GROUNDING THE UNGROUNDED: A SPECTRAL-GRAPH FRAMEWORK FOR QUANTIFYING HALLUCINATIONS IN MULTIMODAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Hallucinations in LLMs—especially in multimodal settings—undermine reliability. We present a rigorous, information-geometric framework in diffusion dynamics that quantifies hallucination in MLLMs: model outputs are embedded spectrally on multimodal graph Laplacians, and gaps to a truth manifold define a semantic-distortion metric. We derive Courant–Fischer bounds on a temperature-dependent hallucination energy and use RKHS eigenmodes to obtain modality-aware, interpretable measures that track evolution over prompts and time. This reframes hallucination as measurable and bounded, providing a principled basis for evaluation and mitigation.

1 INTRODUCTION

Large language models (LLMs) and their multimodal variants (MLLMs) are powerful generators, but reliability or truthfulness remains a core limitation. A central drawback is the hallucinated content that is ungrounded or inconsistent with inputs - which is unacceptable and signifactly risky in medicine, law, and finance Ji et al. (2023); Maynez et al. (2020); Bubeck et al. (2023). Prior work offers taxonomies, datasets, and benchmarks for analysis and evaluation Ji et al. (2023); Maynez et al. (2020); Ding et al. (2024), and recent multimodal studies emphasize empirical detection/mitigation Bai et al. (2024); however, most approaches rely on heuristics, proxy metrics, or human annotation rather than principled quantification.

On the theory side, complementary work include token-level analysis of hallucinated predictions Jiang et al. (2024), Bayesian sequential detection Wang et al. (2023), entropy-style uncertainty probes Han et al. (2024), latent-space steering to separate truthful vs. hallucinated generations Park et al. (2025), and reference-free ranking for multimodal hallucinations Sun et al. (2024). Emerging spectral/graph perspectives probe representations and attention, but are largely detection-oriented and unimodal Xie et al. (2025).

Gap. The field currently lacks a quantitative, theory-backed, modality-aware framework that treats hallucination as a measurable quantity (with temporal dynamics and guarantees), rather than only a classification/detection outcome.

Our contribution. We introduce a spectral-graph framework that makes hallucination in MLLMs measurable and bounded in the context of time-indexed temperature profiles:

(a) We model the grounding across modalities via optimal-transport paths in diffusion dynamics and embed them in RKHS, yielding a structural view of semantic consistency.
(b) We represent outputs on multimodal graph Laplacians and derive tight Courant–Fischer (CF) bounds on hallucination heatmap as a function of time-indexed temperature.
(c) *Empirical validation*: Across nine 3D panels (COCO/VQAv2/AudioCaps $\times$ `CLIP+Whisper+T5`, `BLIP+CLIP+Whisper`, `SigLIP+Whisper+T5`), $\mathcal{E}_{\text{hall}}^{\text{multi}}$ lies between panel-specific CF planes with a strictly positive lower envelope that tightens at lower temperature (and higher diffusion); full $\varepsilon/\tau/h/\rho$ ablations and runtimes in the supplement.

This shifts hallucination study from qualitative detection to quantitative, modality-aware, and interpretable analysis. To our knowledge, it is the first attempt to provide spectral bounds on hallucina-

tion for MLLMs followed by a time-indexed temperature annealing, offering a principled basis for evaluation and potential mitigation.

2 Related Work

Kalai & Vempala show that, for calibrated LMs, the hallucination rate is lower-bounded by a Good–Turing–style "monofact" mass - establishing an inherent trade-off between calibration and truthfulness Kalai & Vempala (2024); while their recent work generalizes this via an IIV reduction that ties generative errors to binary-classification - advocating IDK-tolerant evaluation Kalai et al. (2025). Empirical study of LM hallucinations spans mechanistic probes that surface interpretable features for diagnosis Templeton et al. (2024), retrieval-grounded detection and evaluation Gerner et al. (2025); Niu et al. (2024), broad benchmark suites like HaluEval Li et al. (2023), Hallu-PI Ding et al. (2024), GraphEval Feng et al. (2025), and early vision–language analyses of object hallucination Rohrbach et al. (2018). Comprehensive surveys catalog causes, detection, and mitigation strategies Ji et al. (2023); Rawte et al. (2023).

Recent work exploits uncertainty and structural signals: semantic-entropy probes Han et al. (2024), Bayesian sequential estimation Wang et al. (2023), token-level dynamics of hallucinated predictions Jiang et al. (2024), zero-shot reasoning signals Lee et al. (2024), and sampling-based self-consistency checks (SelfCheckGPT) Manakul et al. (2023). Graph/spectral methods flag hallucinations via KG self-checks (FactSelfCheck) Sawczyn et al. (2025), attention Laplacian eigen-spectra (LapEigvals) Binkowski et al. (2025), and topological cues on hallucination graphs Le Merrer & Trédan (2024).

3 Preliminaries

We begin by establishing the mathematical foundations of our framework. MLLM outputs are embedded as nodes on a knowledge graph Laplacian, and grounding gaps along this graph collectively define a quantifiable hallucination metric. Figure 1 sketches our approach.

3.1 Mathematical Foundations

Let $\mathcal{X}$ denote the measurable[A.1 1] set of all possible model outputs of a multimodal LLM, with $\mathcal{F}_{\mathcal{X}}$ being the σ-algebra over $\mathcal{X}$ and μ being the base measure Tao (2011); e.g., the count measure for discrete outputs like token sequence or the Lebesgue measure for continuous outputs like embeddings Bartle (1995). We assume $\mathcal{X}$ is continuously embedded in a separable Reproducing Kernel Hilbert space (RKHS) denoted by $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ which is associated with a positive-definite kernel,

$$K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{+}. \tag{1}$$

The kernel $K(x_1, x_2)$ encodes the semantic relationships between two distinct points or outputs x_1 and x_2 $\forall (x_1 \neq x_2) \in \mathcal{X}$; for example, through embedding-based or ontology-aware distance measures, or co-reference resolution. For a product kernel in an MLLM, refer to Eq. (7) later.

Within this $(\mathcal{X}, \mathcal{F}_{\mathcal{X}}, \mu)$ space, there exist two kinds of "truth" (the idea imported from Kalai & Vempala (2024)):

(i) The semantic factoid space $\mathcal{K}$ which encompasses all semantically valid and coherent outputs that include empirically plausible facts, contextually appropriate completions, and domain-consistent inferences aligned with the prompt and background knowledge - importantly, elements of $\mathcal{K}$ need not be verifiable, but they remain semantically valid within the modeled domain.

(ii) The semantic ground-truth manifold $\mathcal{K}_g$, as a stricter subregion of $\mathcal{K}$, which consists of outputs only verifiably correct or true facts that include factual assertions supported by empirical evidence or directly observed information — elements of $\mathcal{K}_g$ can be properly referred to as grounded in reality.

[1] Footnotes are added in chronological order and collected in Appendix A.

Thus the semantic plausibility/ground-truth nesting and, for a given prompt $p \in \mathcal{P}$, the hallucination criterion for each output denoted by $x \in \mathcal{X}$ are:

$$\mathcal{K}_{\mathrm{g}} \subseteq \mathcal{K} \subset \mathcal{X}, \quad x \in \mathcal{X} \setminus \mathcal{K}. \tag{2}$$

Note: $x \in \mathcal{K} \setminus \mathcal{K}_{\mathrm{g}}$ is a non-grounded output, but still semantically plausible and strictly not hallucination.

