Unveiling Hallucination in Text, Image, Video, and Audio Foundation Models: A Comprehensive Review

Anonymous ACL submission

Abstract

The rapid advancement of foundation models (FMs) across language, image, audio, and video domains has shown remarkable capabilities in diverse tasks. However, the proliferation of FMs brings forth a critical challenge: the potential to generate hallucinated outputs, particularly in high-stakes applications. The tendency of foundation models to produce hallucinated content arguably represents the biggest hindrance to their widespread adoption in real-world scenarios, especially in domains where reliability and accuracy are paramount. This survey paper presents a comprehensive overview of recent developments that aim to identify and mitigate the problem of hallucination in FMs, spanning text, image, video, and audio modalities. By synthesizing recent advancements in detecting and mitigating hallucination across various modalities, the paper aims to provide valuable insights for researchers, developers, and practitioners. Essentially, it establishes a clear framework encompassing definition, taxonomy, and detection strategies for addressing hallucination in multimodal foundation models, laying the foundation for future research and development in this pivotal area.

1 Introduction

The rapid progress in large-scale foundation models (FMs), spanning language, image, audio, and video domains, has revolutionized the field of artificial intelligence (AI). Models such as GPT-3 (Brown et al., 2020), MiniGPT-4 (Zhu et al., 2023), AudioLLM (Borsos et al., 2023), and LaViLa (Zhao et al., 2022) have demonstrated remarkable abilities across diverse tasks, from text generation to multimodal understanding. However, as these models are increasingly deployed in high-stakes domains, understanding and mitigating their potential to generate hallucinated outputs – content that appears plausible but is factually incorrect or inconsistent with the input – has become a critical priority. Hallucination, often unintended, can arise due to factors like biases in training data, limited access to current information, or the model's constraints in understanding and generating contextually precise responses. Deploying these models without addressing their hallucination tendencies may result in misinformation, incorrect conclusions, and adverse consequences. Thus, it is imperative to comprehensively study and understand the hallucination behavior of FMs across different modalities.

1.1 Motivation and Contributions

Most of the existing survey papers have explored hallucination in the context of large language models (LLMs) (Huang et al., 2023), (Tonmoy et al., 2024). Recent studies have shown that hallucination can also occur in vision, audio, and video foundation models, highlighting the need for a comprehensive understanding of this challenge across multiple modalities (Liu et al., 2024a), (Sahoo et al., 2024), (Rawte et al., 2023b). To address this gap, the present survey aims to provide a holistic and multimodal perspective on the hallucination challenge in FMs. This review comprehensively examines existing research across language, vision, video, and audio domains to understand the mechanisms, detection methods, and mitigation strategies for hallucination in FMs. It serves as a vital resource for researchers and practitioners, aiding in the development of more robust AI solutions. Additionally, it includes a detailed taxonomy diagram in Fig. 1 and a summarized Table 1 illustrating recent advancements across different modalities. Please refer to Table 9.1 of the appendix. The contributions of this survey paper are as follows:

- Establish a precise definition and structured taxonomy of hallucination in the context of large-scale foundation models.
- Identify the key factors and mechanisms that contribute to the emergence of hallucination

across different modalities.

- Explore the various detection and mitigation strategies that have been proposed to address the hallucination problem in a multimodal setting.
- Highlight the open challenges and future research directions in this critical area.

2 Hallucination in Large Language Models

Despite the progress of LLMs, a notable challenge persists in their proneness to hallucinate, impeding their practical implementation. For instance, the illustration in Figure 2 exemplifies the generated response by the LLM, showcasing indications of hallucination.

2.1 Hallucination Detection and Mitigation

Detecting hallucinations in LLMs is crucial for ensuring the credibility and reliability of their results, especially in scenarios requiring factual correctness. Existing fact-checking methods often rely on complex modules or external databases, requiring either output probability distributions or interfacing with external sources. The SelfCheckGPT method proposed by (Manakul et al., 2023) offers a zero-resource black-box solution for detecting hallucinations in any LLM without relying on external resources. This method operates on the principle that an LLM familiar with a topic will produce consistent and comparable facts in its responses. In contrast, randomly sampled responses from an unfamiliar topic are likely to contain contradicting and hallucinated facts. Continuing the exploration of methods for passage-level hallucination detection, (Yang et al., 2023) proposed a novel self-check approach based on reverse validation, aiming to automatically identify factual errors without external resources. They introduced a benchmark, Passage-level Hallucination Detection(PHD) , generated using ChatGPT and annotated by human experts, to assess different methods. Assessing the accuracy of long text generated by LLMs is challenging because it often contains both accurate and inaccurate information, making simple quality judgments insufficient. To address this, (Min et al., 2023) introduced FACTSCORE (Factual Precision in Atomicity Score), a new evaluation method that breaks down text into individual facts and measures their reliability. The study (Huang and Chang, 2023) introduced a unique strategy to mitigate hallucination risks in LLMs by drawing parallels with established web systems. They identified the absence of a "citation" mechanism in LLMs, which refers to acknowledging or referencing sources or evidence, as a significant gap.

Addressing the need to identify factual inaccuracies in LLM-generated content, (Rawte et al., 2024b) developed a multi-task learning (MTL) framework, integrating advanced long text embeddings like e5-mistral-7b-instruct, along with models such as GPT-3, SpanBERT, and RoFormer. This MTL approach demonstrated a 40% average improvement in accuracy on the FACTOID benchmark compared to leading textual entailment methods. Hallucination mitigation efforts have predominantly relied on empirical methods, leaving uncertainty regarding the possibility of complete elimination. To tackle this challenge, (Xu et al., 2024b) introduced a formal framework defining hallucination as inconsistencies between computable LLMs and a ground truth function. Through this framework, the study examines existing hallucination mitigation strategies and their practical implications for real-world LLM deployment. The study (Rawte et al., 2024c) introduces the Sorry, Come Again (SCA) prompting technique to address hallucination in contemporary LLMs. SCA enhances comprehension through optimal paraphrasing and injecting [PAUSE] tokens to delay LLM generation. It analyzes linguistic nuances in prompts and their impact on the hallucinated generation, emphasizing how prompts with lower readability, formality, or concreteness pose challenges. (Rawte et al., 2023a) investigates how LLMs respond to factually correct and incorrect prompts, categorizing their hallucinations into mild, moderate, and alarming subcategories. Additionally, the paper introduces the Hallucination eLiciTation dataset, comprising 75,000 text snippets annotated by humans, and introduces a novel Hallucination Vulnerability Index metric.

