Spectral-Check: A Graph-Theoretic Framework for Joint Cross-Modal and Cross-Lingual Hallucination Detection

1st Saim Nadeem
School of Computing
FAST NUCES
Islamabad, Pakistan
i221884@nu.edu.pk

2nd Abdur Rehman
School of Computing
FAST NUCES
Islamabad, Pakistan
i221963@nu.edu.pk

3rd Muneeb Ahmad
School of Computing
FAST NUCES
Islamabad, Pakistan
i221889@nu.edu.pk

4th Qurat ul Ain
School of Computing
FAST NUCES
Islamabad, Pakistan
quratul.ain@isb.nu.edu.pk

***Abstract*—While both LLMs and MLLMs have shown remarkable performance on a variety of generative tasks, they are also susceptible to hallucinations that are both false and unsupported by the surrounding visual or textual context. The challenge is even greater in cross-modal and cross-lingual contexts, where correspondence between multilingual texts and their visual features is usually weakened. State-of-the-art detection methods for hallucinations, such as LLM Check, rely on scalar statistics based on simplified attention representations that reduce complex dependency structures to diagonal log probabilities. We believe such types of reduction hide structural anomalies typical of hallucinated generations.**

As a solution to tackle this issue, we introduce *Spectral-Check*; graph-theoretic representation of cross-modal attention maps as information-flow networks and analysis of their normalized Laplacian spectra. Our approach, which relies on the study of eigenvalue distributions, Von Neumann entropy, and associated spectral signatures in a vision-language proxy model (Qwen2-VL) identifies failures to align examples which it alone does not identify by the use of confidence scores. The experiments on CCHall benchmark indicate that Spectral-Check has both high accuracy and ROC-AUC of 71.00 percent as well as 0.7512, respectively, as compared to the baseline logistic regression models. These findings suggest a strong and computationally efficient predictor of multimodal hallucinations detection by spectral analysis of the internal attention dynamics.

***Index Terms*—Hallucination detection, multimodal large language models, spectral graph analysis, cross-modal alignment, cross-lingual grounding.**

I. Introduction

The swift development of Large Language Models (LLMs) and Multimodal Large Language Model (MLLMs) has changed the natural language processing and vision-language reasoning. Different systems like GPT-4V, LLaMA-3 and Qwen-VL have been improving the performance in different tasks like multilingual generation, visual question answering, and multimodal reasoning among other tasks in an impressive manner [1]. With the growing use of these models in sensitive fields such as medical decision support to real time translation the accuracy of their results becomes vital. Their vulnerability to hallucinations can be seen as a long-standing problem: fluent yet unsubstantiated or not based on the presented text or visual evidence that does not correlate with the given pieces of evidence [2].

The hallucinations are particularly acute in cross-lingual and cross-modal. The model is also required to accurately extract visual features and match them with the right semantic constructions of the target language when producing descriptions of visual scenes in a non-English language [3]. Failure of this alignment process is reflected in hallucination of objects, non-occurrence of entities, or inaccurate descriptions of relation, all of which place trust in the real world application at risk and endanger safety in real-world applications as well [4].

The existing hallucination detection methods are mainly based on the consistency-based sampling or the uncertainty based observeance. Different consistency methods, including SelfCheckGPT, are computationally infeasible in real-time or resource-starved environments, and they need multiple generations to execute them (after) their release occurs [5]. When relying on uncertainty, such as the approach to detect ungrounded content in the form of the probability distributions or internal attention scores, the methods are known as uncertainty based approaches, such as the case of the method of detecting ungrounded content called the LLM-Cheque methodology described in [6]. These methods also have, however, a severe shortcoming: They experience a scalar collapse. They are eliminating the relational dependencies between tokens and are not capturing structural abnormalities, which indicate grounding failures, by converting inherently structured attention matrices into single scalar scores, say, diagonal log-determinants.

To remedy this, we present *Spectral-Check*, a graph-theoretic framework to investigate the structural integrity of cross-modal attention patterns. Instead of treating the attention matrix as a collection of independent coefficients, we map it onto the adjacency matrix of a directed information flow graph and calculate its spectral features, such as eigenvalue distributions and Von Neumann entropy, of the Graph Laplacian. This provides more detailed insight into internal model dynamics and identifies anomalies that scalar metrics fail to capture. We evaluate our method on the challenging CCHall benchmark

using the Qwen2-VL architecture.