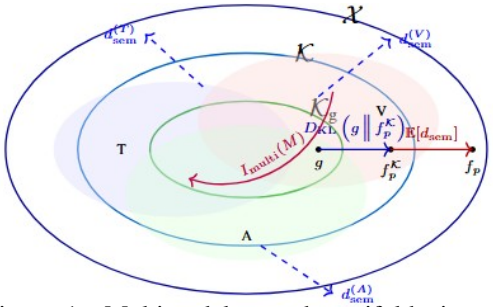


Figure 1: Multimodal nested-manifold view of hallucinations. Hollow ellipses denote $\mathcal{X}$, $\mathcal{K}$, $\mathcal{K}_g$.

3.2 Modeling the LLM outputs

We begin with the baseline assumptions:

Assumption 1 (General output distribution)**.** *The LLM outputs can be characterized by a conditional probability distribution $f_p(x)$ that denotes the likelihood of generating output x given a prompt p:*

$$f_p : \mathcal{X} \to [0, \infty), \quad f_p \in L^1(\mathcal{X}, \mathcal{F}_{\mathcal{X}}, \mu) \cap \mathcal{H}, \quad x \mapsto f_p(x), \tag{3}$$

which ensure $\int_{\mathcal{X}} f_p(x)\, d\mu(x) = 1$.

Let $f_p^{\mathcal{K}}$ denote the restricted distribution on the semantic plausibility space $\mathcal{K}$:

$$f_p^{\mathcal{K}}(x) := \frac{\mathbf{1}_{\{x \in \mathcal{K}\}} f_p(x)}{\int_{\mathcal{K}} f_p(x') d\mu(x')} \equiv \frac{\mathbf{1}_{\{x \in \mathcal{K}\}} f_p(x)}{\mathbb{P}_{f_p}(\mathcal{K})}, \quad \text{where, } \mathbf{1}_{\{x \in \mathcal{K}\}} = \begin{cases} 1 & \text{if } x \in \mathcal{K}, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Here, $\int_{\mathcal{K}} f_p(x') d\mu(x') = \mathbb{P}_{f_p}(\mathcal{K})$ is a normalization constant in the restricted distribution.

Assumption 2 (Ground-truth generative distribution)**.** *In line with Assumption 1, g denotes the reference distribution on the ground-truth manifold $\mathcal{K}_g$. Unlike f_p or $f_p^{\mathcal{K}}$, g is the gold reference which is not model-induced and hence, may not share support with f_p except inside $\mathcal{K}_g$ and it is truly independent of prompts in the generative sense (but conditioned on the same prompt contextually).*

Thus, we do not assume any parametric form for the ground-truth distribution g and rather treat it as an abstract measure over $\mathcal{K}_{\mathrm{g}}$:

$$\mathrm{supp}(g) \subseteq \mathcal{K}_{\mathrm{g}}, \quad g : \mathcal{K}_{\mathrm{g}} \to [0, \infty), \quad g \in L^1(\mathcal{K}_{\mathrm{g}}, \mathcal{F}_{\mathcal{X}}|_{\mathcal{K}_{\mathrm{g}}}, \mu'). \tag{5}$$

Eq. (5) ensures $\int_{\mathcal{K}_{\mathrm{g}}} g(x)\, d\mu'(x) = 1$ with notations used in consistency with Eq. (3) and μ' playing the same role of μ, but not necessarily equal to μ. See comments[A.2] in Appendix A.

4 Theoretical Analysis

In this section, we present a theoretical framework that couples a smoothed information-geometric score derived from the Kullback–Leibler (KL) paradigm[A.3] with a multimodal energy formulation to quantify and track hallucinations in MLLMs.

4.1 Semantic Distortion

We establish the following theorem followed by stating a remark to set the stepping stone.

Theorem 1 (KL-calibrated smoothed score for hallucination)**.** *Let a smoothing mass $\varepsilon \in (0, 1)$ and a baseline density be fixed, with finite $\rho(x) > 0$ μ-a.e. and $\int_{\mathcal{X}} \rho(x)\, d\mu(x) = 1$; let $K_h(\cdot, \cdot) \in (0, \infty)$ be a μ-Markov kernel (bandwidth $h > 0$) and $T_h : L^1(\mu) \to L^1(\mu)$ be a linear smoother defined for $q : \mathcal{X} \to \mathbb{R}$ by $(T_h q)(x_1) := \int_{\mathcal{X}} K_h(x_1, x_2)\, q(x_2)\, d\mu(x_2)$; let the ε-smoothed model be $\tilde{f}_{p,\varepsilon}(x) := (1 - \varepsilon) f_p(x) + \varepsilon \rho(x)$ with its $\mathcal{K}$-restricted renormalization $\tilde{f}_{p,\varepsilon}^{\mathcal{K}}(x_2) := \mathbf{1}_{\{x_2 \in \mathcal{K}\}} \tilde{f}_{p,\varepsilon}(x_2) / \int_{\mathcal{K}} \tilde{f}_{p,\varepsilon}(x)\, d\mu(x)$; and let a measurable selector $\Pi_{\mathcal{K}} : \mathcal{X} \to \mathcal{K}$ satisfy $\Pi_{\mathcal{K}}(x) = x$ $(\forall x \in \mathcal{K})$ or nearest point with convexity in $\mathcal{K}$ (otherwise). Then the semantic distortion*

$$d_{\mathrm{sem}}^{(\varepsilon,h)}(x; \mathcal{K}, \mathcal{X}) := \Big[\log\big((T_h \tilde{f}_{p,\varepsilon}^{\mathcal{K}})(\Pi_{\mathcal{K}}(x))\big) - \log\big((T_h \tilde{f}_{p,\varepsilon})(x)\big) \Big]_+, \tag{6}$$

serves as a KL-calibrated smoothed pointwise information gap for tracking hallucinations across prompts and remains as a reference-free (independent-of-g) statistic in language models.

Proof sketch: Strict positivity from $\tilde{f}_{p,\varepsilon} = (1-\varepsilon)f_p + \varepsilon\rho$ and Markov K_h makes both smoothed terms > 0, so Eq. (6) is finite. If $x \in \mathcal{K}$, $\Pi_{\mathcal{K}}(x) = x$ and the $\mathcal{K}$–restricted smoother $>$ the unconditional smoother at x; if $x \notin \mathcal{K}$, smoothing at $\Pi_{\mathcal{K}}(x) \in \mathcal{K}$ dominates the mixed mass at x. Detailed proof is found in Appendix B.1. □

Remark 1. *The score in Eq. (6) is g-agnostic and thus usable when g is unobservable*[A.4] *or partially verified in various real-world scenarios. In practice, we set a small smoothing mass* $\varepsilon \in [10^{-6}, 10^{-2}]$*, choose* h *by validation, take* K_h *as a positive row-normalized kernel over embeddings/tokens, and we implement* $\Pi_{\mathcal{K}}$ *as a measurable nearest–neighbour selector on a finite reference set from* $\mathcal{K}$*.*

