APPENDIX

In this section, we provide elaboration on footnotes, extended derivations of our Theorems, some supplementary mathematical results, and details of experimental validation.

A TECHNICAL NOTES AND EXTENDED COMMENTS

Here, we provide elaboration on footnotes/ some extended explanations.

A.1 MEASURABLE SETS AND σ-ALGEBRA

Any mathematical set can be equipped with a σ-algebra to form a measurable space, say, $\mathcal{X}$. The common choices are: (i) the power set if $\mathcal{X}$ is countable/ finite, (ii) the Borel σ-algebra if $\mathcal{X}$ is a topological space (e.g., continuous embeddings), (iii) Product σ-algebra if $\mathcal{X}$ is a product of spaces (e.g., sequences of tokens or multimodal outputs). For a measure space $(\mathcal{X}, \mathcal{F}_{\mathcal{X}}, \mu)$ and $1 \leq p < \infty$ (where p is the integrability exponent, not to be clashed with "prompts"), the space $L^p(\mathcal{X}, \mathcal{F}_{\mathcal{X}}, \mu)$ consists of (equivalence classes of) μ-measurable $q : \mathcal{X} \to \mathbb{R}$ with $\int_{\mathcal{X}} |q(x)|^p \, d\mu(x) < \infty$; the norm is $\|q\|_{L^p} := \left(\int |q|^p d\mu \right)^{1/p}$. For $p = \infty$, L^∞ denotes essentially bounded functions with norm $\|q\|_{L^\infty} := \operatorname{ess\,sup}_x |q(x)|$. In particular, L^1 denotes integrable functions ($p = 1$).

A.2 GROUND TRUTH DISTRIBUTION

The absence of an exact analytical expression of $g(x)$ limits the direct interpretability, but provides a flexible framework for comparing the model outputs to the ground-truth via the functional and spectral metrics. This is used only as a theoretical reference for calibration/fidelity analyses representing the (idealized) generative distribution of facts/outputs as seen in Kalai & Vempala (2024).

A.3 DEFINITION OF KL-DIVERGENCE

For any two probability distributions $P_1(x)$ and $P_2(x)$, say defined over the same space $x \in \mathcal{X}$, the functional operator $D_{\text{KL}} \in \mathbb{R}_{\geq 0}$ refers to the KL divergence of $P_2(x)$ from the "true" reference or actual distribution $P_1(x)$ as:

$$D_{\text{KL}}\left(P_1(x) \parallel P_2(x)\right) = \sum_{x \in \mathcal{X}} P_1(x) \log \frac{P_1(x)}{P_2(x)}.$$

When x is a continuous random variable, $\sum_{x \in \mathcal{X}}$ is evidently replaced by $\int_{x=-\infty}^{\infty}$ with $P_1(x)$ & $P_2(x)$ by respective probability densities. More generally, if P_1 & P_2 are probability measures on a measurable space $\mathcal{X}$, then

$$D_{\text{KL}}\left(P_1 \parallel P_2\right) = \int_{x=-\infty}^{\infty} P_1(dx) \log \frac{P_1(dx)}{P_2(dx)},$$

where $\frac{P_1(dx)}{P_2(dx)}$ is the Radon–Nikodym derivative of P_1 w.r.t P_2.

A.4 ABSENCE OF THE "GROUND-TURTH"

Even without access to g, one can (i) estimate $\mathbb{P}_{f_p}(\mathcal{K})$ from samples, (ii) compute per-instance distortions via the log–likelihood ratio, and (iii) aggregate these into empirical bounds and diagnostics. In multimodal settings, the same decomposition localizes contributions by modality and by interaction (intra/cross/joint), enabling targeted interventions—e.g., modality-specific calibration, cross-modal consistency constraints, or temperature schedules—and straightforward experimental verification via ablations that track how $\mathbb{P}_{f_p}(\mathcal{K})$ and the induced distortions respond to each mitigation.

