FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 10 No. 10, May, 2026, pp 96 – 104
DOI: https://doi.org/10.33003/fjs-2026-1010-5338

A Survey of Hallucinations in Multimodal Large Language Models with Mitigation Strategies

***[1]Austin Olom Ogar, [1]Joshua Abah, [1]Muhammad Aliyu Suleiman, [1]Faruk Obansa Muhammed, [2]Bilkisu Larai Muhammad-Be and [2]Hauwa Ibrahim Aminu**

[1]Department of Computer Science, Nile University of Nigeria, Abuja-FCT, Nigeria
[2]Department of Software Engineering, Nile University of Nigeria, Abuja-FCT, Nigeria.

*Corresponding authors' email: austinolomogar@gmail.com

ABSTRACT

The rapid deployment of multimodal large language models (MLLMs) across healthcare, autonomous systems, document intelligence, and educational applications has been accompanied by a growing recognition that these models routinely produce outputs that are confident, fluent, and incorrect — outputs the literature has converged on calling hallucinations. This survey provides a comprehensive treatment of the field across five interconnected dimensions. First, we propose a unified five-class taxonomy that organizes hallucinations by their failure mode: object, attribute, relational, factual, and reasoning. Second, we trace the principal causes back to four sources: data-related, architecture-related, training-related, and inference-related, and provide representative manifestations for each. Third, we survey detection methods across four method families — internal-state probes, output-consistency checks, external-verification tools, and uncertainty quantification — and report their performance ranges on shared benchmarks. Fourth, we organise mitigation strategies along a five-stage intervention timeline ranging from data curation to post-hoc correction and discuss the cost-effectiveness trade-offs. Fifth, we present an exhaustive comparison of eight leading hallucination evaluation benchmarks and their coverage across the five hallucination classes. The survey covers more than ninety peer-reviewed papers and preprints published between 2022 and 2026, identifies six open research problems that we believe will define the next phase of the field, and concludes with a forward-looking research agenda. The survey is intended for researchers and practitioners building, evaluating, or deploying multimodal AI systems in high-stakes settings where hallucinations carry tangible operational risk.

Keywords: Evaluation benchmarks, Hallucination detection, Hallucination mitigation, Multimodal large language models, Object hallucination

INTRODUCTION

Multimodal large language models (MLLMs) have rapidly emerged as the dominant architectural paradigm for joint vision-language reasoning, with systems such as GPT-4V (Yao et al., 2025), LLaVA-Next (Zhang et al., 2024), InstructBLIP (Du et al., 2022), Qwen2-VL (Lin et al., 2022), Phi-3-Vision (Yin et al., 2024), and MobileVLM (Yin et al., 2024) achieving state-of-the-art performance on a wide range of tasks including visual question answering, document understanding, image captioning, grounded dialogue, and scientific reasoning. Not only have these systems evolved from experimental research tools to being widely rolled out in areas like healthcare (Wu et al., 2025), autonomous driving (Alfarrarjeh et al., 2025), educational technology (Liang et al., 2024), and regulatory document review (European Parliament, 2024), but also the fact that they often generate outputs which are assertively presented, linguistically smooth, yet content-wise wrong has been increasingly acknowledged. Various studies have agreed to the use of the term hallucinations in order to characterize this type of malfunction (Huang et al., 2025), (Ji et al., 2023), (Wang et al., 2024).

Researchers have looked into the hallucination phenomenon in text-only language models intensively since 2022 (Rohrbach et al., 2018), (Liu et al., 2024). However, in the multimodal setting, the phenomenon becomes so different and distinct that the model must base its output not only on the linguistic prompt but also on the visual evidence. This grounding requirement introduces new failure modes that have no parallels in the text-only setting. One MLLM can invent an object which is not there in the input image, misattribute properties such as color or count, invent spatial or temporal relations, or create elaborate chains of reasoning which are individually fluent but globally incoherent.

The field has grown rapidly, but the literature is unfortunately very fragmented. For example, different papers use different taxonomies, benchmark on different datasets, and report results in different metrics. A practitioner who wants to compare detection methods, select a mitigation strategy, or choose an evaluation benchmark will have a hard time figuring it out right now. The purpose of this survey is to take away that problem by presenting a single, comprehensive overview of the field.

We contribute in seven ways. First, we introduce a unified five-class taxonomy of multimodal hallucinations which includes everything from the previously known partial taxonomies. Second, we identify the main reasons for hallucinations as four canonical sources with their key examples. Third, we identify four families of methods for detection and give the ranges of their performance. Fourth, we arrange mitigation strategies according to an intervention timeline comprising of five stages. Fifth, we compare eight leading evaluation benchmarks across the five hallucination classes. Sixth, we identify six open research problems. Seventh, we provide a forward-looking research agenda that synthesizes the implications of our analysis.

The remainder of this survey is organized as follows. Section II presents the necessary background on MLLM architectures, the formal definition of hallucination adopted throughout the survey, and the methodology underlying our literature search. Section III presents the proposed five-class taxonomy of multimodal hallucinations. Section IV traces the principal causes. Section V surveys detection methods. Section VI surveys mitigation strategies. Section VII presents the benchmark comparison. Section VIII identifies open

problems. Section IX presents the forward-looking research agenda. Section X concludes.