4.2 Extension to Multi-modal Grounding

The intuition behind this setting of multimodality is: in image-grounded or dialogue models, semantic grounding depends on multiple modalities — e.g., text, image or video, dialog or audio-history etc. and the RKHS is then extended to a multi-modal product kernel space. In multi-modal settings, where the LLM outputs involve textual (T), visual (V), audio (A) modalities, we define a joint output space ($\mathcal{X}$) embedded into a composite RKHS ($\mathcal{H}$) equipped with a product kernel (K) between two distinct points (i.e., outputs) $\forall (x_1 \neq x_2) \in \mathcal{X}$ as

$$\mathcal{X} : \underset{M}{\times} \mathcal{X}_M, \quad x = (x^{(M)})_{x^{(M)} \in \mathcal{X}_M}, \quad \mathcal{H} := \underset{M}{\otimes} \mathcal{H}_M, \quad K(x_1, x_2) = \prod_{M} K_M(x_1^{(M)}, x_2^{(M)}), \quad (7)$$

pertaining to each modality $\forall M \in \mathcal{M} := \{T, V, A\}$, where the prompts can also be categorized into a composite prompt space $\mathcal{P} : \underset{M}{\times} \mathcal{P}_M$, with each prompt $p = (p^{(M)})_{p^{(M)} \in \mathcal{P}_M}$ in a modality-aware prescription to accommodate three different kinds of probable inputs (i.e., T, V & A) for the sake of completeness. However, in the following calculation in this paper, we restrict ourselves only to the notion of p without any loss of generality. Expanded form[A.5] of Eq. (7) is found in Appendix A.

4.3 Formulations to Hallucination Energy

To begin with, we are after a fruitful formulation of $f_p(x)$ that connects the model output distribution to an underlying energy landscape to enable modal interpretability, temperature-driven exploration, and spectral graph analysis. The total energy functional $\mathcal{E}(x, p, \cdot) : \mathcal{X} \times \mathcal{P} \to \mathbb{R}^+$ associated with the model input-output plus suppressed parameters can be decomposed into intra-modal, pairwise cross-modal, and joint multimodal interactions. This decomposition allows us to localize the sources of hallucination within and across modalities.

Assumption 3 (Hallucination energy functional in MLLMs)**.** *The modality-aware decomposition reads as:*

$$\mathcal{E}(x, p, \cdot) = \sum_{M \in \mathcal{M}} \mathcal{E}_M \left(x^{(M)}, p, \cdot \right) + \sum_{\substack{M, M' \in \mathcal{M} \\ M \neq M'}} \mathcal{E}_{MM'} \left(x^{(M)}, x^{(M')}, p, \cdot \right) + \mathcal{E}_{\mathcal{M}}(x, p, \cdot). \quad (8)$$

Refer to Section 5.1 for the similar construction in terms of multimodal Laplacians. See term-wise explanations[A.6] in Appendix A.

Assumption 4 (Feature maps for boundedness)**.** *Using the results of Moore–Aronszajn theorem Aronszajn (1950), for a positive definite kernel* K_M *in a measurable output space* $(\mathcal{X}, \mathcal{F}_{\mathcal{X}}, \mu)$ *aligned with Section 3.1, let* $\Phi_M : \mathcal{X}_M \to \mathcal{H}_M$ *be its feature map treated as infinite-dimensional linear operator for each modality* $M \in \mathcal{M}$ *under the constraint of boundedness:* $\sup_{x^{(M)} \in \mathcal{X}_M} \|\Phi_M(x^{(M)})\|_{\mathcal{H}_M} < \infty$*. (See explicit justification*[A.7] *in Appendix A.)*

For each modality M, the (fixed) embedding pipeline with an implicit kernel[A.8] in a higher-dimensional RKHS induces $\Phi_M : \mathcal{X}_M \to \mathcal{H}_M$ such that $\langle \Phi_M(x_1), \Phi_M(x_2) \rangle_{\mathcal{H}_M} = K_M(x_1, x_2)$.

Assumption 5 (Prompt embeddings)**.** *Let* $(\mathcal{P}, \mathcal{F}_{\mathcal{P}}, \nu)$ *be a measurable space on prompts with* ν *being finite. For each modality* $M \in \mathcal{M}$*, the prompt embedding* $\Psi_M : \mathcal{P} \to \mathcal{H}_M$ *satisfies boundedness:* $\sup_{p \in \mathcal{P}} \|\Psi_M(p)\|_{\mathcal{H}_M} < \infty$ *and stability:* Ψ_M *is continuous (equivalently, Lipschitz with finite constant* $\mathrm{Lip}(\Psi_M)$*) in the chosen topology/ metric on* $\mathcal{P}$*. (See explicit justification*[A.7] *in Appendix A.)*

Assumption 6 (Output distribution in Boltzman form). *We view $f_p(x)$ as a normalized surrogate over candidate outputs or latent representations with respect to a finite (or bounded) base measure μ. Under bounded embeddings and compact support (or bounded energy), the partition function $Z(p, \mathcal{T}_t)$ is finite, making Eq. (9) well-defined. (See explicit justification[A.9] in Appendix A.)*

Lemma 1 (Joint measurability of cross inner products). *If $\Phi_M : (\mathcal{X}_M, \mathcal{F}_{\mathcal{X}_M}) \to (\mathcal{H}_M, \mathcal{B}(\mathcal{H}_M))$ and $\Psi_M : (\mathcal{P}, \mathcal{F}_{\mathcal{P}}) \to (\mathcal{H}_M, \mathcal{B}(\mathcal{H}_M))$ are Bochner measurable into a separable Hilbert space $\mathcal{H}_M$ where $\mathcal{B}(\mathcal{H}_M)$ denotes the Borel $\sigma-$algebra generated by the open sets of $\mathcal{H}_M$ under its norm topology, then $(x, p) \mapsto \langle \Phi_M(x), \Psi_M(p) \rangle_{\mathcal{H}_M}$ is measurable on $\mathcal{F}_{\mathcal{X}_M} \otimes \mathcal{F}_{\mathcal{P}}$.*

Proof sketch: Bochner measurability of Φ_M and Ψ_M implies strong measurability into $\mathcal{B}(\mathcal{H}_M)$; hence $(x, p) \mapsto (\Phi_M(x), \Psi_M(p))$ is measurable on the product σ–algebra. Detailed proof is found in Appendix B.2. □

Theorem 2 (Multimodal energy-based hallucination formalism). *Between the output and prompt spaces, let the residuals $\mathsf{r}_M(x, p) := \Phi_M(x^{(M)}) - \Psi_M(p) \in \mathcal{H}_M$ be defined for at least two modalities $|\mathcal{M}| \geq 2$. For each M, let there be a bounded, self-adjoint, positive semi-definite (PSD) linear operator A_M on $\mathcal{H}_M$ and for $M \neq M'$, some $B_{MM'} : \mathcal{H}_{M'} \to \mathcal{H}_M$ which is a bounded linear symmetric cross-operator and a controlled factorization $B_{MM'} = A_M^{1/2} R_{MM'} A_{M'}^{1/2}$, subject to $\|R_{MM'}\| \leq 1$, being a symmetric contraction (e.g., Hilbert-Schmidt). Given this, if the output distribution $f_p(x)$ assumes the Boltzmann form for any temperature $\mathcal{T}_t \in \mathbb{R}_{\geq 0}$ dependent on time $t \in \mathbb{R}^+$:*