Benchmark Evaluation: In certain instances, LLMs engage in a phenomenon termed "hallucination snowballing," where they fabricate false claims to rationalize prior hallucinations, despite acknowledging their inaccuracy. To empirically explore this phenomenon, (Zhang et al., 2023a) devised three question-answering datasets spanning diverse domains, wherein ChatGPT and GPT-4 often furnish inaccurate answers alongside explanations fea-

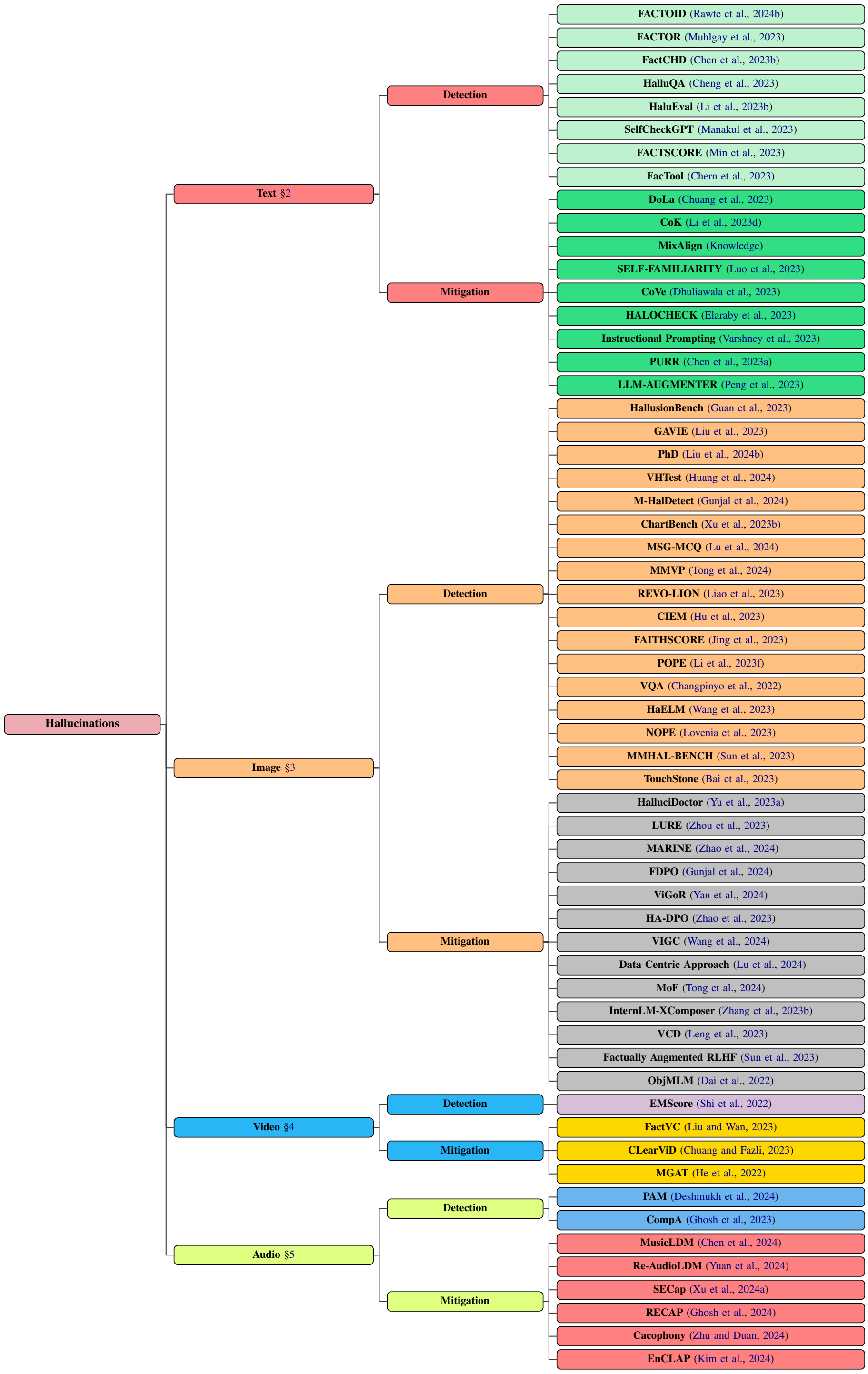

Figure 1: Taxonomy of hallucination in large foundation models, organized around detection and mitigation techniques.

turing at least one false claim. Significantly, the study suggests that the language model can discern these false claims as incorrect. Another benchmark dataset FactCHD (Chen et al., 2023b), was introduced to detect fact-conflicting hallucinations within intricate inferential contexts. It encompasses a range of datasets capturing different factuality patterns and integrates fact-based evidence chains to improve assessment accuracy. The study by (Li et al., 2023b) introduced a dataset to assess the performance of LLMs in recognizing hallucinations. The outcomes highlighted ChatGPT's inclination to produce hallucinated content, particularly on certain topics, introducing unverifiable information.

Figure 2: LLM responses showing the types of hallucinations, highlighted in red, green, and blue (Zhang et al., 2023d).

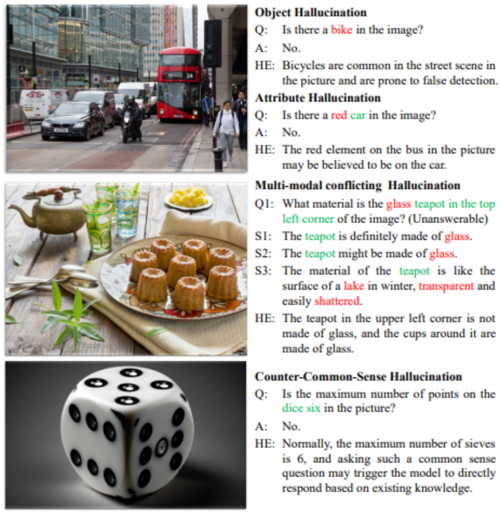

Figure 3: Four IVL-Hallu examples in Prompted Hallucination Dataset(PhD) (Liu et al., 2024b) including visuals and the matching question-answer pairs and hallucination elements (HE). While words annotated in red do not exist or do not match within the image, words annotated in green have correspondences within the image. Question, Answer, and Statement are denoted by the letters Q, A, and S respectively.

3 Hallucination in Large Vision-Language Models

Large Vision-Language Models (LVLMs) have garnered significant attention in the AI community for their capacity to handle visual and textual data simultaneously. Nonetheless, similar to LLMs, LVLMs also confront the issue of hallucination. Figure 3 illustrates an example of visual hallucination.