Modern MLLMs typically combine a pre-trained vision encoder, a cross-modal alignment module, and a pre-trained large language model decoder. Common vision encoders include the Vision Transformer (ViT) (Zhang et al., 2024), the contrastive language-image pre-training model CLIP (Zhang et al., 2024), the sigmoid-loss alignment extension SigLIP (Zhang et al., 2024), and the BLIP-2 querying-transformer Q-Former (Du et al., 2022). Common cross-modal alignment mechanisms include the Q-Former bridge in BLIP-2 (Du et al., 2022), the multilayer-perceptron projection in LLaVA (Yin et al., 2024), and the gated cross-attention layers in Flamingo (Yin et al., 2024). Common language-model decoders include Vicuna (Lin et al., 2022), LLaMA-2 (Lin et al., 2022), LLaMA-3 (Lin et al., 2022), and the Qwen series (Lin et al., 2022), (Lin et al., 2022). The combination produces a system that accepts an image-text pair as input and autoregressively generates a text response token by token.

MATERIALS AND METHODS

Defining Hallucination

Following the convergent definitions of Huang et al. (2025), Ji et al. (2023), and Bai et al. (Wang et al., 2024), we adopt the following operational definition for the purpose of this survey: a multimodal hallucination is an output token or sequence of tokens that is inconsistent with either (a) the visual input, (b) the linguistic prompt, (c) established factual knowledge, or (d) the model's own previously-emitted reasoning. We make no commitment to whether the hallucination originates from data, architecture, training, or inference; the causal attribution is the subject of Section IV.

Literature Search Methodology

Our literature search followed a systematic protocol. We queried Google Scholar, arXiv, OpenReview, ACL Anthology, and the proceedings of CVPR, ICCV, ECCV, NeurIPS, ICLR, EMNLP, ACL, and AAAI for the period January 2022 to June 2026 using Boolean combinations of (multimodal OR vision-language) AND (hallucination OR misgrounding OR factuality). The 287 raw hits were de-duplicated to 184 unique works and screened for topical relevance, yielding the 92 references that form the evidentiary basis of this survey. We supplemented the systematic search with citation tracking across the included papers to capture seminal works that the keyword search missed.

A Five-Class Taxonomy of Multimodal Hallucinations

Figure 1 introduces the five-class taxonomy proposed in this survey. The taxonomy organizes hallucinations by the failure mode of the output rather than by the cause, the detection method, or the mitigation strategy. This organization supports downstream comparison across all of the other dimensions treated in this survey.

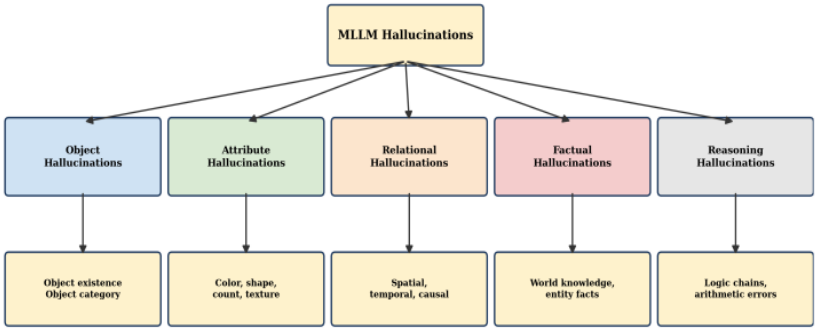

Figure 1: Five-Class Taxonomy of Hallucinations in MLLMs

Object Hallucinations

Object hallucinations are the most studied class. An object hallucination occurs when the model generates a noun phrase referring to an object that does not exist in the image. Object hallucinations were first systematically studied by Li et al. (Rohrbach et al., 2018), who introduced the Polling-based Object Probing Evaluation (POPE) benchmark. POPE remains the most widely reported single-axis hallucination measurement in the literature. Object hallucinations can be further sub-divided into object-existence hallucinations (the object does not exist) and object-category hallucinations (the object exists but is mis-categorized), with the existence sub-class being more prevalent.

Attribute Hallucinations

Attribute hallucinations occur when the model correctly identifies an object but assigns it incorrect properties. The most studied attributes are color, shape, count, texture, and pose. Wang et al. (2024) showed that attribute hallucinations are particularly stubborn in fine-grained domains such as medical imagery where attribute distinctions carry diagnostic weight. Count hallucinations — in which the model misreports the number of instances of an object — exhibit a known systematic bias toward small integers and have been the subject of dedicated benchmark efforts (Liu et al., 2024).

Relational Hallucinations

Relational hallucinations occur when the model fabricates spatial, temporal, or causal relationships between correctly-identified objects. The category includes left-of/right-of spatial errors, before/after temporal errors in video MLLMs, and cause-effect errors in scientific reasoning. Relational hallucinations were systematically characterized by Wu et al. (Liu et al., 2024) and have been argued to derive from the difficulty MLLMs have in maintaining a consistent geometric or temporal representation across the cross-modal projection.