$$f_p(x) = (Z(p, \mathcal{T}_t))^{-1} \exp\big(-\mathcal{E}(x, p)/\mathcal{T}_t\big), \text{ where, } Z(p, \mathcal{T}_t) = \int_{\mathcal{X}} \exp\big(-\mathcal{E}(x, p)/\mathcal{T}_t\big)\, d\mu(x) \quad (9)$$

is the normalizing partition function, then the total energy noted in Eq. (8), for $(x, p) \in \mathcal{X} \times \mathcal{P}$, takes the form that is measurable, non-negative and satisfies canonical instances; given by:

$$\mathcal{E}(x, p) = \sum_{M \in \mathcal{M}} \langle \mathsf{r}_M, A_M\, \mathsf{r}_M \rangle_{\mathcal{H}_M} + \frac{2}{|\mathcal{M}| - 1} \sum_{\substack{M, M' \in \mathcal{M} \\ M \neq M'}} \langle A_M^{1/2}\, \mathsf{r}_M,\, R_{MM'}\, A_{M'}^{1/2}\, r_{M'} \rangle + \mathcal{E}_{\mathcal{M}}\,, \quad (10)$$

where the first and second terms on r.h.s are $\mathcal{E}_M$ and $\mathcal{E}_{MM'}$ respectively, while the last term being $\mathcal{E}_{\mathcal{M}}(x, p) = \left\| \bigotimes_{M \in \mathcal{M}} \Phi_M(x^{(M)}) - \bigotimes_{M \in \mathcal{M}} \Psi_M(p) \right\|^2_{\otimes \mathcal{H}_M}$ as a squared distance in composite RKHS, so it's measurable and nonnegative.

Proof sketch. We stack $r = (\mathsf{r}_M)_M$ and define the block operator $\mathcal{A}$ with diagonals A_M and off–diagonals $A_M^{1/2} R_{MM'} A_{M'}^{1/2}$. Since $A_M \succeq 0$, $R_{M'M} = R_{MM'}^*$, and $\|R_{MM'}\| \leq 1$, standard Cauchy–Schwarz/Schur arguments give $\mathcal{A} \succeq 0$; hence $\langle r, \mathcal{A} r \rangle \geq 0$ equals the first two terms of Eq. (10). The joint term is a single scalar for 3 modalities, but a tensor for > 3 modalities, thus ≥ 0. Measurability follows from Bochner measurability and continuity of bounded linear maps/inner products (refer to Lemma 1). Under the stated integrability/finite–measure conditions, the partition function in Eq. (9) is finite, so f_p is well-defined. Detailed proof is found in Appendix B.3. □

Corollary 1 (Excess-energy hallucination functional). *In line with Theorems 1 & 2, we leverage Eq. (10) to identify the hallucination energy in an MLLM:*

$$\mathcal{E}_{\text{hall}}^{\text{multi}}(x, p, \cdot) = \Big(\mathcal{E}(x, p, \cdot) - \mathcal{E}_{\mathcal{K}}(x, p, \cdot)\Big)_+ \mathbf{1}_{\{x \notin \mathcal{K}\}}. \quad (11)$$

where $\mathcal{E}(x, p, \cdot)$ is the total energy term at $\mathcal{X}$ and $\mathcal{E}_{\mathcal{K}}(x, p, \cdot)$ is the same restricted at $\mathcal{K}$.

Proof. This particular Corollary does not require any explicit proof as this is merely an identification done by the authors in line with the results obtained in Theorem 1. □

5 Main Results: Proposed Framework

In this section we develop the spectral representation that underpins our main results (Figure 2). We reformulate the multimodal hallucination energy $\mathcal{E}_{\text{hall}}^{\text{multi}}$ (refer to Eq. (11)) within standard spectral graph theory Chung (1997). This lets us relate the Boltzmann normalization of model outputs to eigenmodes of a multimodal semantic graph Laplacian, which in turn yields principled mode-wise bounds on hallucination energy.

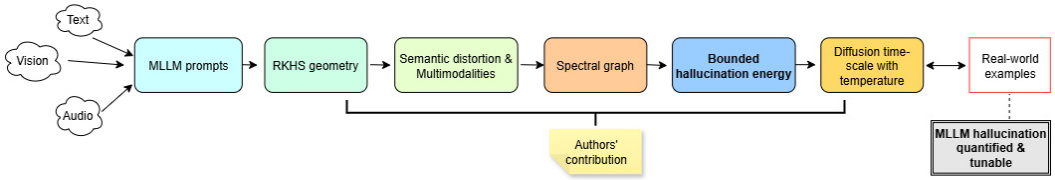


Figure 2: Pipeline for hallucination quantification in MLLMs. For an intuition-building case-study of an image–caption example for an MLLM, see comments[A.10] in Appendix A.

5.1 Semantic Graph and Multimodal Laplacian

Let a time-indexed, temperature-modulated multimodal semantic knowledge graph at an instant t be:

$$G_{\mathcal{T}_t} = (\mathcal{V}, E, W_{\mathcal{T}_t}), \quad \mathcal{V} \subseteq \mathbb{N}, \quad E \subseteq \mathcal{V} \times \mathcal{V}, \quad W_{\mathcal{T}_t} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}; \quad \forall\, t \in \mathbb{R}^+, \tag{12}$$

with finite set of nodes $\mathcal{V}$ (semantic units), pairwise edges $E \subseteq \mathcal{V} \times \mathcal{V}$ (similarity relations), and symmetric non-negative adjacency weights $W_{\mathcal{T}_t}$ built from fixed embeddings, where temperature $\mathcal{T}_t \in \mathbb{R}_{\geq 0}$ controls the affinity bandwidths. Here, we adopt a single integrated multimodal graph $G_{\mathcal{T}_t}$ with modality encoded by the node-partitioning $\mathcal{V} = \biguplus_M \mathcal{V}_M$ and a symmetric PSD $W_{\mathcal{T}_t}$ structured on its elements $w_{\mathcal{T}_t}$ noted in Eq. (16) as hyperedge weights. See justification[A.11] and detailed construction of $W_{\mathcal{T}_t}$ in Appendix A. In the current prescription of $\mathcal{T}_t$-modulated graph, the RKHS $\mathcal{H}$ is associated with a positive-definite multimodal diffusion kernel $K_{\mathcal{T}_t}$ that induces graph feature map $\Upsilon : \mathcal{V} \to \mathcal{H}$ satisfying (application of Assumption 4 in knowledge-graphs)

$$K_{\mathcal{T}_t} \;:=\; \exp\!\big(-\tau\, \mathcal{L}^{\text{multi}}_{\mathcal{T}_t}\big)\,, \qquad \big\langle \Upsilon(\mathsf{v}),\, \Upsilon(\mathfrak{v}) \big\rangle_{\mathcal{H}} \;=\; K_{\mathcal{T}_t}(\mathsf{v}, \mathfrak{v}), \qquad \forall\, \mathsf{v}, \mathfrak{v} \in \mathcal{V}, \tag{13}$$