3.1 Hallucination Detection and Mitigation

(Dai et al., 2022) investigate the issue of object hallucinations in Vision-Language Pre-training (VLP) models, where textual descriptions generated by these models contain non-existent or inaccurate objects based on input images. (Li et al., 2023f) reveal widespread and severe object hallucination issues and suggests that visual instructions may influence hallucination, noting that objects frequently appearing in visual instructions or co-occurring with image objects are more likely to be hallucinated. To enhance the evaluation of object hallucination, they introduce a polling-based query method called POPE, which demonstrates improved stability and flexibility in assessing object hallucination. The absence of a standardized metric for assessing object hallucination has hindered progress in understanding and addressing this issue. To address this gap, (Lovenia et al., 2023) introduce NOPE (Negative Object Presence Evaluation), a novel benchmark for evaluating object hallucination in vision-language (VL) models through visual question answering (VQA). Utilizing LLMs, the study generates 29.5k synthetic negative pronoun (NegP) data for NOPE. It extensively evaluates the performance of 10 VL models in discerning the absence of objects in visual questions, alongside their standard performance on visual questions across nine other VQA datasets. Most existing efforts focus primarily on object hallucination, overlooking the diverse types of LVLM hallucinations. The study by (Liu et al., 2024b) delves into Intrinsic Vision-Language Hallucination (IVL-Hallu) and proposes several novel IVL-Hallu tasks categorized into four types: attribute, object, multi-modal conflicting, and counter-common-sense hallucination. To assess and explore IVL-Hallu, they introduce a challenging benchmark dataset and conduct experiments on five LVLMs, revealing their incapacity to effectively address the proposed IVL-Hallu tasks. To mitigate object hallucination in LVLMs without resorting to costly training or API reliance, (Zhao et al., 2024) introduces MARINE, which is both training-free and API-free. MARINE enhances the visual understanding of LVLMs by integrating existing open-source vision models and utilizing guidance without classifiers to integrate object grounding features, thereby improving the precision of the generated outputs. Evaluations across six LVLMs reveal MARINE's effectiveness in reducing hallucinations and enhancing output detail, validated through assessments using GPT-4V.

HalluciDoctor (Yu et al., 2023a) tackles hallucinations in Multi-modal Large Language Models (MLLMs) by using human error detection to identify and eliminate various types of hallucinations. Through rebalancing data distribution via counterfactual visual instruction expansion, they successfully mitigate 44.6% of hallucinations while maintaining competitive performance. Despite proficiency in visual semantic comprehension and meme humor, MLLMs struggle with chart analysis and understanding. Addressing this, (Xu et al., 2023b) proposed ChartBench, a benchmark assessing chart comprehension. ChartBench exposes MLLMs' limited reasoning with complex charts, prompting the need for novel evaluation metrics like Acc+ and a handcrafted prompt, ChartCoT. The study (Zhang et al., 2023b) introduced InternLM-XComposer, an LVLM aimed at designed to address the challenge of hallucination in image-text comprehension and composition. The performance of InternLM-XComposer's text-image composition is evaluated through a robust procedure involving both human assessment and comparison to GPT4-Vision, with the model demonstrating competitive performance against solutions like GPT4-V and GPT3.5.

3.2 Benchmark Evaluation

The current methods of developing LVLMs rely heavily on annotated benchmark datasets, which can exhibit domain bias and limit model generative capabilities. To address this, (Li et al., 2023e) proposed a novel data collection approach that synthesizes images and dialogues synchronously for visual instruction tuning, yielding a large dataset of image-dialogue pairs and multi-image instances. Another study (Huang et al., 2024) introduced VHTest, a benchmark dataset with 1,200 diverse visual hallucinations (VH) instances across 8 VH modes. Evaluation of three SOTA MLLMs showed varying performance, with GPT-4V exhibiting lower hallucination than MiniGPT-v2. The study (Rawte et al., 2024a) categorizes visual hallucination in VLMs into eight orientations and introduces a dataset of 2,000 samples covering these types. They propose three main categories of methods to mitigate hallucination: data-driven approaches, training adjustments, and post-processing techniques. (Wang et al., 2024) propose the Visual Instruction Generation and Correction (VIGC) framework to address the scarcity of high-quality instruction-tuning data for MLLMs. VIGC enables MLLMs to generate diverse instruction-tuning data while iteratively refining its quality through Visual Instruction Correction (VIC), mitigating hallucination risks. The framework produces diverse, high-quality data for fine-tuning models, validated through evaluations, improving benchmark performance, and overcoming language-only data limitations.

4 Hallucinations in Large Video Models

Large Video Models (LVMs) represent a significant advancement, allowing for the processing of video data at scale. Despite their potential for various applications like video understanding and generation, LVMs face challenges with hallucinations, where misinterpretations of video frames can result in artificial or inaccurate visual data. This issue arises due to the complexity of video data, which requires the model to thoroughly process and comprehend it. Figure 4 demonstrates the instances of hallucination observed in LVMs.

4.1 Hallucination Detection and Mitigation

The intricate task of dense video captioning, involving the creation of descriptions for multiple events within a continuous video, necessitates a thorough understanding of video content and contextual reasoning to ensure accurate description generation. However, this endeavor faces numerous challenges, potentially resulting in instances of inaccuracies and hallucinations (Iashin and Rahtu, 2020), (Suin and Rajagopalan, 2020). Traditional methods detect event proposals first, then caption subsets, risking hallucinations due to overlooking temporal dependencies. To address this, (Mun et al., 2019) introduces a novel approach to modeling temporal dependencies and leveraging context for coherent storytelling. By integrating an event sequence generation network and a sequential video captioning network trained with reinforcement learning and two-level rewards, the model captures contextual information more effectively, yielding coherent and accurate captions while minimizing the risk of hallucinations. Another study (Liu and Wan, 2023) in-

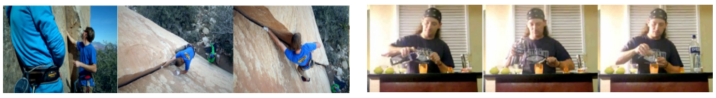

Figure 4: A video featuring descriptions generated by VLTinT model and ground truth (GT) with description errors highlighted in red italics. (Chuang and Fazli, 2023).

troduces a novel weakly-supervised, model-based factuality metric called FactVC, which outperforms previous metrics. Furthermore, they provide two annotated datasets to promote further research in assessing the factuality of video captions. (Wu and Gao, 2023) proposed a context-aware model that incorporates information from past and future events to conditionally influence the description of the current event. Their approach utilizes a robust pre-trained context encoder to encode information about the surrounding context events, which is then integrated into the captioning module using a gate-attention mechanism. Experimental findings on the YouCookII and ActivityNet datasets demonstrate that the proposed context-aware model outperforms existing context-aware and pre-trained models by a significant margin. To enhance dense video captioning, (Zhou et al., 2024) introduced a streaming model comprising a memory module for long video handling and a streaming decoding algorithm enabling predictions before video completion. This approach notably boosts performance on dense video captioning benchmarks such as ActivityNet, YouCook2, and ViTT.

