selection ensures that our evaluation remains consistent and modality-adaptive, allowing a meaningful performance comparison of our multi-agent framework across different tasks.

Evaluations. We follow the evaluation settings of UNIHD, computing precision, recall, and Micro-F1 scores separately for both hallucinatory and non-hallucinatory categories at the claim level to ensure fine-grained hallucination detection analysis. Additionally, we report accuracy and macro-averaged F1 scores, maintaining consistency with prior work.

6.2 Results and Analysis.

Overall Performance. Our method consistently outperforms all baselines across six multimodal hallucination detection tasks, demonstrating its effectiveness in both comprehension and generation settings. As shown in Table 2, our multi-agent framework achieves the highest scores in all tasks, consistently surpassing self-check baselines and existing approaches. These results highlight the advantage of integrating structured claim verification with multi-agent collaboration, enabling precise hallucination detection across diverse modalities.

Performance Comparison Across Modalities. The performance comparison across modalities, as shown in Table 2, reveals a trend: image-based tasks achieve the highest scores, followed by video-based tasks, while audio-based tasks perform the worst. This pattern aligns with the current capabilities of MLLMs, where static image understanding is the most mature. In contrast, video comprehension introduces additional challenges due to the need for temporal reasoning, leading to slightly lower performance. The most pronounced limitations are observed in audio-based tasks, where hallucination detection remains challenging due to the inherent ambiguities in sound interpretation and the weaker alignment.

Performance on Fine-grained Hallucination Types. Our fine-grained analysis reveals a clear difficulty hierarchy among hallucination types, as shown in Figure 4. Object hallucinations are the easiest to detect. Attribute hallucinations are more challenging, requiring fine-grained semantic understanding. Event hallucinations introduce additional complexity, as they involve temporal information. Relation hallucinations are the most difficult, relying on complex spatial, temporal, and causal reasoning. Despite these challenges, our method consistently outperforms the strongest baseline across all categories, with the largest improvements in relation hallucinations.

Ablation Study. To assess the contributions of key components in our multi-agent framework, we conduct an ablation study. As shown in Table 3, removing Atomic Claim Decomposition (ACD)

Task	**Full Multi-Agent**	**w/o ACD**	**w/o MV**
Image-to-Text	82.12	75.02	77.64
Text-to-Image	80.92	74.31	76.55
Video-to-Text	76.58	70.10	72.34
Text-to-Video	74.78	68.42	70.71
Audio-to-Text	73.18	67.05	69.24
Text-to-Audio	72.58	66.81	68.92

Table 3: Ablation study on our multi-agent framework.

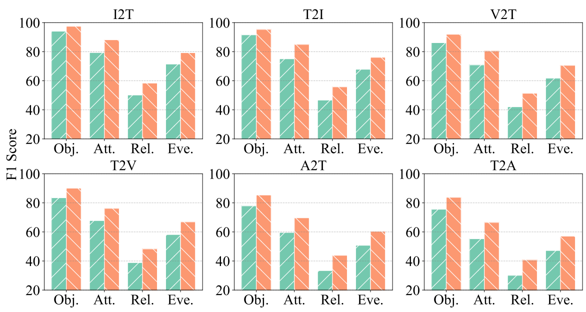

Figure 4: Comparison of detection performance across hallucination types. 'Obj.', 'Att.', 'Rel.', 'Eve.' denote object, attribute, relation and event respectively.

leads to a large performance drop, with F1 scores decreasing by 7.1%–8.5% across tasks. This underscores the importance of structured decomposition for accurate hallucination detection. Without ACD, the model struggles to isolate hallucinations in long-form responses, increasing false negatives and reducing precision. Removing Majority Voting (MV) results in a notable F1 decline of 5.0%–5.5%, showing the benefits of aggregating multiple expert results instead of relying on a single model.

7 Conclusion

We introduce **OmniHallu**, a unified hallucination detection framework designed to address hallucinations across comprehension and generation tasks in multiple modalities of MLLMs. To support comprehensive evaluation, we construct **OmniHallu-Bench**, a large-scale, high-quality benchmark covering diverse multimodal scenarios. We design a novel multi-agent hallucination detection architecture, which systematically decomposes outputs into atomic claims, verifies them using expert models, and consolidates results through structured reasoning. Extensive evaluations demonstrate that our method significantly improves hallucination detection across all settings. By providing a robust and interpretable hallucination detection framework, this work lays a solid foundation for advancing the development of more reliable MLLMs.

Potential Limitations

Limited Hallucination Taxonomy. While our framework covers commonly observed hallucination types, including object, attribute, relation, and event hallucinations, different modalities may exhibit additional hallucination patterns that remain unexplored. Expanding the taxonomy to incorporate more modality-specific errors, such as temporal inconsistencies in video or prosodic misinterpretations in audio, is an important direction for future research.