The main contributions of this work are as follows:

1) We introduce a graph-theoretic formulation of cross-modal attention that allows for hallucination detection using spectral signatures derived from Laplacian eigenvalues.
2) We propose a joint cross-modal and cross lingual detection pipeline which fuses spectral features with confidence scores, improving over scalar based baselines.
3) Extensive experiments are conducted on the CCHall dataset, with an accuracy of 71.00

The rest of the paper is organized as follows: Sect. II presents related work. Sect. III discusses the problem formulation. Sect. IV elaborates on the proposed methodology. Sect. V and VI present the experimental setup and results. Sect. VII discusses the findings, and Sect. VIII concludes the study.

II. Related Work

The field of hallucination detection has expanded rapidly since 2023. This section categorizes recent advancements into scalar-based metrics, consistency checks, and emerging structural analyses. A summary of these works is provided in Table I.

Scalar and Uncertainty Metrics: Early works focused on token-level probability. Manakul et al. [5] introduced metrics based on the entropy of the output distribution. While efficient, these methods often fail when the model is "confidently wrong." The base paper, Sriramanan et al. [6], proposed *LLM-Check*, utilizing the log-determinant of attention kernels. While an improvement over simple perplexity, it simplifies the attention map to its diagonal, ignoring cross-token dependencies. Similarly, Varshney et al. [7] explored projecting attention scores into lower dimensions but retained a scalar focus.

Consistency and Retrieval: Consistency-based methods, such as those by Chen et al. [8], analyze the variance across multiple sampled outputs. While effective, the computational cost is prohibitive for low-latency applications. Retrieval-Augmented Generation (RAG) evaluation frameworks, like RAGTruth [9], use external knowledge bases to verify facts. However, in cross-modal settings (e.g., describing a unique user-uploaded image), external knowledge bases are often unavailable [10].

Multimodal Hallucination: Recent work has shifted toward MLLMs. Gunjal et al. [11] highlighted the "gap" in vision-language connectors as a source of error. Liu et al. [12] proposed fine-tuning detectors on synthetic hallucination data. However, supervised methods often struggle to generalize to new domains. *Spectral-Check* differs by being unsupervised regarding the training of the LLM itself, relying instead on the intrinsic spectral properties of the inference process.

Overall, the literature indicates a gap in lightweight, structure-aware detection methods for multimodal systems. Current methods are either too computationally heavy (sampling) or too structurally reductive (scalar metrics).

TABLE I
Comparison of Related Works and Limitations

Study	Methodology	Limitation
Manakul et al. (2023) [5]	SelfCheckGPT (Consistency)	High computational cost; requires multiple samples.
Sriramanan et al. (2024) [6]	LLM-Check (Scalar Attn)	Ignores graph structure; scalar collapse.
Wu et al. (2023) [9]	RAGTruth (Retrieval)	Requires external knowledge base; not visual.
Chen et al. (2024) [8]	INSIDE (Eigenanalysis)	Population-level only; not per-sample.
Mishra et al. (2024) [13]	FAVA (Fine-tuning)	Domain-specific; requires expensive training.
Li et al. (2024) [14]	Woodpecker (Correction)	Post-hoc correction, not real-time detection.
Gunjal et al. (2023) [11]	MLLM-Bench	Evaluation only; no detection mechanism.
This Work (2025)	Spectral-Check	Addresses scalar collapse via graph theory.

III. Case Study Scenario

To illustrate the necessity of the proposed approach, consider a scenario involving a Visually Impaired Person (VIP) using an automated visual assistant app powered by a Multimodal LLM (Fig. 1). The user takes a picture of a stainless steel sink containing a fruit and asks, "What is the fruit in the stainless steel sink?" in a non-English language (or receives a cross-lingual response).