Video infilling and prediction tasks are crucial for assessing a model's ability to comprehend and anticipate the temporal dynamics within video sequences (Höppe et al., 2022). To address this, (Himakunthala et al., 2023) introduced an inference-time challenge dataset containing keyframes with dense captions and structured scene descriptions. This dataset contains keyframes supplemented with unstructured dense captions and structured FAMOUS: *(Focus, Action, Mood, Objects, and Setting)* scene descriptions, providing valuable contextual information to support the models' understanding of the video content. They employed various language models like GPT-3, GPT-4, and Vicuna with greedy decoding to mitigate hallucination risks. Prominent developments in video inpainting have been observed recently, especially in situations where explicit guidance like optical flow helps to propagate missing pixels across frames (Ouyang et al., 2021). However, difficulties and constraints occur from a lack of cross-frame information. The study (Yu et al., 2023b) aims to tackle the opposite issue rather than depending on using pixels from other frames. The suggested method presents a Deficiency-aware Masked Transformer (DMT), a dual-modality-compatible inpainting framework. This approach improves handling scenarios with incomplete information by pre-training an image inpainting model to serve as a prior for training the video model.

Understanding scene affordances, which involve potential actions and interactions within a scene, is crucial for comprehending images and videos. (Kulal et al., 2023) introduced a method for realistically inserting people into scenes. The model seamlessly integrates individuals into scenes by deducing realistic poses based on the context and ensuring visually pleasing compositions. (Chuang and Fazli, 2023) introduced CLearViD, a transformer-based model that utilizes curriculum learning techniques to enhance performance. By adopting this approach, the model acquires more robust and generalizable features. Furthermore, CLearViD incorporates the Mish activation function to address issues like vanishing gradients, thereby reducing the risk of hallucinations by introducing nonlinearity and non-monotonicity. Extensive experiments and ablation studies validate the effectiveness of CLearViD, with evaluations on ActivityNet Captions and YouCook2 datasets showcasing significant improvements over existing SOTA models in terms of diversity metrics.

4.2 Benchmark Evaluation

The study (Zhang et al., 2006) created a novel two-level hierarchical fusion method to hallucinate facial expression sequences from training video samples using only one frontal face image with a neutral expression. To effectively train the system, they introduced a dataset specifically designed for facial expression hallucination, which included 112 video sequences covering four types of facial expressions (happy, angry, surprise, and fear) from 28 individuals, resulting in the generation of reasonable facial expression sequences in both the temporal and spatial domains with less artifact. In the realm of video understanding, the development of end-to-end chat-centric systems has become a growing area of interest. (Zhou et al., 2018) assembled the YouCook2 dataset, an extensive set of cooking videos with temporally localized and described procedural segments, to facilitate procedure learning tasks. The study by (Li et al., 2023c) introduces "VideoChat", a novel approach integrating video foundation models and LLMs through a learnable neural interface to enhance spatiotemporal reasoning, event localization, and causal relationship inference in video understanding. The researchers constructed a video-centric instruction dataset with detailed descriptions and conversations, emphasizing spatiotemporal reasoning and causal relationships. To counteract model hallucination, they employed a multi-step process to condense video descriptions into coherent narratives using GPT-4 and refined them to improve clarity and coherence. To explore the challenge of deducing scene affordances, (Kulal et al., 2023) curated a dataset of 2.4M video clips, showcasing a variety of plausible poses that align with the scene context.

5 Hallucinations in Large Audio Models

Large audio models (LAMs) have emerged as a powerful tool in the realm of audio processing and generation, with a wide range of applications like speech recognition, music analysis, audio synthesis, and captioning (Latif et al., 2023), (Hussain et al., 2023), (Ghosal et al., 2023). Although these models have demonstrated remarkable capabilities across various domains, they are susceptible to hallucinations. These anomalies can take several forms, from creating unrealistic audio by piecing together fabricated snippets to injecting false information, such as quotes or facts, into summaries. Additionally, they may fail to accurately capture the inherent features of audio signals, such as timbre, pitch, or background noise (Shen et al., 2023).

5.1 Hallucination Detection and Mitigation

In the realm of audio captioning, where natural language descriptions for audio clips are automatically generated, a significant challenge arises from the over-reliance on the visual modality during the pre-training of audio-text models. This reliance introduces data noise and hallucinations, ultimately undermining the accuracy of the resulting captions. To address this issue, (Xu et al., 2023a) introduced an AudioSet tag-guided model designed to bootstrap large-scale audio-text data (BLAT). Notably, this model sidesteps the incorporation of video, thus minimizing noise associated with the visual modality. The experimental findings across a range of tasks, including retrieval, generation, and classification, validate the effectiveness of BLAT in mitigating hallucination issues.

Speech emotions play a crucial role in human communication and find extensive applications in areas such as speech synthesis and natural language understanding. However, traditional categorization approaches may fall short of capturing the nuanced and intricate nature of emotions conveyed in human speech (Jiang et al., 2019), (Han et al., 2021), (Ye et al., 2021). SECap (Xu et al., 2024a), a framework designed for speech emotion captioning. It aims to capture the intricate emotional nuances of speech using natural language. SECap utilizes various components, including LLaMA as the text decoder, HuBERT as the audio encoder, and Q-Former as the Bridge-Net, to generate coherent emotion captions based on speech features. Audio-language models, despite their capability for zero-shot inference, confront challenges like hallucinating task-specific details despite strong performance. To address this, (Elizalde et al., 2024) introduces the Contrastive Language-Audio Pretraining (CLAP) model. Pre-trained with 4.6 million diverse audio-text pairs, CLAP features a dual-encoder architecture, enhancing representation learning for improved task generalization across sound, music, and speech domains.

5.2 Benchmark Evaluation

To address the scarcity of data in the specific domain of music captioning, (Doh et al., 2023) introduced LP-MusicCaps, a comprehensive dataset comprising 0.5 million audio clips accompanied by approximately 2.2 million captions. Leverag-

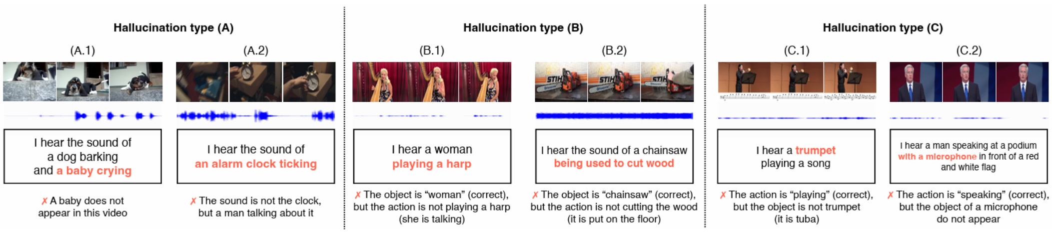

Figure 5: Audio hallucination examples for each classes - Type A: *Involving hallucinations of both objects and actions* Type B: *Featuring accurate objects but hallucinated actions* Type C: *Displaying correct actions but hallucinated objects* (Nishimura et al., 2024).