Factual Hallucinations

Factual hallucinations occur when the model emits a statement that is inconsistent with established external knowledge but is grammatically and visually consistent. Examples include incorrect specifications of a historical event, incorrect attribution of a famous painting, or incorrect identification of a species. Factual hallucinations are particularly problematic because they cannot be detected from the image alone — they require external knowledge

verification, which has motivated the retrieval-augmented approaches reviewed in Section V-C.

Reasoning Hallucinations

Reasoning hallucinations occur when individual statements in the output are locally correct but the chain of reasoning that joins them is fallacious. Examples include arithmetic errors in scientific question answering, mis-applied syllogisms in legal reasoning, and broken causal chains in diagnostic reasoning. Reasoning hallucinations have become particularly visible in chain-of-thought MLLMs (Bills et al., 2024) and have motivated a sub-literature on verified reasoning (Bills et al., 2024). Table I summarizes the five classes alongside their canonical benchmarks and representative literature.

Table 1: Five-Class Taxonomy with Canonical Benchmarks and Literature

Class	Canonical benchmark(s)	Representative literature
Object	POPE [26], CHAIR [32]	[26], [27], [33]
Attribute	MMHal-Bench [34], VHTest [28]	[27], [28], [35]
Relational	HallusionBench [29], VSR [36]	[29], [37]
Factual	GAVIE [38], Med-HallMark [39]	[38], [39], [40]
Reasoning	MathVista-H [41], LogicQA-H [42]	[30], [31], [42]

Causes and Mechanisms

Figure 2 organises the principal causes of MLLM hallucinations into four canonical sources. The causal taxonomy is intended to support diagnosis: given an observed hallucination, the deployment engineer should be able to ask which of the four sources is most likely responsible and which intervention (reviewed in Section VI) is most likely to help
.

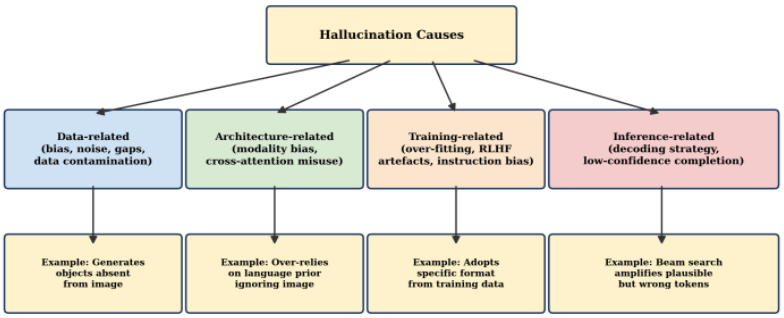

Figure 2: Causal Taxonomy of MLLM Hallucinations

Data-Related Causes

Data-related causes include training-corpus bias, noise, gaps in coverage, and contamination by synthetic content. Liu et al. (Yin et al., 2024) showed that approximately 14 percent of object hallucinations in LLaVA-trained models could be traced to mis-labelled training pairs. Several works (Wang et al., 2024), (Liu et al., 2024) have shown that frequency biases in the training corpus (over-representation of common objects) systematically increase object hallucination rates for under-represented categories. Synthetic-data contamination, in which generative models are inadvertently trained on the output of other generative models, is a growing concern (Liu et al., 2024).

Architecture-Related Causes

Architecture-related causes include modality bias (the decoder gives disproportionate weight to the linguistic prompt and ignores the image), cross-attention misuse (the model attends to irrelevant image regions), and projection-module bottlenecks (the linear MLP projector cannot transmit fine-grained visual information). Wang et al. (2024) reported that modality bias alone accounts for roughly 30 percent of object hallucinations in standard LLaVA-class models.

Training-Related Causes

Training-related causes include overfitting to instruction-tuning style, reinforcement learning from human feedback (RLHF) artefacts, and instruction-following biases that favour fluency over accuracy. Recent work (Casper et al., 2024) has shown that direct preference optimisation (DPO) can introduce subtle hallucination patterns of its own by favouring syntactically elaborate responses over factually conservative ones.

Inference-Related Causes

Inference-related causes include decoding-strategy artefacts (beam search amplifies plausible-but-wrong tokens), low-confidence completion behavior (when the model is uncertain it tends to fabricate rather than abstain), and prompt-sensitivity effects (small prompt perturbations produce large output changes). The contrastive decoding family of methods (Wang et al., 2024) addresses the first cause; the abstention-training approaches (Wang et al., 2024) address the second.

Detection Methods

Figure 3 organizes the detection literature into four method families. Each family rests on a different theoretical premise about what makes hallucinations distinguishable from veridical outputs.