A.5 Modalities in Expanded Forms

In multi-modal settings, the LLM outputs involve textual (T), visual (V), audio (A) modalities and, for better understanding, Eq. (7) can also be re-written as:

$$\begin{aligned}
&\mathcal{X} : \mathcal{X}_T \times \mathcal{X}_V \times \mathcal{X}_A, \qquad x = (x^{(T)}, x^{(V)}, x^{(A)}), \qquad \mathcal{H} := \mathcal{H}_T \otimes \mathcal{H}_V \otimes \mathcal{H}_A, \\
&K(x_1, x_2) = K_T\left(x_1^{(T)}, x_2^{(T)}\right) \cdot K_V\left(x_1^{(V)}, x_2^{(V)}\right) \cdot K_A\left(x_1^{(A)}, x_2^{(A)}\right), \\
&\mathcal{P} : \mathcal{P}_T \times \mathcal{P}_V \times \mathcal{P}_A, \qquad p = (p^{(T)}, p^{(V)}, p^{(A)}).
\end{aligned} \tag{22}$$

A.6 Explanation on hallucination energy

As noted in Eq.(8) in Section 4.3, three terms are: (i) $\mathcal{E}_M$ encodes the intra-modal contributions, (ii) $\mathcal{E}_{MM'}$ captures the pairwise cross-modal terms, while (iii) $\mathcal{E}_{\mathcal{M}}$ being the joint contribution of all three modalities combined. For three modalities, (i) & (ii) form an energy matrix of order 3 with diagonals $\mathcal{E}_M$ and off-diagonals $\mathcal{E}_{MM'}$, while $\mathcal{E}_{\mathcal{M}}$ is a single joint term. With > 3 modalities, $\mathcal{E}_{\mathcal{M}}$ becomes a higher order tensor. This structure not only reveals which modality interactions contribute the most to the semantic drift $d_{\text{sem}}(x; \mathcal{K}, \mathcal{X})$, also enables deriving tight spectral bounds on hallucination energy, which would be impossible under a monolithic energy formulation.

A.7 Explicit justification for Assumptions 4 & 5

For Assumptions 4: this mirrors common practice: modern encoders (CLIP, BERT-style, vision backbones) apply normalization or LayerNorm, and we L2-normalize final vectors so magnitudes stay well-behaved. Bounded features make cosine/similarity scores comparable across modalities, prevent numerical outliers, and keep spectral/energy measures meaningful. In deployment, this is easy to enforce (normalize outputs) and verify (log histograms/max norms and alert on drift). Production stacks (vector DBs, ANN indices, faiss/scann) expect bounded vectors so cosine similarity behaves predictably and distances are comparable across batches and time.

- **Why we need this:** For numerical stability to prevents overflow/NaNs and keep the dot products/similarities in a usable range during training and evaluation and comparability across modalities to handle text & image embeddings simultaneously.
- **Real-world example:** Modern vision–language encoders (e.g., CLIP) explicitly L2-normalize image/text embeddings and use cosine similarity with temperature-scaled softmax, so representation norms are controlled by design; this makes cross-modal scoring numerically stable and comparable out of the box Radford et al. (2021a;b); Zhang et al. (2025).

For Assumption 5: it is reasonable to assume small prompt edits should not cause large representational jumps - matching real product needs for predictable UX, reproducible evaluation, and reduced prompt-sensitivity exploits. In practice, prompt encoders are compositions of linear layers + pointwise activations + norm layers; we also L2-normalize the final embedding.

- **Why we need this:** If "Adding a comma" or "Swapping a synonym" flips the model's answer, the system feels brittle. Stability is essential for predictability and debuggability.
- **Real-world example:** Text prompts are tokenized into a finite vocabulary (BPE/WordPiece/SentencePiece), and the transformer encoder maps these tokens through a sequence of standard layers to probabilities via softmax, yielding well-defined distributions on a discrete space—hence measurability is immediate and commonplace Vaswani et al. (2017); Sennrich et al. (2016); Kudo & Richardson (2018). Length caps, normalization, and regularization used in real systems keep prompt embeddings within reasonable ranges and make small paraphrases produce small representational changes, which is precisely the stability we assume.

A.8 Feature Maps in RKHS

RKHS theory is rooted in Hilbert space theory (inner product spaces of functions) and uses results like the Moore–Aronszajn theorem Aronszajn (1950)). In Measure Theory & Probability, when

kernels are used for distributions (e.g., kernel mean embeddings), the feature map connects to integration theory and probabilistic representations. In Machine Learning, the feature maps are used in kernel methods (in practice: SVMs, Gaussian processes, etc.), making this concept central to the theory of statistical learning (e.g., RKHS regularization). Let Φ_M be a feature map (i.e., identified as a function) such that

$$K_M\left(x_1^{(M)}, x_2^{(M)}\right) = \left\langle \Phi_M(x_1^{(M)}),\, \Phi_M(x_2^{(M)}) \right\rangle_{\mathcal{H}_M}, \tag{23}$$

embedding raw objects, say outputs (x_1, x_2), into the modality-specific RKHS $\mathcal{H}_M$. Instead of just outputs, it can very well mix with the inputs as well meaning: (x, p). Eq. (23) makes this RKHS $\mathcal{H}_M$ unique up to isometry according to the Moore–Aronszajn theorem.