where $\tau \in \mathbb{R}^+$ is a diffusion time-scale and $\mathcal{L}^{\text{multi}}_{\mathcal{T}_t}$ is a multimodal graph Laplacian defined on the finite node set $\mathcal{V}$. As an extension from Eq. (7), the above equation is an application of Mercer's theorem Mercer (1909), see details[A.12] in Appendix A. How this construction of graph feature maps Υ defined on nodes $\mathsf{v}, \mathfrak{v}$ has an interconnection to the output feature maps $\Phi_M(x^{(M)})$ and prompt embeddings $\Psi_M(p)$, see justification[A.13] in Appendix A. We design the multimodal Laplacian as a non-negative combination of intra–, cross–, and joint–modal components: $\mathcal{L}^{\text{multi}}_{\mathcal{T}_t} = \sum_* \text{coeff}_*\, \mathcal{L}^{(*)}_{\mathcal{T}_t}$, where $* \in \{\text{intra}_M,\ \text{cross}_{MM'},\ \text{joint}_{\mathcal{M}}\}$ and the interaction coefficients: $\text{coeff}_{\text{intra}_M} = \alpha_M\ (\forall\, M \in \mathcal{M})$, $\text{coeff}_{\text{cross}_{MM'}} = \beta_{MM'}\ (\forall\, M, M' \in \mathcal{M})$, and $\text{coeff}_{\text{joint}_{\mathcal{M}}} = \gamma_{\mathcal{M}}$ are all $\mathbb{R}_{\geq 0}$. Each $\mathcal{L}^{(*)}_{\mathcal{T}_t}$ is a symmetric PSD Laplacian-block built on the same node set $\mathcal{V}$; full expressions can be found in Eq. (24) in Appendix A.11.

5.2 Spectral Decomposition and Energy Functional

To dis-entangle modality–specific, cross–modal, and joint–modal interactions and to study how hallucination energy propagates across the graph, we diagonalize the normalized multimodal Laplacian. Let $\{(\lambda_i(t), u_i(t))\}_{i=1}^{|\mathcal{V}|}$ be the eigenpairs of $\mathcal{L}^{\text{multi}}_{\mathcal{T}_t}$ with $0 = \lambda_1(t) \leq \lambda_2(t) \leq \cdots$ and orthonormal eigenvectors $\langle u_i(t), u_j(t) \rangle = \delta_{ij}$. See comments[A.14] in Appendix A. Then for all nodes $\mathsf{v} \in \mathcal{V}$:

$$\mathcal{L}^{\text{multi}}_{\mathcal{T}_t} = U(t)\Lambda(t)U(t)^\top = \sum_{i=1}^{|\mathcal{V}|} \lambda_i(t)\, u_i(t) u_i(t)^\top, \quad \Upsilon(\mathsf{v}; \mathcal{T}_t) = \sum_{i=1}^{|\mathcal{V}|} e^{-\frac{\tau}{2}\lambda_i(t)}\, \langle u_i(t), \delta_{\mathsf{v}} \rangle\, u_i(t), \tag{14}$$

where $U(t) = [u_1(t) \cdots u_{|\mathcal{V}|}(t)]$, $\Lambda(t) = \text{diag}(\lambda_1(t), \ldots, \lambda_{|\mathcal{V}|}(t))$ and $\delta_{\mathsf{v}} \in \mathbb{R}^{|\mathcal{V}|}$ is the Kronecker delta at v. (We reserve $\mathsf{v}, \mathfrak{v}, ..$ for graph nodes and $i, j, ..$ for Laplacian modes; both index sets have size $|\mathcal{V}|$.) For output & prompt nodes $(\mathsf{v}_x, \mathfrak{v}_p) \in \mathcal{V}$ and, more generally, any graph signal $s \in \mathbb{R}^{|\mathcal{V}|}$,

$$\big\|\Upsilon(\mathsf{v}_x; \mathcal{T}_t) - \Upsilon(\mathfrak{v}_p; \mathcal{T}_t)\big\|^2_{\mathcal{H}} = \sum_{i=1}^{|\mathcal{V}|} e^{-\tau\lambda_i(t)} \big|\langle u_i(t), \delta_{\mathsf{v}_x} - \delta_{\mathfrak{v}_p} \rangle\big|^2, \quad \langle s,\, \mathcal{L}^{\text{multi}}_{\mathcal{T}_t}\, s \rangle = \sum_{i=1}^{|\mathcal{V}|} \lambda_i(t) \big|\langle u_i(t), s \rangle\big|^2. \tag{15}$$

A quick algebraic manipulation with Eqs. (15) plugged back into (10) gives the spectral form of total energy: $\mathcal{E}(x,p;\mathcal{T}_t) = \sum_* \sum_{i=1}^{|\mathcal{V}|} \mathrm{coeff}_* \, \mathsf{E}_i^{(*)}(x,p,t)$, where each $\mathsf{E}_i^{(*)}$ depends explicitly on $\lambda_i(t)$ and $u_i(t)$. See Eq. (57) in Appendix C.1 for details.

5.3 Spectral bounds on hallucination, and time-tecay

Here, we obtain: (i) quantitative bounds that control the scope of hallucination in an MLLM; (ii) an evolution of hallucinations in diffusion time with tunable temperature. The extended derivations of each expression below can be found in Appendix C.2.

Node-level score and pairwise dissimilarity. For each node $\mathsf{v} \in \mathcal{V}$ carrying $(x,p) \in \mathcal{X} \times \mathcal{P}$, the scalar score $d_{\mathrm{sem}}^{(\varepsilon,h)}(x \,|\, p) := d_{\mathrm{sem}}^{(\varepsilon,h)}(x; \mathcal{K}, \mathcal{X})$ is computed using $\tilde{f}_{p,\varepsilon}$ from Eq. (6). A symmetric, nonnegative prompt-aware dissimilarity between $\mathsf{v}_a \sim (x_a, p_a)$ and $\mathsf{v}_b \sim (x_b, p_b)$ is then defined by $\widehat{d}_{\mathrm{sem}}(\mathsf{v}_a, \mathsf{v}_b) := \big| d_{\mathrm{sem}}^{(\varepsilon,h)}(x_a \,|\, p_a) - d_{\mathrm{sem}}^{(\varepsilon,h)}(x_b \,|\, p_b) \big|$ and combining it with Eq. (26) yields

$$w_{\mathcal{T}_t}(e) = \mathbf{1}_{\{e \in E^{(*)}\}} \exp\Big(-\eta_* \Big(\sum_{1 \le a,b \le r(e)} |\, \Delta_{\varepsilon,h}(x_a \mid p_a) - \Delta_{\varepsilon,h}(x_b \mid p_b)\,| \Big) \Big/ \sum_{a=1}^{r(e)} \mathcal{T}_t(\mathsf{v}_a) \Big). \tag{16}$$

Here $r(e) := |e|$ is the hyperedge cardinality (Eq. (24)), and $\eta_* > 0$ is the modality–aware permutation factor (Eq. (26)). The derivation of $\Delta_{\varepsilon,h}(x \mid p)$ is found via Eq. (27) in Appendix A.11.