Coverage of Multimodal Tasks. OmniHallu primarily focuses on core multimodal tasks involving comprehension and generation. While this covers a broad range of applications, more specialized tasks such as video question answering (VQA) or long-form video understanding may require additional adaptations to our framework. Future work should explore extending OmniHallu to handle these complex scenarios while preserving its interpretability and robustness.

Dependence on Existing MLLMs. Hallucination detection in video and audio modalities requires global semantic understanding, yet the absence of advanced expert models or tools limits fine-grained perception capabilities. As a result, our multi-agent framework primarily relies on MLLMs, complemented by self-reflection and majority voting to enhance verification accuracy. However, these strategies remain inadequate, underscoring the gap between current MLLMs and human-level comprehension. Future advancements in domain-specific expert models and cross-modal verification techniques are crucial for addressing this limitation.

Ethics Statement

The datasets used in our study are sourced from publicly available or ethically curated materials, ensuring compliance with data usage policies. Additionally, our dataset includes AI-generated content for evaluation purposes, which is transparently documented and rigorously verified through detailed human review to ensure accuracy and reliability. We acknowledge the broader implications of hallucination detection in AI systems and advocate for responsible model development that prioritizes reliability, fairness, and interpretability.

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE.

Jisheng Bai, Haohe Liu, Mou Wang, Dongyuan Shi, Wenwu Wang, Mark D. Plumbley, Woon-Seng Gan, and Jianfeng Chen. 2024a. Audiosetcaps: An enriched audio-caption dataset using automated generation pipeline with large audio and language models.

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

Eslam Mohamed Bakr, Pengzhan Sun, Xiaoqian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. 2023. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19984–19996.

Haoran Chen, Jianmin Li, Simone Frintrop, and Xiaolin Hu. 2022. The msr-video to text dataset with clean annotations. *Computer Vision and Image Understanding*, 225:103581.

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.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zhixuan Chu, Lei Zhang, Yichen Sun, Siqiao Xue, Zhibo Wang, Zhan Qin, and Kui Ren. 2024. Sora detector: A unified hallucination detection for large text-to-video models.

Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. 2023. Plausible may not be faithful: Probing object hallucination in vision-language pre-training.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2024. Pengi: An audio language model for audio tasks.

Soham Deshmukh, Benjamin Elizalde, and Huaming Wang. 2022. Audio retrieval with wavtext5k and clap training.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2019. Clotho: An audio captioning dataset.

Hao Fei, Yuan Yao, Zhuosheng Zhang, Fuxiao Liu, Ao Zhang, and Tat-Seng Chua. 2024. From multimodal llm to human-level ai: Modality, instruction, reasoning, efficiency and beyond. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*, pages 1–8.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2022. Fsd50k: An open dataset of human-labeled sound events.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Qichen Han, Weiqiang Yuan, Dong Liu, Xiang Li, and Zhen Yang. 2021. Automated audio captioning with weakly supervised pre-training and word selection methods. In *Workshop on Detection and Classification of Acoustic Scenes and Events*.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2025. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–17.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2024. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. 2023a. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, Yi Ren, Zhou Zhao, and Shinji Watanabe. 2023b. Audiogpt: Understanding and generating speech, music, sound, and talking head.

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

Vladimir Iashin and Esa Rahtu. 2020. Multi-modal dense video captioning.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

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

A. Sophia Koepke, Andreea-Maria Oncescu, João F. Henriques, Zeynep Akata, and Samuel Albanie. 2023. Audio retrieval with natural language queries: A benchmark study. *IEEE Transactions on Multimedia*, 25:2675–2685.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer.

Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, et al. 2024b. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. Halueval: A large-scale hallucination evaluation benchmark for large language models.

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

Huadai Liu, Rongjie Huang, Yang Liu, Hengyuan Cao, Jialei Wang, Xize Cheng, Siqi Zheng, and Zhou Zhao. 2024a. Audiolcm: Text-to-audio generation with latent consistency models.

Hui Liu and Xiaojun Wan. 2023. Models see hallucinations: Evaluating the factuality in video captioning.

Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhanhui Kang, and Xirong Li. 2024b. Phd: A chatgpt-prompted visual hallucination evaluation dataset.

Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. 2023. Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. *ArXiv*, abs/2311.01813.

Meng Luo, Hao Fei, Bobo Li, Shengqiong Wu, Qian Liu, Soujanya Poria, Erik Cambria, Mong-Li Lee, and Wynne Hsu. 2024. Panosent: A panoptic sextuple extraction benchmark for multimodal conversational aspect-based sentiment analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7667–7676.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Ning Miao, Yee Whye Teh, and Tom Rainforth. 2023. Selfcheck: Using llms to zero-shot check their own step-by-step reasoning.