In classical ML, we use "features" to describe the structured attributes of the input data (e.g., pixel values, word embeddings etc.). In the theory of kernels, the feature maps are abstract (possibly infinite), but they play the same role: they represent the data in a space where linear methods (dot products) can capture nonlinear similarities. Thus, Φ_M allows nonlinear learning algorithms to operate in a high-dimensional feature space of an MLLM via the kernel trick.

In practice, implementations typically compute K_M directly—or via finite approximations like Nyström Williams & Seeger (2001) or Random Fourier Features Rahimi & Recht (2007) - so Φ_M need not be explicitly materialized.

A.9 Justification for Assumption 6

In practice, an MLLM scores a finite candidate set $C(x, p)$ (beam/nucleus/reranked hypotheses) via logits or similarity, so with counting measure and energy $\mathcal{E} = -\text{logit}$ (or a bounded margin), the induced softmax probability $\text{prob.}(c \mid x, p; \mathcal{T}_t) \propto \exp(-\mathcal{E}(c)/\mathcal{T}_t)$ is exactly a Boltzmann distribution with finite partition function $Z = \sum_{c \in C} \exp(-\mathcal{E}(c)/\mathcal{T}_t)$—hence both operationally realistic and mathematically well-posed.

A.10 An Example (image–caption pair)

One can consider an MLLM generating a caption for an image. Let $\mathcal{X}$ be the space of all captions, with $\mathcal{K} \subseteq \mathcal{X}$ denoting those grounded in the image (e.g., "A cat on a sofa"), while f_p may also assign mass outside $\mathcal{K}$ to hallucinated captions (e.g., "A dog playing with a ball"). The hallucination divergence $D_{\mathrm{KL}}(g \,\|\, f_p)$ quantifies this deviation.

In this paper, as a part of our main theoretical contributions, we define a multimodal graph whose nodes are caption tokens T and image patches V, with edge weights $W_{\mathcal{T}_t}(i, j)$ computed from the fixed embeddings and modulated by a time-varying temperature $\mathcal{T}_t$. From these weights, we will define the normalized multimodal Laplacian $\mathcal{L}^{\text{multi}}_{\mathcal{T}_t}$ associated with a spectral grounding energy as the quadratic form of $\mathcal{L}^{\text{multi}}_{\mathcal{T}_t}$ evaluated on the residual feature field induced by our energy prescription. It helps reveal how hallucination energy is distributed across the modes (e.g., textual vs. cross-modal misalignment).

A.11 Graph notations and Adjacency Weights

In Eq. (12) noted in Section 5.1, $\mathcal{V}$ is the finite set of nodes, E is the set of edges, and $W_{\mathcal{T}_t}$ is a temperature-modulated, symmetric, non-negative, weighted adjacency matrix (zero diagonal) introduced to assign different weights to the edges (indexed by E). We consider either a node-wise local schedule $\mathcal{T}_t : \mathcal{V} \to \mathbb{R}^+$ in which the edge temperatures are combined symmetrically to keep $W_{\mathcal{T}_t}$ symmetric or a global scalar schedule ($\mathcal{T}_t$ constant over $\mathcal{V}$). Here, each node represents a semantic unit (e.g., concepts, tokens, ideas), and edges represent the semantic similarity. The multimodal structure is represented by a disjoint partition of the node set $\mathcal{V} = \biguplus_{M \in \mathcal{M}} \mathcal{V}_M$ and corresponding within- and cross-modal blocks of $W_{\mathcal{T}_t}$ which is constructed from fixed modality embeddings via temperature-controlled similarity functions. Lower $\mathcal{T}_t$ yields more localized (sharper) affinities; higher $\mathcal{T}_t$ diffuses those (or, in other words, induces more "noise"). This is a standard property under any temperature–scaled affinity constructions - e.g., Gaussian/RBF kernels with bandwidth proportional to $\mathcal{T}_t$ or softmax similarities with temperature $\mathcal{T}_t$ Ng et al. (2002); Coifman & Lafon (2006);

Zelnik-Manor & Perona (2004); Hinton et al. (2015); Chung (1997). Thus, the temperature $\mathcal{T}_t$ dynamically modulates the graph edge connectivity and semantic distortion d_{sem} noted in Theorem 1 and, being a time-indexed function, captures the semantic evolution or uncertainty drift across the graph nodes as knowledge updates over time t.