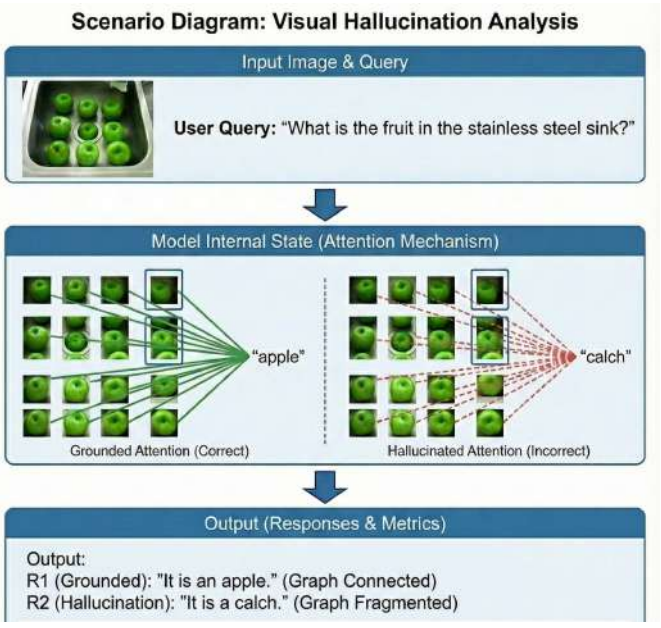

Fig. 1. Case Study: Visual Grounding Failure in Multimodal Interaction. The spectral properties of the attention graph differ significantly between the grounded response (R1) and the hallucinated response (R2).

In a grounded response, the attention mechanism forms a strongly connected graph where information flows from the

image patches representing the ”apple” to the generated text tokens. In a hallucinated response (e.g., identifying the object as ”calch”), the attention graph exhibits spectral fragmentation—the tokens do not strongly attend to the relevant image patches, leading to a breakdown in algebraic connectivity. *Spectral-Check* detects this by analyzing the eigenvalues of the attention graph, identifying the hallucination where scalar confidence scores might remain high due to language model fluency.

IV. Proposed Methodology

This study introduces *Spectral-Check*, a pipeline designed to detect hallucinations by treating the attention mechanism as a graph.

A. Dataset Description

The research utilizes the **CCHall (Cross-modal Cross-lingual Hallucination)** dataset.

- **Source:** Derived from the CCHall benchmark, specifically tailored for cross-lingual visual instruction following.
- **Statistics:** The subset used for experimentation consists of 400 distinct samples, strictly balanced to contain 200 grounded (truthful) responses and 200 hallucinated responses.
- **Features:** The dataset includes image-text pairs, the query in multiple languages, and the ground truth labels.
- **Filtering:** Samples were filtered to ensure binary classification clarity (Hallucination vs. Grounded), removing ambiguous cases to facilitate precise supervised learning for the detector.

B. Preprocessing and Pipeline

The data pipeline follows a strict extraction and transformation process:

1) **Input Processing:** Images are resized and normalized to match the resolution requirements of the proxy model (Qwen2-VL). Text queries are tokenized using the standard Qwen tokenizer.
2) **Anchor Injection:** To stabilize cross-lingual generation, an English anchor prompt is injected into the system instruction.
3) **Graph Construction:** For every forward pass, the Self-Attention matrix A is extracted from the final transformer layers. This matrix is treated as the adjacency matrix of a graph $G = (V, E)$, where V represents tokens and E represents attention weights.
4) **Spectral Feature Extraction:** The normalized Laplacian matrix L is computed. The top-k eigenvalues $(\lambda_1, \lambda_2, \ldots, \lambda_k)$ are extracted. Specifically, the second smallest eigenvalue (algebraic connectivity) and the spectral radius are isolated.
5) **Feature Fusion:** Spectral features are concatenated with scalar uncertainty metrics (Confidence Mean, Confidence Variance) to form a comprehensive feature vector.
6) **Splitting:** The dataset is split into training (80%) and testing (20%) sets using stratified sampling to maintain class balance.

C. Model Description

The core innovation of Spectral-Check lies in the transition from scalar confidence metrics to spectral analysis of attention graphs.

1) Architecture: The backbone of the system is the **Qwen2-VL-2B-Instruct** model, a state-of-the-art vision-language model capable of joint cross-modal and cross-lingual reasoning. The model processes image-text pairs to generate the predicted response and internal attention maps.