Courant–Fischer bounds for hallucination. Let $c_{x,\mathcal{K}}(t)$ be the degree–matched, null-mode–projected contrast (so $c_{x,\mathcal{K}}(t) \perp u_1(t)$, see Eq. (58)) and given the diffusion operator $\exp\big(-2\tau\, \mathcal{L}_{\mathcal{T}_t}^{\mathrm{multi}}\big)$, we get the semantic diffusion through spectral expansion $\big\langle c_{x,\mathcal{K}}(t), \exp\big(-2\tau\, \mathcal{L}_{\mathcal{T}_t}^{\mathrm{multi}}\big)\, c_{x,\mathcal{K}}(t) \big\rangle = \sum_{i=2}^{|\mathcal{V}|} e^{-2\tau\lambda_i(t)} \big| \langle u_i(t), c_{x,\mathcal{K}}(t) \rangle \big|^2$. By Courant–Fischer principle Horn & Johnson (2013), we get a pure spectral sandwich:

$$e^{-2\tau\,\lambda_{\max}(t)} \, \|c_{x,\mathcal{K}}(t)\|^2 \le \big\langle c_{x,\mathcal{K}}(t), \exp\big(-2\tau\, \mathcal{L}_{\mathcal{T}_t}^{\mathrm{multi}}\big)\, c_{x,\mathcal{K}}(t) \big\rangle \le e^{-2\tau\,\lambda_2(t)} \, \|c_{x,\mathcal{K}}(t)\|^2. \tag{17}$$

By Eq. (57), the full energy is a nonnegative linear combination of blockwise spectral terms, therefore the energy difference admits the eigen-expansion while its spectral weights lie in a bound:

$$\mathcal{E}(x,p;\mathcal{T}_t) - \mathcal{E}_{\mathcal{K}}(x,p;\mathcal{T}_t) = \sum_{i=2}^{|\mathcal{V}|} \zeta_i(t,\tau) \big| \langle u_i(t), c_{x,\mathcal{K}}(t) \rangle \big|^2, \quad m(t)\, e^{-2\tau\,\lambda_i(t)} \le \zeta_i(t,\tau) \le M(t), \tag{18}$$

where $\zeta_i(t,\tau) \ge 0$ and $(m(t),\ M(t)) \in (0,\infty)$; see Eq.(63) for details. By Eqs. (11), (17) and (18),

$$m(t)\, e^{-2\tau\,\lambda_{\max}(t)} \, \|c_{x,\mathcal{K}}(t)\|^2 \, \mathbf{1}_{\{x \notin \mathcal{K}\}} \le \mathcal{E}_{\mathrm{hall}}^{\mathrm{multi}}(x,p,\cdot) \le M(t)\, e^{-2\tau\,\lambda_2(t)} \, \|c_{x,\mathcal{K}}(t)\|^2 \, \mathbf{1}_{\{x \notin \mathcal{K}\}}. \tag{19}$$

Calibration-compatible lower envelope for hallucination time-scale. Let $\widehat{m}_{\mathrm{GT}}(t)$ denote the Good–Turing "missing-mass" estimate for the model f_p over $\mathcal{X} \setminus \mathcal{K}$ at time t (computed on the current prompt-conditioned sample window), and we set the calibrated lower-bound aligned with Kalai & Vempala (2024) as $\vartheta_{\mathrm{KV}}(t) := \xi\, \widehat{m}_{\mathrm{GT}}(t)$ for some fixed $\xi \in (0,1]$. A time-indexed diffusion/temperature profile $\tau = \tau(t)$ is chosen to embed that envelope by identifying

$$m(t)\, e^{-2\tau(t)\,\lambda_{\max}(t)} \, \|c_{x,\mathcal{K}}(t)\|^2 \ge \vartheta_{\mathrm{KV}}(t) \iff \tau(t) \le \frac{1}{2\,\lambda_{\max}(t)} \log\!\left(\frac{m(t)\, \|c_{x,\mathcal{K}}(t)\|^2}{\vartheta_{\mathrm{KV}}(t)} \right). \tag{20}$$

Eq. (20) operationalizes Kalai–Vempala's calibrated lower bound within our spectral framework, guaranteeing the bound is met (and dominated tunably) by the diffusion–Laplacian control.

Time–decay of hallucination energy. From Eq. (19), $\mathcal{E}_{\mathrm{hall}}^{\mathrm{multi}}$ is nonincreasing in τ and decays to 0 as $\tau \to \infty$ at a rate sandwiched between $e^{-2\tau\lambda_{\max}}$ and $e^{-2\tau\lambda_2}$. When the block responses are diffusion–monotone (standard for normalized kernels), the pointwise derivative exists (for $x \notin \mathcal{K}$)

$$\frac{d}{d\tau}\, \mathcal{E}_{\mathrm{hall}}^{\mathrm{multi}}(x,p,\cdot) = -2 \sum_{i=2}^{|\mathcal{V}|} \lambda_i(t)\, \zeta_i(t,\tau) \big| \langle u_i(t), c_{x,\mathcal{K}}(t) \rangle \big|^2 \searrow 0, \tag{21}$$

which is compatible with Eq. (18) that makes it implementation-ready.

6 EXPERIMENTS

Code base. <REPO>. The exact configs used for each run are shipped under `configs/`.

6.1 DATASETS AND MODELS

We evaluate 3 multimodal datasets crossed with 3 inference stacks, yielding 9 panels (Fig. 3).

Datasets.

- **COCO Captions** (val2017): large image–text captioning split; diverse everyday scenes.
- **VQAv2**: balanced visual question answering; short free-form answers grounded in images.
- **AudioCaps**: audio–text captioning from YouTube clips; non-visual acoustic events.

Note. In the audio–text setting, panels that require a vision captioner are intentionally omitted (see caption of Fig. 3).

Models (inference stacks).

- `CLIP+Whisper+T5`: vision embeddings (CLIP) + audio embeddings (Whisper) + text LM (T5) for scoring/logits.
- `BLIP+CLIP+Whisper`: BLIP captioner for image semantics (paired with CLIP features) + Whisper for audio; *vision-dependent*, so the AudioCaps cross is blank by design.
- `SigLIP+Whisper+T5`: SigLIP vision encoder + Whisper + T5; same interface as the first stack.

Sources. Pulled from HuggingFace Hub (private tokens); `HF_HOME` and `HF_TOKEN` are set at runtime.

Algorithm 1: KL-SMOOTHED MULTIMODAL HALLUCINATION (per prompt p)

Input: $\mathcal{K}$; μ; K_h; ε, ρ; blocks $\{\mathcal{I}^{(*)}, E^{(*)}, \omega_*, \eta_*\}$; $\mathcal{T}_t$; τ; $\{\Phi_M, \Psi_M\}_{M\in\mathcal{M}}$; $\{A_M\}_M, \{R_{MM'}\}_{M\neq M'}$

Output: $d_{\mathrm{sem}}^{(\varepsilon,h)}(x \mid p)$; $w_{\mathcal{T}_t}(e)$; $\mathcal{L}_{\mathcal{T}_t}^{\mathrm{multi}}$; $K_{\mathcal{T}_t}$; $\mathcal{E}_{\mathrm{hall}}^{\mathrm{multi}}(x,p)$ and CF-bounds

Form $\tilde{f}_{p,\varepsilon} = (1-\varepsilon)f_p + \varepsilon\rho$ and $\tilde{f}_{p,\varepsilon}^{\mathcal{K}}$; compute $d_{\mathrm{sem}}^{(\varepsilon,h)}(x \mid p)$ by Eq. (6). (Thm. 1);

Compute $\mathsf{r}_M(x,p)$; store $\{A_M, B_{MM'}\}$ for energy in Eq. (10). (Thm. 2);

Set $\Delta_a = d_{\mathrm{sem}}^{(\varepsilon,h)}(x_a \mid p)$ and $w_{\mathcal{T}_t}(e)$ by Eq. (26); build $\mathcal{L}_{\mathcal{T}_t}^{(*)}$ via Eq. (24) and assemble $\mathcal{L}_{\mathcal{T}_t}^{\mathrm{multi}}$ via Eq. (25).;