Here, we drop the explicit modality subscripts in Eq. (12), as the modality information is carried by a fixed partition of the vertex set $\mathcal{V} = \biguplus_{M\in\mathcal{M}} \mathcal{V}_M$ together with the block structure of the temperature–modulated weights $W_{\mathcal{T}_t}$, so we do not maintain separate graphs per modality. We assume $W_{\mathcal{T}_t}$ to be symmetric, non-negative, and zero on the diagonal, with $\mathcal{T}_t$ acting as a bandwidth/temperature schedule that controls the locality of affinities. From $W_{\mathcal{T}_t}$, we define the normalized multimodal Laplacian $\mathcal{L}^{\text{multi}}_{\mathcal{T}_t}$ in Section 5.1 and design it to be symmetric and PSD by construction; its spectral decomposition yields an orthonormal basis of eigenmodes together with nonnegative eigenvalues. We interpret each mode by its loadings on the partition $\{\mathcal{V}_M\}_{M\in\mathcal{M}}$: some modes are concentrated on a single modality (text, vision, or audio), while others are cross-modal mixtures that capture interactions between partitions. These modes serve as canonical coordinates for representing the residual signal induced by the energy model and for attributing hallucination energy across modality-specific and cross-modal directions. We use this spectral basis to define propagation in time (via diffusion generated by $\mathcal{L}^{\text{multi}}_{\mathcal{T}_t}$) and to derive mode-wise bounds that connect the Boltzmann formulation to spectral-graph structure in a implementable manner.

Hypergraph blocks and effective pairwise adjacency. To accommodate > 2 modalities, we construct each interaction block via the normalized hypergraph Laplacian Zhou et al. (2006):

$$\mathcal{L}^{(*)}_{\mathcal{T}_t} \;=\; \mathbf{I} \;-\; \big(\mathcal{D}^{(*)}_{\mathsf{v},\mathcal{T}_t}\big)^{-1/2} \underbrace{\big(\mathcal{I}^{(*)}\, W^{(*)}_{\mathcal{T}_t}\, (\mathcal{D}^{(*)}_{e,\mathcal{T}_t})^{-1}\, (\mathcal{I}^{(*)})^{\top}\big)}_{W^{*,\,\text{eff}}_{\mathcal{T}_t}} \big(\mathcal{D}^{(*)}_{\mathsf{v},\mathcal{T}_t}\big)^{-1/2},$$

$$\begin{aligned}
&\mathcal{D}^{(*)}_{\mathsf{v},\mathcal{T}_t} = \operatorname{diag}\big(\{\mathfrak{d}^{(*)}_{\mathcal{T}_t}(\mathsf{v})\}_{\mathsf{v}\in\mathcal{V}}\big), \qquad \mathfrak{d}^{(*)}_{\mathcal{T}_t}(\mathsf{v}) = \sum_{e\in E^{(*)}} w_{\mathcal{T}_t}(e)\, \mathcal{I}^{(*)}(\mathsf{v},e),\\
&\mathcal{D}^{(*)}_{e,\mathcal{T}_t} = \operatorname{diag}\big(\{r(e)\}_{e\in E^{(*)}}\big), \qquad r(e) = |e| \;\text{ (hyperedge cardinality)},\\
&\mathcal{I}^{(*)} \in \{0,1\}^{|\mathcal{V}|\times|E^{(*)}|} \text{ (node–hyperedge incidence)}, \quad W^{(*)}_{\mathcal{T}_t} = \operatorname{diag}\big(\{w_{\mathcal{T}_t}(e)\}_{e\in E^{(*)}}\big),\\
&\forall * \in \{\text{intra}_M,\, \text{cross}_{MM'},\, \text{joint}_{\mathcal{M}}\}, \quad \forall\, \mathsf{v}\in\mathcal{V}\,(\text{graph nodes}).
\end{aligned} \tag{24}$$

Here $\mathbf{I}$ is the $|\mathcal{V}|\times|\mathcal{V}|$ identity. To be noted that

(i) v runs over the graph nodes, and no roles attached yet. Output or prompt embeddings are later designated roles on the nodes: $\mathsf{v}_x, \mathsf{v}_p \in \mathcal{V}$ only while forming the contrast $c_{x,\mathcal{K}}(t)$ seen in Eq.(58). Thus, $\mathcal{L}^{(*)}_{\mathcal{T}_t}$ itself is designed to be role-agnostic.