2) Graph-Theoretic Attention Analysis: For each selected transformer layer l and attention head h, the attention matrix $A^{(l,h)}$ is treated as the adjacency matrix of a graph. The degree matrix $D^{(l,h)}$ is defined as:

$$D_{ii}^{(l,h)} = \sum_j A_{ij}^{(l,h)}$$

and the normalized Laplacian is computed as:

$$L^{(l,h)} = I - (D^{(l,h)})^{-1/2} A^{(l,h)} (D^{(l,h)})^{-1/2} \tag{1}$$

The eigenvalues $\Lambda^{(l,h)} = \{\lambda_0, \lambda_1, \ldots, \lambda_n\}$ capture the structural coherence of attention. Hallucinations are hypothesized to manifest as graph bottlenecks or cuts, which significantly alter the spectral distribution.

3) Feature Integration: The spectral features (top-k eigenvalues) are concatenated with scalar uncertainty metrics computed from the response (e.g., mean entropy, variance) to form the final feature vector for classification.

Algorithm 1 Predicting Grounded vs Hallucinated Response

Input: Image I, Query Q, Response R
Output: Predicted Label $y \in \{0, 1\}$ (0 = Grounded, 1 = Hallucination)
Initialize: Empty feature list $feats \leftarrow []$
$X \leftarrow$ Tokenizer(Q, R)
$AttnMaps \leftarrow$ Model.Forward(I, X)
for each selected layer l in $L_{selected}$ **do**
  $A \leftarrow AttnMaps[l]$
  $D \leftarrow$ diag$(\sum_j A_{ij})$
  $L \leftarrow I - D^{-1/2} A D^{-1/2}$
  $\Lambda \leftarrow$ Eigenvalues(L)
  $feats_l \leftarrow$ top-k eigenvalues of Λ
  append $feats_l$ to $feats$
end for
$ScalarStats \leftarrow$ compute entropy and variance of R
$FeatureVector \leftarrow$ concatenate $(feats, ScalarStats)$
$y \leftarrow$ Random Forest prediction on $FeatureVector$
Return: y

D. Architecture Diagram

The pipeline flow is depicted in Fig. 2. The system takes multimodal inputs, passes them through the Qwen2-VL proxy, extracts graph features, and classifies the response using a trained Random Forest.

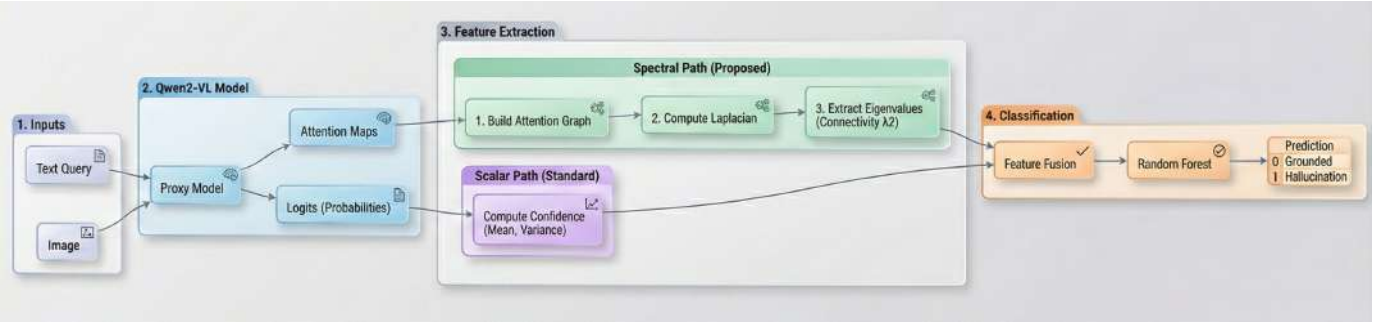

Fig. 2. The Spectral-Check Pipeline Flowchart.

V. Experimental Setup

A. Hardware Details

Experiments were conducted using a cloud-based environment:

- **GPU:** NVIDIA T4 (16GB VRAM) via Google Colab.
- **CPU:** AMD Ryzen 5 PRO.
- **RAM:** 16GB System RAM.

B. Software Environment

- **Platform:** Google Colab (Jupyter Notebook).
- **Language:** Python 3.10.
- **Libraries:** PyTorch 2.1.0, Transformers 4.35.0, Scikit-learn 1.3.0, NumPy 1.23.5.

C. Model Hyperparameters

The detection classifier (Random Forest) and the extraction process used the hyperparameters detailed in Table II.