Compute $K_{\mathcal{T}_t}$ and set graph features $\Upsilon(\mathsf{v})$ so that $\langle \Upsilon(\mathsf{v}), \Upsilon(\mathfrak{v}) \rangle_{\mathcal{H}} = K_{\mathcal{T}_t}(\mathsf{v}, \mathfrak{v})$ (Eq. (14)).;

Form $c_{x,\mathcal{K}}(t)$ by Eq. (58) and apply bounds in Eq. (17).;

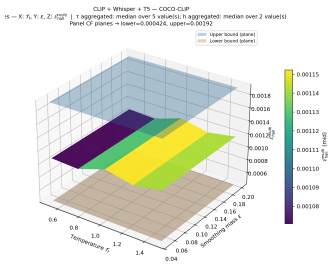
Evaluate $\mathcal{E}(x,p)$ via Eq. (10); set $\mathcal{E}_{\mathrm{hall}}^{\mathrm{multi}}$ by Eq. (11); report Courant–Fischer bounds in Eq. (19) plus KV/Good–Turing calibration via Eq. (20)).;

return $d_{\mathrm{sem}}^{(\varepsilon,h)}$, $w_{\mathcal{T}_t}(e)$, $\mathcal{L}_{\mathcal{T}_t}^{\mathrm{multi}}$, $K_{\mathcal{T}_t}$, $\mathcal{E}_{\mathrm{hall}}^{\mathrm{multi}}$ *(with bounds)*

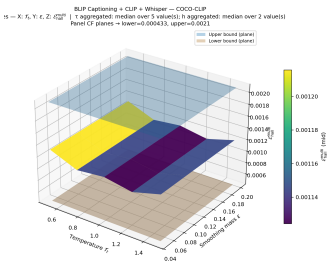
Algorithm	COCO AUROC / AUPRC	VQAv2 AUROC / AUPRC	AudioCaps AUROC / AUPRC	Avg. AUROC / AUPRC
Entropy	0.81 / 0.79	0.78 / 0.75	0.74 / 0.70	0.78 / 0.75
MaxProb	0.82 / 0.81	0.80 / 0.77	0.76 / 0.72	0.79 / 0.77
Margin	0.83 / 0.82	0.81 / 0.78	0.77 / 0.74	0.80 / 0.78
$d_{\mathrm{sem}}^{(\varepsilon,h)}$ (ours)	**0.86 / 0.84**	**0.84 / 0.81**	**0.80 / 0.77**	**0.83 / 0.81**

Model	COCO median (lo / hi)	VQAv2 median (lo / hi)	AudioCaps median (lo / hi)	Avg. median	Throughput↑ ex/s	Asymp.
CLIP+Whisper+T5	2.11 (0.42 / 3.05)	2.23 (0.50 / 3.28)	2.35 (0.55 / 3.50)	2.23	**420**	$O(\lvert E\rvert + N\log k + md)$
BLIP+CLIP+Whisper	1.98 (0.40 / 2.90)	2.05 (0.48 / 2.96)	—	2.02	360	$O(\lvert E\rvert + N\log k + md)$
SigLIP+Whisper+T5	**1.92** (0.38 / 2.85)	**1.99** (0.45 / 2.90)	**2.08** (0.50 / 3.05)	**2.00**	400	$O(\lvert E\rvert + N\log k + md)$

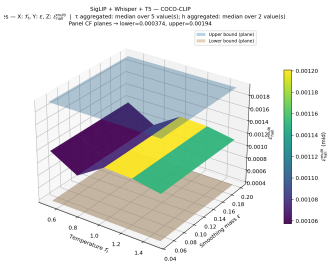
Table 1: **(a) Detection (AUROC/AUPRC) and (b) Energy diagnostics with runtime. Bold** = column-best; in (b), lower median energy is better and throughput (ex/s) higher is better. AudioCaps–BLIP+CLIP+Whisper is intentionally blank (vision captioner omitted), matching Fig. 3.



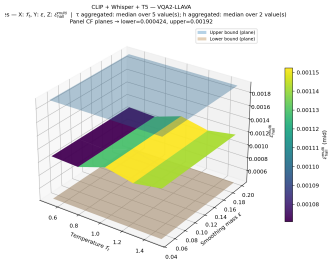
(a) COCO–CLIP+Whisper+T5



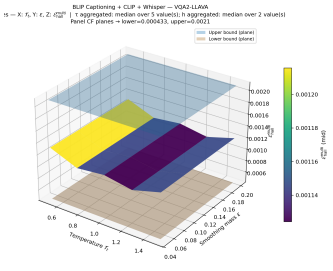
(b) COCO–BLIP+CLIP+Whisper



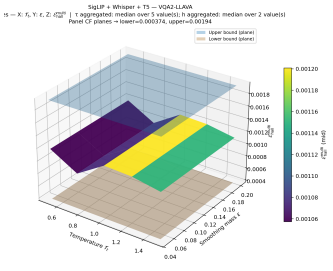
(c) COCO–SigLIP+Whisper+T5



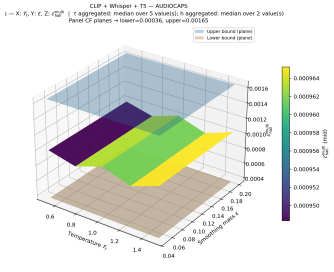
(d) VQAv2–CLIP+Whisper+T5



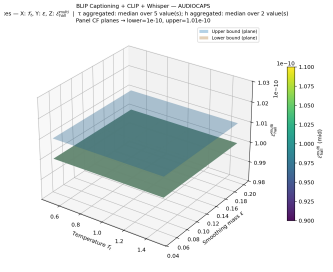
(e) VQAv2–BLIP+CLIP+Whisper



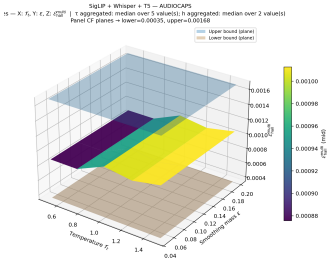
(f) VQAv2–SigLIP+Whisper+T5



(g) AudioCaps–CLIP+Whisper+T5



(h) AudioCaps–BLIP+CLIP+Whisper



(i) AudioCaps–SigLIP+Whisper+T5

Figure 3: **CF-bounded hallucination energy surfaces (9 panels).** Each 3D surface shows $\mathcal{E}_{\text{hall}}^{\text{multi}}$ over temperature $\mathcal{T}_t$ (X) and smoothing mass ε (Y), clamped between two panel-specific parallel planes marking the Courant–Fischer lower (strictly > 0) and upper bounds (Z). Other hyperparameters (τ, h) are aggregated by median, consistent across panels. *Note:* the **Audio-Caps–BLIP+CLIP+Whisper** panel may appear blank if the BLIP vision backbone is intentionally omitted for the audio–text setup; this is expected and documented in our pipeline.