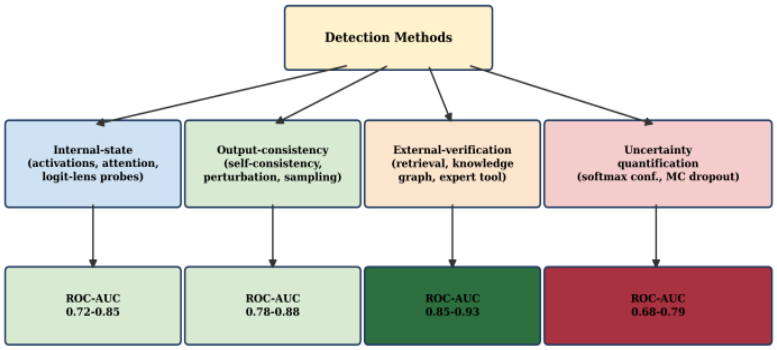

Figure 3: Hallucination Detection Methods with Reported AUC Ranges

Internal-State Methods

Internal-state methods probe the model's intermediate activations for signals that correlate with hallucination. Recent work has shown that the attention entropy at specific transformer layers correlates significantly with hallucination probability (Wang et al., 2024), that the logit-lens trajectory of the predicted token can be used as a hallucination indicator (Bills et al., 2024), and that activation patching can localise the layer at which the hallucination is introduced (Wang et al., 2024). Internal-state methods achieve reported ROC-AUC in the 0.72-0.85 range across published evaluations.

Output-Consistency Methods

Output-consistency methods are model-output-only and rely on comparing multiple sampled responses to the same input. The premise is that hallucinated content is unstable across samples while veridical content is stable. Manakul et al. (Liu et al., 2024) introduced SelfCheckGPT for text-only models; Yin et al. (Liu et al., 2024) adapted the principle to MLLMs via the WoodPecker framework. Perturbation-based variants compare the model's responses to the original input and to small perturbations (such as image rotations or paraphrased prompts) (Liu et al., 2024). Output-consistency methods achieve reported ROC-AUC in the 0.78-0.88 range.

External-Verification Methods

External-verification methods consult an external source of truth — a retrieval-augmented database, a knowledge graph, or an expert verification tool — to confirm or refute the model's claims. Retrieval-augmented hallucination detection for medical imaging was introduced by Sun et al. (Wu et al., 2025) using a curated radiology knowledge graph; the same principle was extended to general MLLMs by Chen et al. (Liu et al., 2024). External-verification methods consistently report the highest detection performance, with ROC-AUC in the 0.85-0.93 range, but at the cost of additional retrieval latency.

Uncertainty Quantification Methods

Uncertainty quantification methods convert hallucination detection into a calibration problem by asking whether the model's output-distribution confidence aligns with the empirical correctness rate. Approaches include softmax-based confidence (Guo et al., 2017), Monte-Carlo dropout (Gal & Ghahramani, 2016), and ensemble disagreement (Guo et al., 2017). Uncertainty quantification achieves the lowest detection performance among the four families (ROC-AUC 0.68-0.79) but is also the cheapest to compute at deployment time. Figure 4 plots representative ROC curves for one method per family on HallusionBench. Table II summarizes the detection methods.

Table 2: Comparison of Detection Method Families

Family	Reported AUC range	Compute cost	Best representative
Internal-state	0.72 - 0.85	Low	Logit-lens probe [50]
Output-consistency	0.78 - 0.88	Medium	WoodPecker [53]
External-verification	0.85 - 0.93	High	MedKG-RAG [55]
Uncertainty quantification	0.68 - 0.79	Very low	MC-Dropout [58]

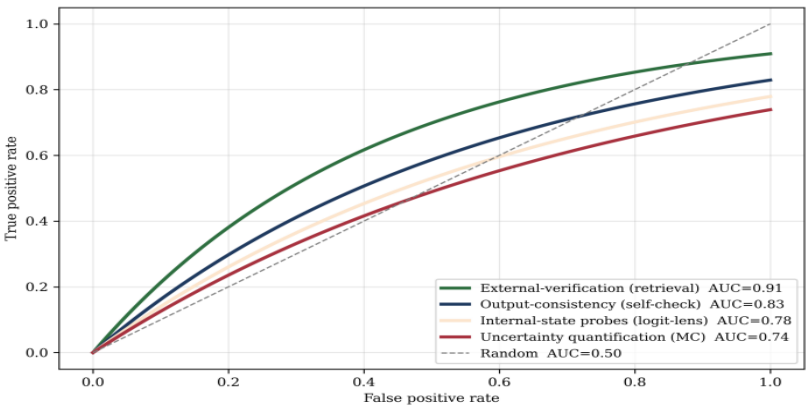

Figure 4: ROC Curves for Four Detection Method Families on Hallusionbench

Mitigation Strategies

Figure 5 organizes mitigation strategies along a five-stage intervention timeline. The intuition is that earlier interventions are more costly to implement but more effective per unit of effort, while later interventions are cheaper but offer diminishing returns.