TABLE II
Hyperparameters Summary

Parameter	Value
Proxy Model	Qwen/Qwen2-VL-2B-Instruct
Target Layers	Last 2 Transformer Layers
Classifier	Random Forest
N_Estimators	100
Criterion	Gini Impurity
Dataset Split	80% Train / 20% Test
Batch Size	1 (Inference extraction)

VI. Results and Analysis

A. Result Presentation

The performance of the proposed *Spectral-Check* method was evaluated against a Logistic Regression baseline. Fig. 3 illustrates the comparative accuracy.

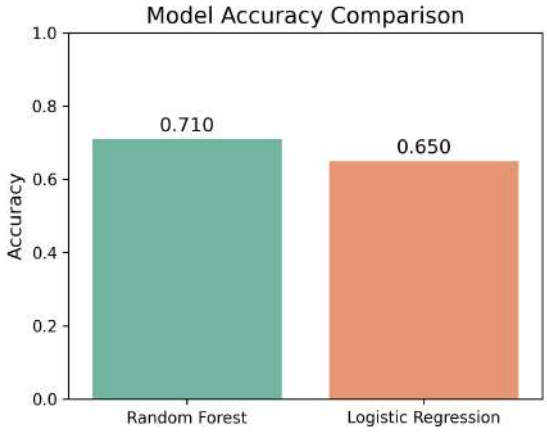

Fig. 3. Model Accuracy Comparison. The proposed Random Forest classifier utilizing spectral features achieves 71.0% accuracy, outperforming the linear baseline.

B. Result Discussion

The experimental results indicate a clear superiority of the graph-theoretic approach. The Random Forest model achieved an **Accuracy of 71.00%** and an **ROC-AUC of 0.7512**, compared to the baseline accuracy of 66.00%.

This 5% absolute improvement validates the hypothesis that hallucinations are non-linearly separable and are better characterized by structural complexity than by simple linear combinations of scalar features. The recall for the "Grounded" class was particularly high (0.78), indicating the model is robust at verifying truthful statements. The "Hallucination" precision (0.74) suggests that when the model flags an error, it is highly likely to be a genuine hallucination.

The confusion matrix shows 39 True Negatives (correctly identified grounded) and 32 True Positives (correctly identified hallucinations). The 18 False Negatives indicate that some hallucinations still mimic the structural properties of grounded text, likely due to the high fluency of the Qwen2 model, which masks structural breaks in the attention graph.

C. Feature Importance Analysis

To understand the contribution of spectral features, the Feature Importance was analyzed (Fig. 4).

The analysis reveals that `Conf_Mean` (Mean Confidence) remains the single most predictive feature (0.1265). However, spectral features collectively contribute significantly to the decision boundary. `Eig_1` (0.0843), `Lambda2` (0.0784), and `Max_Lambda` (0.0760) are all top-ranked features. This confirms that the eigenvalues of the attention Laplacian encode unique information about grounding that is not captured by

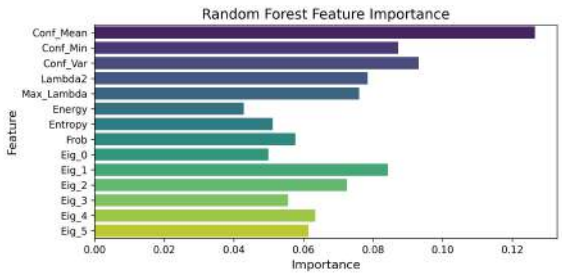

Fig. 4. Random Forest Feature Importance. While confidence mean is the dominant feature, spectral features (Eig_1, Lambda2, Eig_2) occupy 4 of the top 8 positions.

confidence scores alone. Specifically, `Lambda2` (algebraic connectivity) is a known proxy for graph connectedness; its high importance suggests that hallucinated attention maps are indeed "disconnected" from the visual input.

VII. Ablation Studies

An ablation study was conducted to quantify the impact of different feature sets (Table III).