ing LLMs, they train a transformer-based music captioning model with the dataset and assess its performance under zero-shot and transfer-learning scenarios, demonstrating its superiority over supervised baseline models. Another author (Nishimura et al., 2024) investigates audio hallucinations in large audio-video language models, where audio descriptions are generated primarily based on visual information, neglecting audio content. They have classified these hallucinations into three distinct types such as *Involving hallucinations of both objects and actions*, *Featuring accurate objects but hallucinated actions*, and *Displaying correct actions but hallucinated objects* as represented in Fig. 5. In their investigation, they gathered 1000 sentences by soliciting audio information and then annotated them to determine whether they contained auditory hallucinations, further categorizing the type of hallucination if detected. To assess compositional reasoning in LAMs, (Ghosh et al., 2023) introduced CompA, consisting of two expert-annotated benchmarks primarily focused on real-world audio samples. This benchmark is employed to fine-tune CompA-CLAP with a novel learning approach, enhancing its compositional reasoning skills and demonstrating substantial improvement over all the baseline models in tasks requiring compositional reasoning.

6 Hallucination: Good or Bad?

Hallucinations in large-scale models present a complex interplay between creativity and uncertainty. On one hand, the ability to traverse beyond conventional data boundaries can lead to the generation of novel and innovative outputs. Hallucinations can spark exploratory learning, revealing unexpected patterns and features within the data. They can also serve as a form of stress testing, improving the model's robustness and adaptability. Furthermore, these unexpected outputs can even inspire human creativity, serving as a springboard for new ideas and perspectives (Rawte et al., 2023b). However, this dual nature of hallucinations also introduces significant drawbacks. The quality and coherence of hallucinatory outputs can be questionable, posing challenges in applications where accuracy and reliability are paramount. Hallucinations can also propagate misinformation and biases present in the model's training data, potentially reinforcing existing prejudices and eroding user trust. The reduced interpretability of these outputs can further undermine the model's credibility and adoption. Ethical concerns arise when hallucinations produce inappropriate, offensive, or harmful content. Careful monitoring and control mechanisms are essential to prevent the generation of outputs that could cause harm or distress to users. Navigating this intricate balance between exploration and fidelity is crucial for maximizing the utility of large models while mitigating the risks associated with unexpected outputs. Overall, the phenomenon of hallucinations in large-scale models highlights the need for a nuanced understanding and strategic management of these capabilities.

7 Conclusion and Future Directions

This survey paper systematically categorizes existing research on hallucination within FMs, providing comprehensive insights into critical aspects such as detection, mitigation, tasks, datasets, and evaluation metrics. It addresses the pressing issue of hallucination in FMs, acknowledging its widespread impact across various domains. By examining recent advancements in detection and mitigation techniques, the paper underscores the importance of addressing this challenge, given FMs' indispensable role in critical tasks. Its primary contribution lies in presenting a structured taxonomy for classifying hallucination in FMs, spanning text, image, video, and audio domains.

8 Limitation

Previous survey papers primarily focused on hallucination in Large Language Models and did not extensively cover hallucinations in vision, audio, and video modalities. In this survey paper, our aim is to provide a comprehensive overview of hallucinations across all modalities, considering that hallucinations can occur in any large foundation model. Despite our efforts to provide a comprehensive summary of recent advancements related to hallucination techniques in all foundational models, we acknowledge that we may miss some relevant work in the field.

References

Muhammad Aurangzeb Ahmad, Ilker Yaramis, and Taposh Dutta Roy. 2023. Creating trustworthy llms: Dealing with hallucinations in healthcare ai. *arXiv preprint arXiv:2311.01463*.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when its lying. *arXiv preprint arXiv:2304.13734*.

Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. 2023. Touchstone: Evaluating vision-language models by language models. *arXiv preprint arXiv:2308.16890*.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2023. Audiolm: a language modeling approach to audio generation.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Soravit Changpinyo, Linting Xue, Idan Szpektor, Ashish V Thapliyal, Julien Amelot, Michal Yarom, Xi Chen, and Radu Soricut. 2022. Maxm: Towards multilingual visual question answering. *arXiv preprint arXiv:2209.05401*.

Anthony Chen, Panupong Pasupat, Sameer Singh, Hongrae Lee, and Kelvin Guu. 2023a. Purr: Efficiently editing language model hallucinations by denoising language model corruptions. *arXiv preprint arXiv:2305.14908*.

Ke Chen, Yusong Wu, Haohe Liu, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2024. Musicldm: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1206–1210. IEEE.

Xiang Chen, Duanzheng Song, Honghao Gui, Chengxi Wang, Ningyu Zhang, Fei Huang, Chengfei Lv, Dan Zhang, and Huajun Chen. 2023b. Unveiling the siren's song: Towards reliable fact-conflicting hallucination detection. *arXiv preprint arXiv:2310.12086*.

Qinyuan Cheng, Tianxiang Sun, Wenwei Zhang, Siyin Wang, Xiangyang Liu, Mozhi Zhang, Junliang He, Mianqiu Huang, Zhangyue Yin, Kai Chen, et al. 2023. Evaluating hallucinations in chinese large language models. *arXiv preprint arXiv:2310.03368*.

I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.

Cheng-Yu Chuang and Pooyan Fazli. 2023. Clearvid: Curriculum learning for video description. *arXiv preprint arXiv:2311.04480*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *arXiv preprint arXiv:2401.01301*.

Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. 2022. Plausible may not be faithful: Probing object hallucination in vision-language pre-training. *arXiv preprint arXiv:2210.07688*.

Soham Deshmukh, Dareen Alharthi, Benjamin Elizalde, Hannes Gamper, Mahmoud Al Ismail, Rita Singh, Bhiksha Raj, and Huaming Wang. 2024. Pam: Prompting audio-language models for audio quality assessment. *arXiv preprint arXiv:2402.00282*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. 2023. Lp-musiccaps: Llm-based pseudo music captioning. *arXiv preprint arXiv:2307.16372*.

Mohamed Elaraby, Mengyin Lu, Jacob Dunn, Xueying Zhang, Yu Wang, and Shizhu Liu. 2023. Halo: Estimation and reduction of hallucinations in open-source weak large language models. *arXiv preprint arXiv:2308.11764*.

Benjamin Elizalde, Soham Deshmukh, and Huaming Wang. 2024. Natural language supervision for general-purpose audio representations. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 336–340. IEEE.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. 2022. Rarr: Researching and revising what language models say, using language models. *arXiv preprint arXiv:2210.08726*.

Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. 2023. Text-to-audio generation using instruction-tuned llm and latent diffusion model. *arXiv preprint arXiv:2304.13731*.

Sreyan Ghosh, Sonal Kumar, Chandra Kiran Reddy Evuru, Ramani Duraiswami, and Dinesh Manocha. 2024. Recap: retrieval-augmented audio captioning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1161–1165. IEEE.