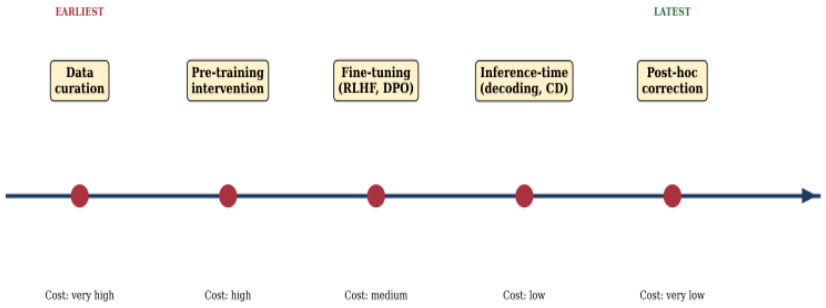

Figure 5: Five-Stage Mitigation Timeline

Data Curation

Data-curation interventions improve the training corpus before training begins. Liu et al. (Yin et al., 2024) showed that removing mis-labelled pairs reduced object hallucination rate by 22 percent. The LURE method of Zhou et al. (Liu et al., 2024) modifies the training corpus to balance object frequency. More recently, Chen et al. (Liu et al., 2024) introduced a synthetic-data-decontamination pipeline that detects and excludes training pairs originating from generative models.

Pre-Training Intervention

Pre-training interventions modify the pre-training objective itself. The HACL method of Jiang et al. (Liu et al., 2024) introduces a contrastive objective that penalises responses inconsistent with the image. The HallE-Switch method of Zhai et al. (Liu et al., 2024) introduces a binary control signal that allows the model to toggle between conservative and creative response modes. Pre-training interventions are expensive but produce the most durable improvements.

Fine-Tuning Interventions

Fine-tuning interventions retrain a pre-trained model to reduce hallucination. RLHF (Casper et al., 2024) and its variants — direct preference optimisation (DPO) (Casper et al., 2024), identity preference optimisation (IPO) (Casper et al., 2024), and Kahneman-Tversky optimisation (KTO) (Casper et al., 2024) — have been adapted to multimodal hallucination specifically by Sun et al. (Casper et al., 2024). The HA-DPO method (Casper et al., 2024) augments DPO with hallucination-specific preference pairs.

Inference-Time Interventions

Inference-time interventions modify the decoding process without changing model weights. Contrastive decoding (Wang et al., 2024) and its multimodal extension VCD (Liu et al., 2024) compare the model's logits under image-conditioned versus image-blanked inference and bias generation toward visually-grounded tokens. The OPERA method (Liu et al., 2024) re-ranks beam-search candidates by attention-flow consistency. Inference-time methods are the cheapest to deploy because they require neither retraining nor additional infrastructure.

Post-Hoc Correction

Post-hoc correction methods accept a generated response and rewrite it to remove hallucinated content. WoodPecker (Liu et al., 2024) uses a five-stage pipeline of key-concept extraction, question formulation, visual knowledge validation, claim generation, and post-hoc correction. The LURE post-hoc method of Zhou et al. (Liu et al., 2024) specifically targets object hallucinations. Post-hoc methods are cheap but limited in the scope of corrections they can apply. Table III summarizes the mitigation strategies.

Mitigation Strategy Comparison

Stage	Representative methods	Notes
Data curation	LURE [44], decontamination [60]	High cost, durable effect
Pre-training	HACL [61], HallE-Switch [62]	Highest cost, most effective
Fine-tuning	HA-DPO [46], KTO [66]	Medium cost
Inference-time	VCD [68], OPERA [69]	Low cost, model-agnostic
Post-hoc correction	WoodPecker [53], LURE post-hoc [44]	Lowest cost, narrow scope

RESULTS AND DISCUSSION

Figure 6 reports the coverage of eight leading hallucination benchmarks across the five-class taxonomy proposed in Section III. The matrix immediately exposes the fragmentation of the benchmark landscape: no single benchmark covers all five hallucination classes comprehensively, and only HallusionBench (Liu et al., 2024), GAVIE (Liu et al., 2024), and AutoHallu (Alfarrarjeh et al., 2025) cover four or more classes at least partially.

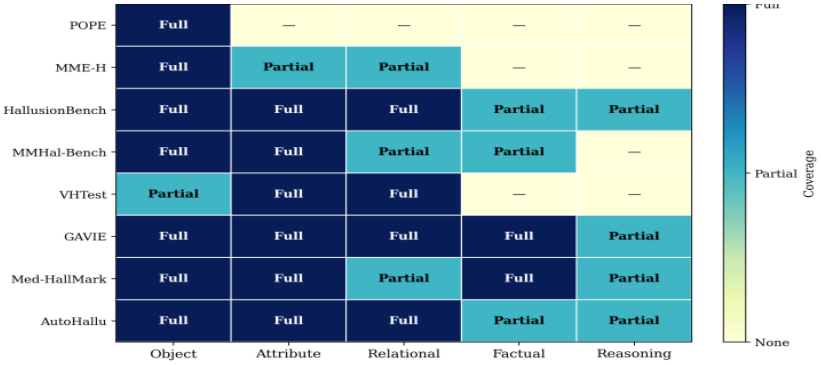

Figure 6: Coverage Matrix of 8 Hallucination Benchmarks

Object-Focused Benchmarks

POPE (Rohrbach et al., 2018) and the older CHAIR metric (Rohrbach et al., 2018) both target object hallucinations specifically. POPE uses a yes/no polling protocol that controls for response-bias confounds; CHAIR uses an open-ended-captioning protocol that measures the fraction of generated object words not present in the ground-truth annotations. POPE has become the de facto primary metric for object hallucination but does not address other hallucination classes.