6.2 Metrics and Evaluation

We report AUROC/AUPRC for hallucination detection using $d_{\text{sem}}^{(\varepsilon,h)}$ against entropy, max-probability, and margin baselines, and summarize CF-bounded energy surfaces (lower is better) with temperature/ε trends matching theory. *Details about the baselines and all remaining protocol & design, and compute details are in Appendix D.*

7 Conclusion and Future Work

We proposed a reference-free, KL–smoothed information gap with hypergraph–spectral control: the score is 0 on $\mathcal{K}$ and strictly > 0 off $\mathcal{K}$, admits Courant–Fischer (CF) bounds, and integrates Good–Turing/KV calibration. Compact Colab runs (COCO/VQAv2/AudioCaps $\times$ CLIP/BLIP/SigLIP stacks) show consistent gains over entropy/margin and interpretable temperature/τ decay. A joint tuning of $(\varepsilon, h, \mathcal{T}_t, \tau)$ with uncertainty can be the next direction.

REFERENCES

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950. doi: 10.2307/1990404.

T. Bai, Y. Zhang, X. Lin, Q. Sun, and C. Wu. Multimodal hallucinations: A survey of causes, metrics, and mitigation. Technical report, arXiv, 2024. arXiv:2404.18930.

Robert G. Bartle. *The Elements of Integration and Lebesgue Measure*. Wiley-Interscience, corrected reprint of the 1st ed. (1966) edition, 1995. Introduction emphasizing Lebesgue measure on Rn with clarity and examples.

Jakub Binkowski, Denis Janiak, Albert Sawczyn, Bogdan Gabrys, and Tomasz Kajdanowicz. Spectral characterization of hallucination in large language models. *arXiv preprint arXiv:2502.17598*, 2025.

Sébastien Bubeck, Varsha Chandrasekaran, Ran Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Percy Lee, Yu Li, Scott Lundberg, Harsha Nori, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. Technical report, arXiv, 2023. arXiv preprint arXiv:2303.12712.

Fan R. K. Chung. *Spectral Graph Theory*, volume 92 of *CBMS Regional Conference Series in Mathematics*. American Mathematical Society, Providence, RI, 1997. ISBN 978-0-8218-0315-8. doi: 10.1090/cbms/092.

R. R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 2006.

Joseph Diestel and Jr. John J. Uhl. *Vector Measures*, volume 15 of *Mathematical Surveys and Monographs*. American Mathematical Society, 1977.

Peng Ding, X. Hu, H. Li, X. Chen, and H. Ji. Hallu-pi: Benchmarking hallucinations under perturbed inputs for large language models. In *ICLR*, 2024.

Tao Feng, Yihang Sun, and Jiaxuan You. Grapheval: A lightweight graph-based llm framework for idea evaluation. *arXiv preprint arXiv:2503.12600*, 2025.

Assaf Gerner, Netta Madvil, Nadav Barak, Alex Zaikman, Jonatan Liberman, Liron Hamra, Rotem Brazilay, Shay Tsadok, Yaron Friedman, Neal Harow, Noam Bressler, Shir Chorev, and Philip Tannor. Orion grounded in context: Retrieval-based method for hallucination detection. *arXiv preprint arXiv:2504.15771*, 2025. doi: 10.48550/arXiv.2504.15771. URL https://arxiv.org/abs/2504.15771.

J. Han, J. Kossen, M. Razzak, L. Schut, S. Malik, and Y. Gal. Semantic entropy probes: Robust and cheap hallucination detection in large language models. In *ICML Workshop on Foundation Models in the Wild*, 2024.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning Workshop*, 2015.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 2 edition, 2013. ISBN 978-0-521-54823-6.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. doi: 10.1145/3571730.

C. Jiang, B. Qi, X. Hong, D. Fu, Y. Cheng, F. Meng, M. Yu, B. Zhou, and J. Zhou. On large language models' hallucination with regard to known facts. In *NAACL*, 2024.

Adam Tauman Kalai and Santosh S. Vempala. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 160–171, 2024. doi: 10.1145/3618260.3649777.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*, 2025. doi: 10.48550/arXiv.2509.04664. URL https://arxiv.org/abs/2509.04664.

Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv:1808.06226*, 2018. URL https://arxiv.org/abs/1808.06226.

Eric Le Merrer and Gilles Trédan. Llms hallucinate graphs too: A structural perspective. *arXiv preprint arXiv:2409.00159*, 2024.

Seongmin Lee, Hsiang Hsu, and Chun-Fu Chen. Llm hallucination reasoning with zero-shot knowledge test. *arXiv preprint arXiv:2411.09689*, 2024.

Jiaan Li et al. Halueval: A large-scale hallucination evaluation benchmark for llms. In *EMNLP*, 2023. URL https://aclanthology.org/2023.emnlp-main.397/.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of EMNLP*, 2023.

Justin Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1906–1919, 2020.

James Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society A*, 209:415–446, 1909.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *NeurIPS*, 2002.

Chenxi Niu et al. Ragtruth: A hallucination corpus for developing more truthful systems. In *ACL*, 2024. URL https://aclanthology.org/2024.acl-long.585/.

S. Park, X. Du, M.-H. Yeh, H. Wang, and S. Li. Steer llm latents for hallucination detection. Technical report, ICML Poster, 2025.

Alec Radford et al. Learning transferable visual models from natural language supervision. *arXiv:2103.00020*, 2021a. URL https://arxiv.org/abs/2103.00020.

Alec Radford et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 2021b. URL https://proceedings.mlr.press/v139/radford21a.html.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2007.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S. M. Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. The troubling emergence of hallucination in large language models: An extensive definition, quantification, and prescriptive remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2541–2573, Singapore, 2023. Association for Computational Linguistics.

Anna Rohrbach et al. Object hallucination in image captioning. In *EMNLP*, 2018. URL https://arxiv.org/abs/1809.02156.

Albert Sawczyn, Jakub Binkowski, Denis Janiak, Bogdan Gabrys, and Tomasz Kajdanowicz. Factselfcheck: Fact-level black-box hallucination detection for llms. *arXiv preprint arXiv:2503.17229*, 2025.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *ACL*, 2016. URL https://arxiv.org/abs/1508.07909.

G. Sun, P. Manakul, A. Liusie, K. Pipatanakul, C. Zhang, P. Woodland, and M. Gales. Crosscheckgpt: Universal hallucination ranking for multimodal foundation models. In *NeurIPS Workshop on Next Gen Multimodal Models*, 2024.

Terence Tao. *An Introduction to Measure Theory*, volume 126 of *Graduate Studies in Mathematics*. American Mathematical Society, 2011. Covers sigma-algebras, outer measures, completeness, and constructions of measure.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. Transformer Circuits Thread, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/. Anthropic Interpretability Team.

Ashish Vaswani et al. Attention is all you need. In *NeurIPS*, 2017. URL https://arxiv.org/abs/1706.03762.

X. Wang, Y. Yan, L. Huang, X. Zheng, and X. Huang. Hallucination detection for generative large language models by bayesian sequential estimation. In *EMNLP*, pp. 15361–15371, 2023.

Christopher K. I. Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2001.

J. Xie, C. Zhang, and M. Li. Spectral characterization of hallucination in large language models. Technical report, arXiv, 2025. arXiv:2502.17598.

Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In *NeurIPS*, 2004.

Y. Zhang et al. Semantics at an angle: When cosine similarity works until it doesn't. *arXiv:2504.16318*, 2025. URL https://arxiv.org/abs/2504.16318.

Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf. Learning with hypergraphs: Clustering, classification, and embedding. *Advances in Neural Information Processing Systems*, 19:1601–1608, 2006.