Taichi Nishimura, Shota Nakada, and Masayoshi Kondo. 2024. On the audio hallucinations in large audio-video language models. *arXiv preprint arXiv:2401.09774*.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2016. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models.

Vipula Rawte, Sarthak Jain, Aarush Sinha, Garv Kaushik, Aman Bansal, Prathiksha Rumale Vishwanath, Samyak Rajesh Jain, Aishwarya Naresh Reganti, Vinija Jain, Aman Chadha, et al. 2024. Vibe: A text-to-video benchmark for evaluating hallucination in large multimodal models. *arXiv preprint arXiv:2411.10867*.

Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaoke Jiang, Yihao Chen, Yuda Xiong, Hao Zhang, Feng Li, Peijun Tang, Kent Yu, and Lei Zhang. 2024. Grounding dino 1.5: Advance the "edge" of open-set object detection.

Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. 2023. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers.

Maitreya Suin and A. N. Rajagopalan. 2020. An efficient framework for dense video captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12039–12046.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. 2025. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation.

Zhiqing Sun, Yikang Shen, Hongxin Zhang, Qinhong Zhou, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2024. Salmon: Self-alignment with instructable reward models.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Gabriel Tjio, Ping Liu, Joey Tianyi Zhou, and Rick Siow Mong Goh. 2021. Adversarial semantic hallucination for domain generalized semantic segmentation.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020a. Asking and answering questions to evaluate the factual consistency of summaries.

Jing Yi Wang, Nicholas Sukiennik, Tong Li, Weikang Su, Qianyue Hao, Jingbo Xu, Zihan Huang, Fengli Xu, and Yong Li. 2024a. A survey on human-centric llms. *arXiv preprint arXiv:2411.14491*.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. 2024b. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2020b. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research.

Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. 2024c. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arXiv preprint arXiv:2406.16338*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, pages 24824–24837.

Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. 2024. Longvlm: Efficient long video understanding via large language models.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihan Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. 2024b. Cogvideox: Text-to-video diffusion models with an expert transformer.

Zongxin Yang, Guikun Chen, Xiaodi Li, Wenguan Wang, and Yi Yang. 2024c. Doraemongpt: Toward understanding dynamic scenes with large language models. *arXiv preprint arXiv:2401.08392*.

Zhongjie Ye, Helin Wang, Dongchao Yang, and Yuexian Zou. 2021. Improving the performance of automated audio captioning via integrating the acoustic and semantic information. In *Workshop on Detection and Classification of Acoustic Scenes and Events*.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. 2025. Videollama 3: Frontier multimodal foundation models for image and video understanding.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. 2024. Open-sora: Democratizing efficient video production for all.

A Dataset Specifications

Microsoft COCO Captions is a large-scale dataset developed for image captioning research, containing over 330,000 images with more than 1.5 million human-annotated descriptions. Each image is paired with at least five independent captions, ensuring diversity and reliability.The dataset is widely used for training and evaluation of automatic image captioning models . Its core goal is to foster the synergy between computer vision and natural language processing, enhancing models' ability to understand visual scenes.

Nocaps is developed to assess models' proficiency in Novel Object Captioning. It consists of 15,100 images from Open Images validation and test sets, with 166,100 human-annotated captions. Nocaps leverages COCO captions for training while providing image-level labels and object bounding boxes from Open Images. As Open Images covers more object categories than COCO, nearly 400 categories in the test set lack corresponding captions in the training set, hence the name "Nocaps".

Flickr30K is a dataset widely used for image caption generation, containing 31,783 images collected from Flickr, each paired with five human-annotated captions. The dataset evaluates models on their ability to generate captions accurately aligned with real-world image content, following a standard training, validation, and test set partitioning.Flickr30K prioritizes linguistic diversity and naturalness, making it a key benchmark for visual-linguistic tasks.

MVSD is a multimodal video dataset developed to support research on translation and paraphrase generation. It contains 2,089 video clips with 85K English descriptions, along with thousands of descriptions per video across multiple languages. Each video clip is under 10 seconds and depicts a single, unambiguous action or event.The dataset leverages short videos as stimuli to elicit natural linguistic responses, enabling same-language descriptions to function as paraphrases and cross-language descriptions as translations.

MSR-VTT is a video description dataset developed to connect video understanding with natural language processing. It contains 10,000 web video clips totaling 41.2 hours, with 200,000 clip-sentence pairs covering 20 categories of diverse video content.Each clip has approximately 20 human-annotated descriptions by 1,327 AMT workers, ensuring rich linguistic variation. The dataset enables research in video captioning, retrieval, and multimodal learning.