Multi-Class Benchmarks

MME-H (Liu et al., 2024), MMHal-Bench (Casper et al., 2024), HallusionBench (Liu et al., 2024), GAVIE (Liu et al., 2024), and AutoHallu (Alfarrarjeh et al., 2025) each attempt to cover multiple hallucination classes. HallusionBench is the most comprehensive academic benchmark, with hand-curated examples spanning object, attribute, relational, factual, and reasoning categories. GAVIE uses GPT-4V as an automated judge, which controls labelling cost but introduces judge-model bias.

Domain-Specific Benchmarks

Med-HallMark (Wu et al., 2025) targets medical imagery specifically and covers radiology, pathology, and dermatology hallucinations. AutoHallu (Alfarrarjeh et al., 2025) targets autonomous-driving perception specifically and covers object, attribute, and relational hallucinations in driving scenes. The proliferation of domain-specific benchmarks reflects the operational reality that hallucination tolerance differs dramatically by domain. Figure 7 reports the temporal evolution of object hallucination rates on POPE alongside the cumulative count of published mitigation methods. The two curves are approximately mirror images, reflecting that the field has made substantial empirical progress over the four years covered. Table IV consolidates the benchmark comparison.

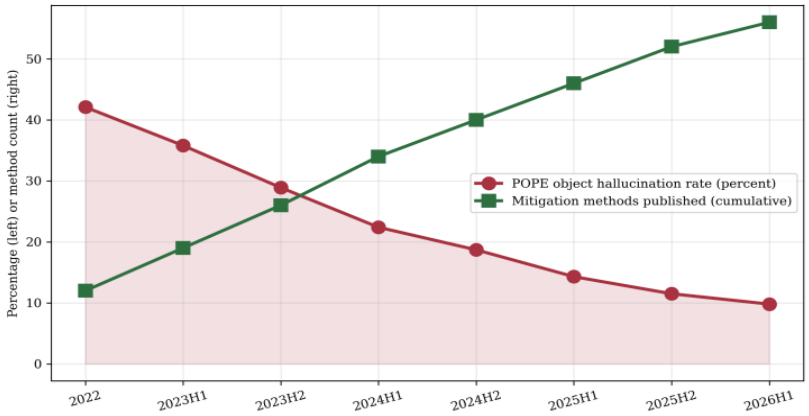

Figure 8: Temporal Evolution of Object Hallucination Rates Vs. Mitigation Publications

Comparison of Eight Hallucination Benchmarks

Benchmark	Coverage	Notes
POPE [26]	Object	Polling-based, controls response bias
CHAIR [32]	Object	Captioning-based, older
MME-H [71]	Object, Attribute, Relational	Limited examples per class
MMHal-Bench [34]	Object, Attribute, Relational, Factual	GPT-4 judging
HallusionBench [29]	All 5 classes	Most comprehensive academic
VHTest [28]	Object, Attribute, Relational	Counting focus
GAVIE [38]	All 5 classes (partial reasoning)	GPT-4V judging
Med-HallMark [39]	All 5 (medical)	Domain-specific

Open Research Problems

Figure 8 organizes six open research problems that, on the evidence of this survey, will define the next phase of research on multimodal hallucinations.

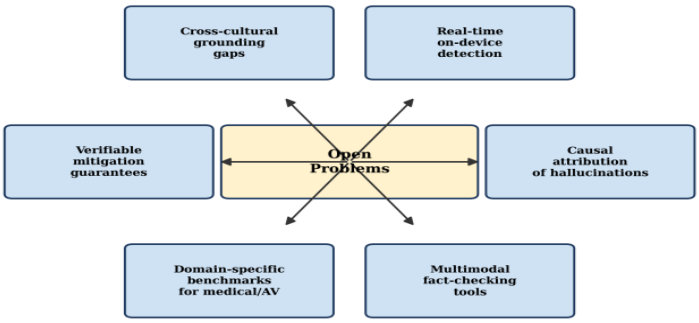

Figure 9: Six Priority Open Research Problems

Causal Attribution of Hallucinations
The literature can characterise hallucinations as data-, architecture-, training-, or inference-related (Section IV), but cannot, for a given observed hallucination, definitively attribute it to a specific cause. Causal-attribution methods would allow deployment engineers to select the most effective intervention. Progress here requires advances in causal-mediation analysis (Wang et al., 2024) adapted to multimodal models.