Sreyan Ghosh, Ashish Seth, Sonal Kumar, Utkarsh Tyagi, Chandra Kiran Evuru, S Ramaneswaran, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2023. Compa: Addressing the gap in compositional reasoning in audio-language models. *arXiv preprint arXiv:2310.08753*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. 2023. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models. *arXiv preprint arXiv:2310.14566*.

Nuno M Guerreiro, Duarte M Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André FT Martins. 2023. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143.

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 *DCASE*, pages 6–10.

Mengge He, Wenjing Du, Zhiquan Wen, Qing Du, Yutong Xie, and Qi Wu. 2022. Multi-granularity aggregation transformer for joint video-audio-text representation learning. *IEEE Transactions on Circuits and Systems for Video Technology*.

Vaishnavi Himakunthala, Andy Ouyang, Daniel Rose, Ryan He, Alex Mei, Yujie Lu, Chinmay Sonar, Michael Saxon, and William Wang. 2023. Let's think frame by frame with vip: A video infilling and prediction dataset for evaluating video chain-of-thought. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 204–219.

Tobias Höppe, Arash Mehrjou, Stefan Bauer, Didrik Nielsen, and Andrea Dittadi. 2022. Diffusion models for video prediction and infilling. *arXiv preprint arXiv:2206.07696*.

Hongyu Hu, Jiyuan Zhang, Minyi Zhao, and Zhenbang Sun. 2023. Ciem: Contrastive instruction evaluation method for better instruction tuning. *arXiv preprint arXiv:2309.02301*.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Citation: A key to building responsible and accountable large language models. *arXiv preprint arXiv:2307.02185*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

Wen Huang, Hongbin Liu, Minxin Guo, and Neil Zhenqiang Gong. 2024. Visual hallucinations of multi-modal large language models. *arXiv preprint arXiv:2402.14683*.

Atin Sakkeer Hussain, Shansong Liu, Chenshuo Sun, and Ying Shan. 2023. M Θ{2} ugen: Multi-modal music understanding and generation with the power of large language models. *arXiv preprint arXiv:2311.11255*.

Vladimir Iashin and Esa Rahtu. 2020. Multi-modal dense video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 958–959.

Susmit Jha, Sumit Kumar Jha, Patrick Lincoln, Nathaniel D Bastian, Alvaro Velasquez, and Sandeep Neema. 2023. Dehallucinating large language models using formal methods guided iterative prompting. In *2023 IEEE International Conference on Assured Autonomy (ICAA)*, pages 149–152. IEEE.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843.

Pengxu Jiang, Hongliang Fu, Huawei Tao, Peizhi Lei, and Li Zhao. 2019. Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition. *IEEE access*, 7:90368–90377.

Liqiang Jing, Ruosen Li, Yunmo Chen, Mengzhao Jia, and Xinya Du. 2023. Faithscore: Evaluating hallucinations in large vision-language models. *arXiv preprint arXiv:2311.01477*.

Haoqiang Kang and Xiao-Yang Liu. 2023. Deficiency of large language models in finance: An empirical examination of hallucination. *arXiv preprint arXiv:2311.15548*.

Jaeyeon Kim, Jaeyoon Jung, Jinjoo Lee, and Sang Hoon Woo. 2024. Enclap: Combining neural audio codec and audio-text joint embedding for automated audio captioning. *arXiv preprint arXiv:2401.17690*.

Grounding Knowledge. The knowledge alignment problem: Bridging human and external knowledge for large language models.

Sumith Kulal, Tim Brooks, Alex Aiken, Jiajun Wu, Jimei Yang, Jingwan Lu, Alexei A Efros, and Krishna Kumar Singh. 2023. Putting people in their place: Affordance-aware human insertion into scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17089–17099.

Siddique Latif, Moazzam Shoukat, Fahad Shamshad, Muhammad Usama, Heriberto Cuayáhuitl, and Björn W Schuller. 2023. Sparks of large audio models: A survey and outlook. *arXiv preprint arXiv:2308.12792*.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2023. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*.

Juncheng B Li, Jackson Sam Michaels, Laura Yao, Lijun Yu, Zach Wood-Doughty, and Florian Metze. 2023a. Audio-journey: Efficient visual+ llm-aided audio encodec diffusion. In *Workshop on Efficient Systems for Foundation Models@ ICML2023*.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023b. 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.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023c. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Lidong Bing, Shafiq Joty, and Soujanya Poria. 2023d. Chain of knowledge: A framework for grounding large language models with structured knowledge bases. *arXiv preprint arXiv:2305.13269*.

Y Li, R Panda, Y Kim, C Chen, R Feris, D Cox, and N Vasconcelos. 2022. Valhalla: Visual hallucination for machine translation. in 2022 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5206–5216.

Yanda Li, Chi Zhang, Gang Yu, Zhibin Wang, Bin Fu, Guosheng Lin, Chunhua Shen, Ling Chen, and Yunchao Wei. 2023e. Stablellava: Enhanced visual instruction tuning with synthesized image-dialogue data. *arXiv preprint arXiv:2308.10253*.

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

Ning Liao, Shaofeng Zhang, Renqiu Xia, Bo Zhang, Min Cao, Yu Qiao, and Junchi Yan. 2023. Revo-lion: Evaluating and refining vision-language instruction tuning datasets. *arXiv preprint arXiv:2310.06594*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.

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

Hui Liu and Xiaojun Wan. 2023. Models see hallucinations: Evaluating the factuality in video captioning. *arXiv preprint arXiv:2303.02961*.

Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhanhui Kang, and Xirong Li. 2024b. Phd: A prompted visual hallucination evaluation dataset. *arXiv preprint arXiv:2403.11116*.

Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. 2023. Negative object presence evaluation (nope) to measure object hallucination in vision-language models. *arXiv preprint arXiv:2310.05338*.

Jiaying Lu, Jinmeng Rao, Kezhen Chen, Xiaoyuan Guo, Yawen Zhang, Baochen Sun, Carl Yang, and Jie Yang. 2024. Evaluation and enhancement of semantic grounding in large vision-language models. In *AAAI-ReLM Workshop*.

Junyu Luo, Cao Xiao, and Fenglong Ma. 2023. Zero-resource hallucination prevention for large language models. *arXiv preprint arXiv:2309.02654*.

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*.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. *arXiv preprint arXiv:2305.14552*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2023. Generating benchmarks for factuality evaluation of language models. *arXiv preprint arXiv:2307.06908*.

Jonghwan Mun, Linjie Yang, Zhou Ren, Ning Xu, and Bohyung Han. 2019. Streamlined dense video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6588–6597.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. arxiv [cs. cl]. 2023.