VATEX is a multilingual video captioning dataset developed for video-language research. It contains 41,250 unique videos and 825,000 high-quality captions in both English and Chinese, including 206,000 English-Chinese parallel translation pairs.Each video is annotated with 10 diverse captions in both English and Chinese by 20 human annotators. VATEX supports multilingual video captioning and video-guided machine translation by leveraging spatiotemporal video context.

AudioCaps is a large-scale dataset developed for audio captioning research, enabling models to generate natural language descriptions for environmental sounds. It consists of 46,000 audio clips, each paired with human-written captions, sourced from AudioSet.Each clip is approximately 10 seconds long, and the dataset includes five captions per clip to ensure linguistic diversity. AudioCaps facilitates audio-based scene understanding and sound event recognition.

Clotho V2 is a dataset developed for audio captioning research, enabling models to generate natural language descriptions of general audio content. It consists of 4,981 audio samples, each lasting 15 to 30 seconds, sourced from Freesound.Each audio sample is paired with five human-written captions, containing 8 to 20 words, ensuring linguistic diversity. Clotho V2 facilitates general sound event recognition and audio scene understanding.

AudioSetCaps is a large-scale dataset developed for automated audio captioning research, containing 1.9 million audio-caption pairs sourced from AudioSet.Captions are generated using a sophisticated pipeline combining audio-language and large language models, ensuring fine-grained and high-quality descriptions. The dataset supports audio-text retrieval, zero-shot audio classification, and automated captioning.

HRS-Bench employs 3,000 prompts per skill to evaluate T2I models across accuracy, robustness, generalization, fairness, and bias, using human annotation and template-based generation. Prompts cover object counting ("Three cats on two chairs"), visual text ("A sign with 'Speed Limit 60'"), paraphrasing ("A cat is on the sofa" vs. "On the sofa, a cat is resting"), typos ("A womn is hollding a cup") and creativity ("A fish flying in the clouds") . Fairness and bias prompts ensure gender neutrality and unbiased representations.

T2I-CompBench++ is a benchmark developed for assessing compositional text-to-image generation. It consists of 8,000 prompts, categorized into attribute binding, object relationships, generative numeracy, and complex compositions. The benchmark evaluates models' capacity to bind attributes correctly (e.g., "A red book and a yellow vase"), generate spatially accurate relationships (e.g., "A cat in front of a chair"), and handle numeracy (e.g., "Four swans and two suitcases").

T2V-Bench utilizes structured prompts to assess T2V models across spatial relationships, motion binding, action binding, object interactions, attribute consistency, dynamic attributes, and numeracy understanding. Prompts are generated using real-world user inputs, predefined templates, and GPT-4-assisted augmentation.Examples include spatial positioning (e.g., "A bird flies to the left of a hot air balloon"), motion dynamics (e.g., "A robot walks on the moon"), and temporal changes (e.g., "A leaf turns from green to yellow"). This design ensures compositional complexity, providing a rigorous evaluation of models' scene comprehension and motion synthesis.

FETV is a benchmark developed for the fine-grained assessment of open-domain T2V generation. It categorizes 619 prompts based on major content, attribute control, and prompt complexity, ensuring a structured assessment.Prompts cover spatial and temporal attributes, including actions, kinetic motions, light changes, fluid motions, speed, motion direction, and event order. The prompts are sourced from existing text-video datasets and manually created scenarios, offering a diverse and rigorous evaluation framework.

WavText5K is a dataset developed for audio-text retrieval research, containing 4,525 audio clips with 4,348 unique descriptions sourced from web-crawled sound effects. Prompts describe isolated audio events with rich contextual details. Unlike generic labels, these prompts provide fine-grained scene descriptions.The dataset supports contrastive learning-based retrieval by aligning natural language queries with sound events, improving the accuracy of audio-text alignment.

FSD50K is an open dataset comprising 51,197 manually annotated audio clips spanning 200 classes, derived from the AudioSet ontology. This dataset was developed for large-scale, multi-label sound event classification.The audio clips, sourced mainly from Freesound, range in duration from 0.3 to 30 seconds. FSD50K utilizes weak labels through clip-level annotations, with its evaluation set undergoing rigorous manual verification to ensure high-quality labeling. The dataset supports various tasks, including audio classification, hierarchical classification, cross-dataset evaluation, and sound separation.

SoundDescs is a benchmark dataset designed for text-based audio retrieval, containing 32,979 audio clips paired with natural language descriptions. The data is sourced from the BBC Sound Effects archive, covering 23 categories, including nature, urban soundscapes, and human activities. Audio durations range from a few seconds to several hours, and descriptions vary in length and complexity, providing a rich resource for evaluating retrieval models. The dataset is split into 23,085 training samples, 4,947 validation samples, and 4,947 test samples.