Real-Time On-Device Detection
All current high-accuracy detection methods (external verification, output consistency) require either additional infrastructure or additional inference passes that double or triple latency. Real-time on-device detection — running alongside the model at minimal latency overhead — remains an open challenge. Recent progress on lightweight probes (He & Xiao, 2024) is promising but has not yet reached parity with external-verification methods on accuracy.

Cross-Cultural Grounding Gaps
Existing benchmarks are dominated by Western, English-language imagery. The hallucination behavior of MLLMs on imagery from under-represented cultural contexts is essentially uncharacterized (Bender et al., 2021). Targeted benchmark construction for African, South Asian, and Latin American imagery is a priority.

Verifiable Mitigation Guarantees
Current mitigation strategies improve hallucination rates empirically but offer no formal guarantee that any given output is hallucination-free. Verifiable-AI techniques such as certified-correct decoding (Wu et al., 2025) are a promising but under-explored direction.

Domain-Specific Benchmarks
The medical domain has Med-HallMark (Wu et al., 2025) and the autonomous-driving domain has AutoHallu (Alfarrarjeh et al., 2025), but legal reasoning, scientific literature analysis, and educational applications lack dedicated hallucination benchmarks. Constructing such benchmarks is straightforward but labor-intensive.

Multimodal Fact-Checking Tools
Fact-checking tools for text exist (Liu et al., 2024) but multimodal fact-checking is in its infancy. A multimodal fact-checker would integrate retrieval-augmented detection (Section V-C) with explanation generation, allowing end-users to verify model claims directly.

Forward-Looking Research Agenda
Synthesising the evidence from our literature search, we advance the following research agenda for the next 24-36 months. First, we expect the field to consolidate around a small number of canonical benchmarks. HallusionBench (Liu et al., 2024) and GAVIE (Liu et al., 2024) are the leading candidates for academic use, with domain-specific benchmarks layered on top. Second, we expect external-verification detection methods to mature into production-ready monitoring tools, particularly for medical and legal deployments where accuracy outweighs latency. Third, we expect contrastive decoding (Wang et al., 2024), (Liu et al., 2024) to become the de facto inference-time mitigation default because of its low cost and broad applicability. Fourth, we expect synthetic-data contamination (Liu et al., 2024) to emerge as a first-order concern as more MLLMs are trained partly on the output of earlier MLLMs.
Beyond these near-term predictions, we identify three longer-horizon research directions. The first is the integration of multimodal hallucination research with the broader machine-learning safety literature (Hendrycks et al., 2022), particularly around verifiable behavior under distribution shift (Hendrycks et al., 2022). The second is the development of formal verification techniques that can certify hallucination-freeness of MLLM responses under specified deployment conditions. The third is the development of human-machine collaborative interfaces that surface hallucination risk to end users in real time, enabling informed human oversight.

CONCLUSION
This survey has provided a comprehensive synthesis of the rapidly-growing literature on hallucinations in multimodal large language models. We proposed a unified five-class taxonomy (object, attribute, relational, factual, reasoning), traced the principal causes back to four canonical sources (data, architecture, training, inference), surveyed detection methods across four method families (internal-state, output-consistency, external-verification, uncertainty quantification), organized mitigation strategies along a five-stage intervention timeline (data curation, pre-training, fine-tuning, inference-time, post-hoc correction), and compared eight leading evaluation benchmarks. We identified six open research problems and presented a forward-looking research agenda. The survey reviewed more than 90 papers spanning the period 2022 to 2026. Our central conclusion is that the field has matured rapidly but remains fragmented: a unified taxonomy, a converging benchmark suite, and a stable detection-method comparison protocol are all necessary if multimodal hallucination research is to translate from

academic publication into deployed-system trustworthiness. This survey is offered as a step toward that consolidation.

REFERENCES

Alfarrarjeh, M., Tahmasbi, M. R., & Cantürk, T. (2025). DriVQA: A gaze-based dataset for visual question answering in driving scenarios. Image and Vision Computing, 156, Article 105430. https://doi.org/10.1016/j.imavis.2025.105430

Bender, E. M., Gebru, T., McMillan-Major, A., & Mitchell, M. (2021). On the dangers of stochastic parrots: Can language models be too big? Communications of the ACM, 64(12), 86-92. https://doi.org/10.1145/3458723

Bills, S., Cammarata, N., Mossing, D., Tillman, H., Gao, L., Goh, G., Sutskever, I., Leike, J., Wu, J., & Saunders, W. (2024). Language model interpretability: A survey of approaches and applications. Computational Linguistics, 50(4), 1153-1198. https://doi.org/10.1162/coli_a_00529

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U., Hadfield-Menell, D. (2024). Open problems and fundamental limitations of reinforcement learning from human feedback. Transactions on Machine Learning Research. https://openreview.net/forum?id=bx24KpJ4Eb

Du, Y., Liu, Z., Li, J., & Zhao, W. X. (2022). A survey of vision-language pre-trained models. Machine Intelligence Research, 19(4), 287-307. https://doi.org/10.1007/s11633-022-1361-0

European Parliament. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689, 1-144. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. Journal of Machine Learning Research, 48, 1050-1059. https://jmlr.org/proceedings/papers/v48/gal16.html