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

Hao Ouyang, Tengfei Wang, and Qifeng Chen. 2021. Internal video inpainting by implicit long-range propagation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14579–14588.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback.(2023). *arXiv preprint cs.CL/2302.12813*.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S. M Towhidul Islam Tonmoy, Aman Chadha, Amit P. Sheth, and Amitava Das. 2023a. The troubling emergence of hallucination in large language models – an extensive definition, quantification, and prescriptive remediations.

Vipula Rawte, Anku Rani, Harshad Sharma, Neeraj Anand, Krishnav Rajbangshi, Amit Sheth, and Amitava Das. 2024a. Visual hallucination: Definition, quantification, and prescriptive remediations. *arXiv preprint arXiv:2403.17306*.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023b. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

Vipula Rawte, SM Tonmoy, Krishnav Rajbangshi, Shravani Nag, Aman Chadha, Amit P Sheth, and Amitava Das. 2024b. Factoid: Factual entailment for hallucination detection. *arXiv preprint arXiv:2403.19113*.

Vipula Rawte, SM Tonmoy, SM Zaman, Prachi Priya, Aman Chadha, Amit P Sheth, and Amitava Das. 2024c. " sorry, come again?" prompting–enhancing comprehension and diminishing hallucination with [pause]-injected optimal paraphrasing. *arXiv preprint arXiv:2403.18976*.

Sohini Roychowdhury. 2024. Journey of hallucination-minimized generative ai solutions for financial decision makers. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 1180–1181.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications.

Jaromir Savelka, Kevin D Ashley, Morgan A Gray, Hannes Westermann, and Huihui Xu. 2023. Explaining legal concepts with augmented large language models (gpt-4). *arXiv preprint arXiv:2306.09525*.

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. *arXiv preprint arXiv:2304.09116*.

Yaya Shi, Xu Yang, Haiyang Xu, Chunfeng Yuan, Bing Li, Weiming Hu, and Zheng-Jun Zha. 2022. Emscore: Evaluating video captioning via coarse-grained and fine-grained embedding matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17929–17938.

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

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *arXiv preprint arXiv:2401.06209*.

SM Tonmoy, SM 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 preprint arXiv:2401.01313*.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by actively validating low-confidence generation.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*.

Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li, Wei Li, Jiaqi Wang, et al. 2024. Vigc: Visual instruction generation and correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5309–5317.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. 2023. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*.

Weilun Wu and Yang Gao. 2023. A context-aware model with a pre-trained context encoder for dense video captioning. In *International Conference on Cyber Security, Artificial Intelligence, and Digital Economy (CSAIDE 2023)*, volume 12718, pages 387–396. SPIE.

Xuenan Xu, Zhiling Zhang, Zelin Zhou, Pingyue Zhang, Zeyu Xie, Mengyue Wu, and Kenny Q Zhu. 2023a. Blat: Bootstrapping language-audio pre-training based on audioset tag-guided synthetic data. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 2756–2764.

Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shi-Xiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. 2024a. Secap: Speech emotion captioning with large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19323–19331.

Zhengzhuo Xu, Sinan Du, Yiyan Qi, Chengjin Xu, Chun Yuan, and Jian Guo. 2023b. Chartbench: A benchmark for complex visual reasoning in charts. *arXiv preprint arXiv:2312.15915*.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024b. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Siming Yan, Min Bai, Weifeng Chen, Xiong Zhou, Qixing Huang, and Li Erran Li. 2024. Vigor: Improving visual grounding of large vision language models with fine-grained reward modeling. *arXiv preprint arXiv:2402.06118*.

Shiping Yang, Renliang Sun, and Xiaojun Wan. 2023. A new benchmark and reverse validation method for passage-level hallucination detection. *arXiv preprint arXiv:2310.06498*.

Zhongjie Ye, Helin Wang, Dongchao Yang, and Yuexian Zou. 2021. Improving the performance of automated audio captioning via integrating the acoustic and semantic information. *arXiv preprint arXiv:2110.06100*.

Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. 2023a. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. *arXiv preprint arXiv:2311.13614*.

Yongsheng Yu, Heng Fan, and Libo Zhang. 2023b. Deficiency-aware masked transformer for video inpainting. *arXiv preprint arXiv:2307.08629*.

Yi Yuan, Haohe Liu, Xubo Liu, Qiushi Huang, Mark D Plumbley, and Wenwu Wang. 2024. Retrieval-augmented text-to-audio generation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 581–585. IEEE.

Jian Zhang, Yueting Zhuang, and Fei Wu. 2006. Video-based facial expression hallucination: A two-level hierarchical fusion approach. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 513–521. Springer.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2023a. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*.

Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. 2023b. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*.

Shuo Zhang, Liangming Pan, Junzhou Zhao, and William Yang Wang. 2023c. Mitigating language model hallucination with interactive question-knowledge alignment. *arXiv preprint arXiv:2305.13669*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023d. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. 2024. Mitigating object hallucination in large vision-language models via classifier-free guidance. *arXiv preprint arXiv:2402.08680*.

Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. 2022. Learning video representations from large language models.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*.

Luowei Zhou, Chenliang Xu, and Jason Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Xingyi Zhou, Anurag Arnab, Shyamal Buch, Shen Yan, Austin Myers, Xuehan Xiong, Arsha Nagrani, and Cordelia Schmid. 2024. Streaming dense video captioning. *arXiv preprint arXiv:2404.01297*.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models.

Ge Zhu and Zhiyao Duan. 2024. Cacophony: An improved contrastive audio-text model. *arXiv preprint arXiv:2402.06986*.

9 Appendix

9.1 Table

We have provided a comprehensive summary of the methodologies pertaining to hallucination techniques in large foundational models in Table 1, detailing their approaches to hallucination detection, mitigation, task considerations, datasets utilized, and evaluation metrics employed. This will offer readers a concise overview of recent advancements in this field.