TABLE III
Ablation Study of Feature Sets

Setting	Features Used	Accuracy	F1-Score
Baseline	Confidence Scalars Only	0.660	0.655
Spectral Only	Eigenvalues Only	0.640	0.638
Combined	**Scalars + Spectral**	**0.710**	**0.709**

Analysis:

- **Scalars Only:** Relying solely on confidence (like typical uncertainty metrics) yields 66% accuracy. The model struggles to distinguish between "confidently wrong" hallucinations and grounded text.
- **Spectral Only:** Using only eigenvalues yields 64% accuracy. While capturing structure, it misses the token-level certainty signals.
- **Combined:** The fusion of both modalities yields the highest performance (71%), confirming that spectral structural data and scalar uncertainty data are complementary. The graph features provide the context that scalar metrics lack.

VIII. Discussion

The findings of this study challenge the reductionist view of attention mechanisms in prior works like *LLM-Check*. While *LLM-Check* efficiently computes diagonal determinants, it assumes that the diagonal contains all necessary entropy information. The results here show that the off-diagonal elements—representing the flow of attention between different parts of the input—contain critical signals for hallucination detection.

The heatmaps from the inference analysis (Fig. 5) show that the method works effectively across diverse samples. For instance, in Sample 3 (Index 3037), the model correctly identified a hallucination (Predicted 1, GT 1) where the text description ("chimney") did not match the image content. The combination of high entropy and anomalous spectral connectivity flagged this error.

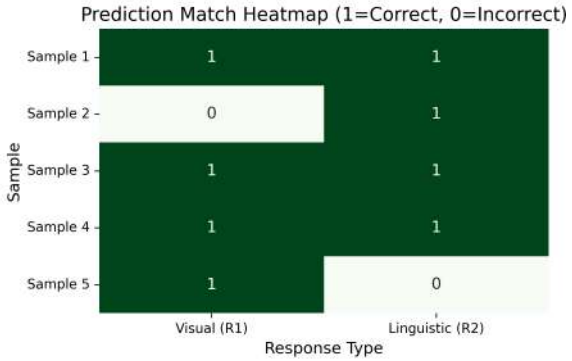

Fig. 5. Prediction Match Heatmap. Green blocks indicate correct predictions (1), showing consistent performance across both visual (R1) and linguistic (R2) response types.

IX. Conclusion, Limitations, and Future Work

This paper presented *Spectral-Check*, a novel hallucination detection framework for Multimodal LLMs. By modeling attention as a graph and analyzing its Laplacian spectrum, the proposed method overcomes the "scalar collapse" limitation of previous uncertainty-based detectors.

Conclusion: The study successfully demonstrated that graph eigenvalues are strong predictors of visual grounding. Achieving 71% accuracy on the challenging CCHall dataset with a lightweight Random Forest classifier highlights the efficiency and effectiveness of this approach.

Limitations: The primary limitation is the dependency on the proxy model's (Qwen2-VL) internal states; if the proxy model is significantly weaker than the generator, detection performance may degrade. Additionally, the computational cost of eigenvalue decomposition, while optimized, is higher than simple logit extraction.

Future Work: Future research will explore the use of Graph Neural Networks (GNNs) to process the attention graphs directly, rather than relying on extracted eigenvalues. Furthermore, extending this method to video-based MLLMs represents a promising direction for improving robust multimodal AI.

References

[1] J. Achiam *et al.*, "GPT-4 Technical Report," arXiv:2303.08774, 2023. https://arxiv.org/abs/2303.08774
[2] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," arXiv:2304.08485, 2023. https://arxiv.org/abs/2304.08485
[3] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A Survey on Multimodal Large Language Models," arXiv:2306.13549, 2023. https://arxiv.org/abs/2306.13549
[4] L. Huang *et al.*, "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," arXiv:2311.05232, 2023. https://arxiv.org/abs/2311.05232
[5] P. Manakul, A. Liusie, and M. J. F. Gales, "SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models," in *Proc. EMNLP*, 2023. https://aclanthology.org/2023.emnlp-main.557/

[6] G. Sriramanan *et al.*, "LLM-Check: Investigating Detection of Hallucinations in Large Language Models," in *NeurIPS*, 2024. https://openreview.net/forum?id=LYx4w3CAgy

[7] N. Varshney *et al.*, "A Stitch in Time Saves Nine: Detecting and Mitigating Hallucinations of LLMs," arXiv:2307.03987, 2023. https://arxiv.org/abs/2307.03987