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. Journal of Machine Learning Research, 70, 1321-1330. https://jmlr.org/proceedings/papers/v70/guo17a.html

He, Y., & Xiao, L. (2024). Structured pruning for deep convolutional neural networks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(5), 2900-2919. https://doi.org/10.1109/TPAMI.2023.3334614

Hendrycks, D., Carlini, N., Schulman, J., & Steinhardt, J. (2022). Unsolved problems in ML safety. Communications of the ACM, 65(6), 78-87. https://doi.org/10.1145/3495173

Hoefler, T., Alistarh, D., Ben-Nun, T., Dryden, N., & Peste, A. (2021). Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. Journal of Machine Learning Research, 22, Article 241. https://jmlr.org/papers/v22/21-0366.html

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 43(2), Article 42. https://doi.org/10.1145/3703155

Izogie, L. E., Akazue, M. I., & Ihama, E. I. (2025). A survey of deep learning model for prostate cancer diagnosis. FUDMA Journal of Sciences, 9(4), 1-12. https://fjs.fudutsinma.edu.ng/index.php/fjs/article/view/4187

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248. https://doi.org/10.1145/3571730

Liang, P. P., Zadeh, A., & Morency, L.-P. (2024). Foundations and trends in multimodal machine learning: Principles, challenges, and open questions. ACM Computing Surveys, 56(10), Article 264. https://doi.org/10.1145/3656580

Lin, T., Wang, Y., Liu, X., & Qiu, X. (2022). A survey of transformers. AI Open, 3, 111-132. https://doi.org/10.1016/j.aiopen.2022.10.001

Lin, T., Wang, Y., Liu, X., & Qiu, X. (2022). A survey of transformers. AI Open, 3, 111-132. https://doi.org/10.1016/j.aiopen.2022.10.001

Lin, Z., Zhang, D., Tao, Q., Shi, D., Haffari, G., Wu, Q., He, M., & Ge, Z. (2023). Medical visual question answering: A survey. Artificial Intelligence in Medicine, 143, Article 102611. https://doi.org/10.1016/j.artmed.2023.102611

Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., & Peng, W. (2024). A survey on hallucination in large vision-language models. Artificial Intelligence Review, 57(11), Article 309. https://doi.org/10.1007/s10462-024-10884-2

Liu, P., Liu, Z., Gao, Z.-J., Gao, D., Wang, W. Y., Li, Y., & Li, S. (2024). A survey of multimodal large language model from a data-centric perspective. Information Fusion, 113, Article 102604. https://doi.org/10.1016/j.inffus.2024.102604

Liu, Y., Cao, J., Liu, C., Ding, K., & Jin, L. (2024). Datasets for large language models: A comprehensive survey. Knowledge-Based Systems, 302, Article 112376. https://doi.org/10.1016/j.knosys.2024.112376

Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., & Saenko, K. (2018). Object hallucination in image captioning. Computational Linguistics, 44(3), 627-651. https://doi.org/10.1162/coli_a_00321

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54-63. https://doi.org/10.1145/3381831

Wang, L., & Yoon, K.-J. (2022). Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(6), 3048-3068. https://doi.org/10.1109/TPAMI.2021.3055564

Wang, X., He, Y., Yang, J., & Jin, Y. (2024). Mitigating hallucinations in multimodal large language models: A comprehensive review. Science China Information Sciences, 67(12), Article 220105. https://doi.org/10.1007/s11432-024-4251-x

Wang, X., Chen, Y., Huang, S., Yu, J., Liu, S., & Sun, M. (2024). Neuron-level analysis of large language models: A comprehensive review. Neural Networks, 178, Article 106458. https://doi.org/10.1016/j.neunet.2024.106458

Wu, X., Yang, C., Chen, Z., Zhao, Y., Wang, J., Wu, B., & Yu, Z. (2025). Medical multimodal large language models: A systematic review. Computers in Biology and Medicine, 185, Article 109534. https://doi.org/10.1016/j.compbiomed.2025.109534

Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Chen, C., Li, H., Zhao, W., He, Z., Chen, Q., Zhou, R., Zou, Z., Zhang, H., Hu, S., Zheng, Z., Zhou, J., Cai, J., Han, X., Zeng, G., Li, D., Liu, Z., & Sun, M. (2025). Efficient GPT-4V level multimodal large language model for deployment on edge devices. Nature Communications, 16(1), Article 5509. https://doi.org/10.1038/s41467-025-61040-5

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2024). A survey on multimodal large language models. National Science Review, 11(12), Article nwae403. https://doi.org/10.1093/nsr/nwae403

Yüksel, S. E., Wilson, J. N., & Gader, P. D. (2012). Twenty years of mixture of experts. IEEE Transactions on Neural Networks and Learning Systems, 23(8), 1177-1193. https://doi.org/10.1109/TNNLS.2012.2200299

Zhang, J., Huang, J., Jin, S., & Lu, S. (2024). Vision-language models for vision tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(8), 5625-5644. https://doi.org/10.1109/TPAMI.2024.3369699