Paper	Detection	Mitigation	Task	Dataset(s)	Evaluation Metric(s)
(Manakul et al., 2023)	Yes	No	QA	Wikibio	Entropy
(Li et al., 2022)	Yes	Yes	QA, Dialog summarization	Halueval	Automatic
(Mündler et al.)	Yes	Yes	Text generation	Manual	F1 Score
(Chen et al., 2023a)	No	Yes	Editing for attribution	MCQ, Dialog	Attribution, Preservation
(Zhang et al., 2023c)	No	Yes	Question knowledge alignment	Fuzzy QA	Attributable to Identified Sources
(Zhang et al., 2023a)	Yes	No	QA	Manual	Accuracy
(Peng et al., 2023)	No	Yes	Task-oriented dialog	News, Customer service	F1 Score, Bleu-4
(Cui et al., 2023)	No	Yes	QA	Manual	Ranking
(Azaria and Mitchell, 2023)	Yes	No	Classification	Manual	Accuracy
(Li et al., 2023d)	Yes	Yes	Knowledge-intensive tasks	Fever, QA	Accuracy
(Elaraby et al., 2023)	Yes	Yes	Consistency, Actuality, QA	Manual NBA domain	Pearson Corelation Coefficient
(Varshney et al.)	Yes	Yes	Text generation	Wikibio	Percentage of mitigated hallucination
(Jha et al., 2023)	Yes	No	Dialog	N/A	N/A
(Pal et al., 2023)	No	No	Reasoning hallucination	Med-Halt	Accuracy, Pointwise Score
(McKenna et al., 2023)	Yes	No	Textual entailment	Altered Directional Inference	Entailment Probability
(Guerreiro et al., 2023)	Yes	Yes	MT	FLores 101, WMT ,TICO	BLEU
(Huang and Chang, 2023)	Yes	Yes	N/A	N/A	N/A
(Luo et al., 2023)	Yes	Yes	Concept extraction	Concept-7	AUC, Accuracy, F1 Score
(Gao et al., 2022)	Yes	Yes	Editing attribution	NQ, SQA	Auto-AIS (Attr$_{auto}$)
(Yang et al., 2023)	Yes	No	Detect factual errors automatically	PHD, WikiBio-GPT3	Precision, Recall, F1 Score, Accuracy
(Min et al., 2023)	Yes	Yes	Fact verification	Manual(Wikipedia)	FActScore
(Rawte et al., 2024b)	Yes	Yes	Factual inaccuracies detection	FACTOID	HV I$_{auto}$
(Ahmad et al., 2023)	Yes	Yes	Hallucination in healthcare	N/A	FActScores
(Ji et al., 2023)	Yes	Yes	Generative and knowledge-intensive	PubMedQA, MEDIQA2019, MedQuAD, and MASH-QA	Unigram F1, ROUGE-L, Med-NLI, and CTRLEval
(Kang and Liu, 2023)	Yes	Yes	Hallucination in finance	N/A	FActScores
(Roychowdhury, 2024)	No	Yes	QA	N/A	N/A
(Savelka et al., 2023)	No	Yes	Factual evaluation in legislation	N/A	N/A
(Dahl et al., 2024)	Yes	No	Legal hallucination	Manual	N/A
(Li et al., 2023f)	Yes	No	Evaluation of object hallucination	MSCOCO	CHAIR, POPE
(Gunjal et al., 2024)	Yes	Yes	VQA	M-Hall Detect	Accuracy
(Dai et al., 2022)	No	Yes	Image captioning	CHAIR	CIDEr
(Lovenia et al., 2023)	Yes	No	Object hallucination	NOPE	METEOR, Exact match accuracy, NegP Accuracy
(Liu et al., 2024b)	Yes	No	Intrinsic vision-language hallucination	PhD	Accuracy
(Zhao et al., 2024)	Yes	Yes	Non-existing object hallucination	MSCOCO	CHAIR, POPE, GPT-4V, recall
(Huang et al., 2024)	Yes	No	Visual hallucination	YNQ, OEQ	Accuracy
(Rawte et al., 2024a)	Yes	No	Video captioning	ActivityNet-Fact, YouCook2-Fact	FactVC
(Wang et al., 2024)	No	Yes	Generate instruction data for vision-language	VIGC-LLaVA-COCO, VIGC-LLaVA-Objects365	Conv, Detail, Complex
(Yu et al., 2023a)	Yes	Yes	Machine-generated visual instruction	LLaVA-Instruction-158K	CHAIR
(Guan et al., 2023)	No	Yes	Visual questions	HallusionBench	Accuracy
(Liu et al., 2023)	Yes	Yes	Vision language	LRV-Instruction	GAVIE
(Xu et al., 2023b)	Yes	No	Evaluation of MLLMs on chart comprehension	ChartBench	Acc+
(Lu et al., 2024)	Yes	Yes	Vision language	MSG-MCQ	Accuracy
(Tong et al., 2024)	Yes	No	Visual question answering	MMVP, VQA	Accuracy
(Liao et al., 2023)	Yes	No	Vision language	REVO-LION	Meta Quality (MQ)
(Hu et al., 2023)	Yes	Yes	Visual captioning, Visual question answering	CIEM	Accuracy, Precision, Recall, F1 Score
(Jing et al., 2023)	Yes	No	Meta-evaluation	LLaVA-1k, MSCOCO-Cap	FAITHSCORE
(Changpinyo et al., 2022)	No	Yes	Multilingual visual question answering	MaXM	Accuracy
(Wang et al., 2023)	Yes	No	Content generation	N/A	Precision, Recall, F1 Score
(Sun et al., 2023)	No	Yes	Visual-language alignment	MMHAL-BENCH	N/A
(Bai et al., 2023)	Yes	No	Evaluate hallucination of vision language model	TouchStone	Hallucination Score
(Zhou et al., 2023)	No	Yes	Hallucination mitigation in LVMs	MSCOCO	CHAIR, BLEU, CLIP
(Yan et al., 2024)	No	Yes	Visual grounding	MMViG	HL, CA, AA, RA, RL, RS, DL
(Zhao et al., 2023)	Yes	Yes	Overcome hallucination in LVMs	POPE, SHR	Accuracy, Precision, F1 Score
(Zhang et al., 2023b)	No	Yes	Image text comprehension and composition	MMBench, SeedBench, QBench, MMBench-CN, Chinese Bench	LR, AR, RR, FP-C, FP-S, CP
(Kulal et al., 2023)	No	Yes	Affordance prediction	Manual	FID, FCKh
(Himakunthala et al., 2023)	No	Yes	Video infilling, Scene prediction	Manual	N/A
(Li et al., 2023c)	No	Yes	Visual dialogue	Manual	N/A
(Zhou et al., 2024)	No	Yes	Video captioning	ActivityNet Captions, YouCook2, ViTT	CIDER, METEOR, SODAc
(Höppe et al., 2022)	Yes	No	Video prediction	BAIR, Kinetics 600, UCF-101	Frechet Video Distance
(Chuang and Fazli, 2023)	No	Yes	Video description	Activity Net Captions, YouCook2	METEOR, ROUGE_L, CIDER, BLEU_4, DIV-2, RE_4
(Li et al., 2023a)	No	Yes	Classification	Manual	Mean avg precsion
(Doh et al., 2023)	No	Yes	Audio captioning	LP- MusicCaps	BLEU
(Xu et al., 2023a)	No	Yes	Caption generation	AudioCaps	R@K, COCO & FENCE
(Liu and Wan, 2023)	No	Yes	Audio captioning	MusciCaps	BLEU
(Nishimura et al., 2024)	Yes	No	Evaluation of LAMs	LAION_CLAP,MS_CLAP	Recall, Precision, F1 Score

Table 1: Overview of the hallucination detection and mitigation landscape in FMs across modalities (Text, Image, Video, and Audio). Each work is categorized based on factors such as detection, mitigation, tasks, datasets, and evaluation metrics.