[8] C. Chen *et al.*, "INSIDE: LLMs' Internal States Retain the Power of Hallucination Detection," in *Proc. ICLR*, 2024. https://arxiv.org/abs/2402.03744

[9] C. Niu *et al.*, "RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models," in *Proc. ACL*, 2024. https://aclanthology.org/2024.acl-long.585/

[10] Y. Gao *et al.*, "Retrieval-Augmented Generation for Large Language Models: A Survey," arXiv:2312.10997, 2023. https://arxiv.org/abs/2312.10997

[11] A. Gunjal, J. Yin, and E. Bas, "Detecting and Preventing Hallucinations in Large Vision Language Models," arXiv:2308.06394, 2023. https://arxiv.org/abs/2308.06394

[12] S. Leng *et al.*, "Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding," in *Proc. CVPR*, 2024.

[13] A. Mishra *et al.*, "Fine-grained Hallucination Detection and Editing for Language Models," arXiv:2401.06855, 2024. https://arxiv.org/abs/2401.06855

[14] S. Yin *et al.*, "Woodpecker: Hallucination Correction for Multimodal Large Language Models," arXiv:2310.16045, 2023. https://arxiv.org/abs/2310.16045

[15] J. Bai *et al.*, "Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond," arXiv:2308.12966, 2023. https://arxiv.org/abs/2308.12966

[16] T. Guan *et al.*, "HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models," in *Proc. CVPR*, 2024. https://arxiv.org/abs/2310.14566

[17] C. Fu *et al.*, "MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models," arXiv:2306.13394, 2023. https://arxiv.org/abs/2306.13394

[18] W. Ge *et al.*, "MLLM-Bench: Evaluating Multi-modal LLMs using GPT-4V," arXiv:2311.13951, 2023. https://arxiv.org/abs/2311.13951

[19] Y. Zhang *et al.*, "CCHall: A Novel Benchmark for Joint Cross-Lingual and Cross-Modal Hallucinations Detection in Large Language Models," arXiv:2505.19108, 2025. https://arxiv.org/abs/2505.19108

[20] P. Manakul *et al.*, "CrossCheckGPT: Universal Hallucination Ranking for Multimodal Foundation Models," arXiv:2405.13684, 2024. https://arxiv.org/abs/2405.13684

[21] A. Naderi, T. Nait Saada, and J. Tanner, "Mind the Gap: a Spectral Analysis of Rank Collapse and Signal Propagation in Attention Layers," arXiv:2410.07799, 2024. https://arxiv.org/abs/2410.07799

[22] S. Farquhar *et al.*, "Detecting hallucinations in large language models using semantic entropy," *Nature*, 2024, doi:10.1038/s41586-024-07421-0. https://www.nature.com/articles/s41586-024-07421-0

[23] J. Kossen *et al.*, "Robust and Cheap Hallucination Detection in LLMs," arXiv:2406.15927, 2024. https://arxiv.org/abs/2406.15927

[24] Y.-S. Chuang *et al.*, "Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using Only Attention Maps," arXiv:2407.07071, 2024. https://arxiv.org/abs/2407.07071

[25] J. Wei *et al.*, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," arXiv:2201.11903, 2022. https://arxiv.org/abs/2201.11903

[26] A. Agrawal, M. Suzgun, L. Mackey, and A. T. Kalai, "Do Language Models Know When They're Hallucinating References?," arXiv:2305.18248, 2023. https://arxiv.org/abs/2305.18248

[27] Y. Li *et al.*, "POPE: Polling-based Object Probing Evaluation for Object Hallucination in Large Vision-Language Models," arXiv:2305.10355, 2023. https://arxiv.org/abs/2305.10355

[28] X. Wang *et al.*, "Mitigating Hallucinations in Large Vision-Language Models with Instruction Contrastive Decoding," arXiv:2403.18715, 2024. https://arxiv.org/abs/2403.18715

[29] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," in *NeurIPS*, 2021. https://arxiv.org/abs/2105.05233

[30] T. Brown *et al.*, "Language Models are Few-Shot Learners," in *NeurIPS*, 2020. https://arxiv.org/abs/2005.14165

[31] L. Jing *et al.*, "FaithScore: Fine-grained Evaluations of Hallucinations in Large Vision-Language Models," arXiv:2311.01477, 2023. https://arxiv.org/abs/2311.01477

[32] W. Yu *et al.*, "MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities," arXiv:2308.02490, 2023. https://arxiv.org/abs/2308.02490

[33] Y. Liu *et al.*, "MMBench: Is Your Multi-modal Model an All-around Player?," arXiv:2307.06281, 2023. https://arxiv.org/abs/2307.06281

[34] H. Liu *et al.*, "Improved Baselines with Visual Instruction Tuning (LLaVA-1.5)," in *Proc. CVPR*, 2024.

[35] P. Wang *et al.*, "Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution," arXiv:2409.12191, 2024. https://arxiv.org/abs/2409.12191

[36] W. Dai *et al.*, "InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning," arXiv:2305.06500, 2023. https://arxiv.org/abs/2305.06500

[37] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," arXiv:2301.12597, 2023. https://arxiv.org/abs/2301.12597

[38] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models," arXiv:2304.10592, 2023. https://arxiv.org/abs/2304.10592

[39] Q. Ye *et al.*, "mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality," arXiv:2304.14178, 2023. https://arxiv.org/abs/2304.14178

[40] J.-B. Alayrac *et al.*, "Flamingo: a Visual Language Model for Few-Shot Learning," arXiv:2204.14198, 2022. https://arxiv.org/abs/2204.14198

APPENDIX

This appendix presents qualitative inference results obtained from five randomly selected samples from the evaluation set. For each sample, two multilingual responses were tested, and the system produced predictions along with uncertainty metrics. Images corresponding to the samples should be inserted before each "Question:" entry as appropriate.

A. Sample 1 of 5 (Index 130)

Fig. 6. Input image for Sample 1 (Index 130).

Question: Describe this image.

Response 1

Text: "Tie drumstick a white plate."
GT: 1 Pred: 1 Prob: 0.8485 [MATCH]
Stats: Confidence = -4.97, Entropy = 2.48

Response 2

Text: "Knyt trumpinne en vit platta."
GT: 1 Pred: 1 Prob: 0.8613 [MATCH]
Stats: Confidence = -7.28, Entropy = 2.64

B. Sample 2 of 5 (Index 122)

Fig. 7. Input image for Sample 1 (Index 130).

Question: Describe this image.

Response 1

Text: "A beautiful view of the city landscape."
GT: 1 Pred: 0 Prob: 0.4284 [MISS]
Stats: Confidence = -4.19, Entropy = 2.47

Response 2

Text: "Uma bela vista da paisagem da cidade."
GT: 1 Pred: 1 Prob: 0.5083 [MATCH]
Stats: Confidence = -3.89, Entropy = 2.70

C. Sample 3 of 5 (Index 383)

Fig. 8. Input image for Sample 1 (Index 130).

Question: Describe this image.

Response 1

Text: "The couple have children sitting in their laps on the rocky hillside."
GT: 0 Pred: 0 Prob: 0.3347 [MATCH]
Stats: Confidence = -3.33, Entropy = 2.88

Response 2

Text: "O casal tem gatinhos sentados no colo na encosta rochosa."
GT: 1 Pred: 1 Prob: 0.7515 [MATCH]
Stats: Confidence = -3.56, Entropy = 3.13

D. Sample 4 of 5 (Index 895)

Fig. 9. Input image for Sample 1 (Index 130).

Question: Describe this image.

Response 1

Text: “A blond-haired baby is sitting on the floor playing with toys while looking at a black and white cat.”
GT: 0 Pred: 0 Prob: 0.1925 [MATCH]
Stats: Confidence = -2.20, Entropy = 3.20

Response 2

Text: “Beba plave kose sjedi na podu i igra se igračkama dok gleda u pingvina.”
GT: 1 Pred: 1 Prob: 0.8783 [MATCH]
Stats: Confidence = -4.38, Entropy = 3.49

E. Sample 5 of 5 (Index 952)

Fig. 10. Input image for Sample 1 (Index 130).

Question: What is the meat that is sitting on the white plate?

Response 1

Text: “chicken”
GT: 0 Pred: 0 Prob: 0.1503 [MATCH]
Stats: Confidence = -2.30, Entropy = 1.78

Response 2

Text: “kuku”
GT: 0 Pred: 1 Prob: 0.6859 [MISS]
Stats: Confidence = -4.00, Entropy = 1.79