(ii) E^* denotes the hyperedge set used to build each interaction block $(*)$ above, while E still remains consistent as per Eq.(12). $r(e)$ is the number of nodes in the hyperedge e; i.e., $e = \{\mathsf{v}_1,\ldots,\mathsf{v}_{r(e)}\} \subset \mathcal{V}$.

(iii) $\mathcal{D}^{(*)}_{\mathsf{v},\mathcal{T}_t}$ is the node–degree matrix (of size $|\mathcal{V}|\times|\mathcal{V}|$) for block $*$: it is diagonal with entries $\big(\mathcal{D}^{(*)}_{\mathsf{v},\mathcal{T}_t}\big)_{\mathsf{vv}} = \mathfrak{d}^{(*)}_{\mathcal{T}_t}(\mathsf{v})$, the temperature–weighted degree of node v computed from the hyperedge weights in that block.

(iv) $\mathcal{D}^{(*)}_{e,\mathcal{T}_t}$ is the hyperedge–cardinality matrix (of size $|E^{(*)}|\times|E^{(*)}|$) for block $*$: it is diagonal with entries $\big(\mathcal{D}^{(*)}_{e,\mathcal{T}_t}\big)_{ee} = r(e)$.

(v) The node set $\mathcal{V}$ is fixed; $r(e)$ is a property of each hyperedge $e\subset\mathcal{V}$ and is independent of $|\mathcal{V}|$ (and of the number of modalities $|\mathcal{M}|$ unless joint hyperedges is specifically chosen to include one node per modality).

The matrix $W^{(*),\,\text{eff}}_{\mathcal{T}_t} = \mathcal{I}^{(*)} W^{(*)}_{\mathcal{T}_t} (\mathcal{D}^{(*)}_{e,\mathcal{T}_t})^{-1} (\mathcal{I}^{(*)})^{\top}$ is the "effective" pairwise adjacency induced by hyperedges (zero diagonal by convention). The pairwise quantities in Eq. (12) are then obtained by summing blocks:

$$W_{\mathcal{T}_t} \;=\; \sum_{*} \omega_*\, W^{(*),\,\text{eff}}_{\mathcal{T}_t}, \qquad \omega_* \ge 0 \text{ (absorbed by interaction coefficients } \alpha_M, \beta_{MM'}, \gamma_{\mathcal{M}}\,). \tag{25}$$

We pick any two nodes: say, $\mathsf{v}_a, \mathsf{v}_b$ in the hyperedge $e = \{\mathsf{v}_1, .., \mathsf{v}_a, .., \mathsf{v}_b, .., \mathsf{v}_{r(e)}\} \subset \mathcal{V}$ to define a symmetric, nonnegative pairwise dissimilarity $\widehat{d}_{\text{sem}}(\mathsf{v}_a, \mathsf{v}_b)$. This quantity captures the semantic distortion at node level.

For some modality-aware permutation factor η_*, a generic choice of $w_{\mathcal{T}_t}(e)$ is

$$w_{\mathcal{T}_t}(e) \;=\; \mathbf{1}_{\{e \in E^{(*)}\}} \exp\left(-\eta_* \frac{\sum_{1 \le \mathsf{v}_a, \mathsf{v}_b \le r(e)} \widehat{d}_{\text{sem}}(\mathsf{v}_a, \mathsf{v}_b)}{\sum_{1 \le \mathsf{v}_a \le r(e)} \mathcal{T}_t(\mathsf{v}_a)} \right), \tag{26}$$

which is permutation–invariant and temperature–scaled.

$$\Delta_{\varepsilon,h}(x \mid p) := \left[\log \left(\frac{\int_{\mathcal{K}} K_h\big(\Pi_{\mathcal{K}}(x), x_2\big) \left[(1-\varepsilon)\, Z\big(p, \mathcal{T}_t\big)^{-1} e^{-\mathcal{E}(x_2,p)/\mathcal{T}_t} + \varepsilon\, \rho(x_2) \right] d\mu(x_2)}{\int_{\mathcal{K}} \left[(1-\varepsilon)\, Z\big(p, \mathcal{T}_t\big)^{-1} e^{-\mathcal{E}(x_2,p)/\mathcal{T}_t} + \varepsilon\, \rho(x_2) \right] d\mu(x_2)} \right) \tag{27}$$

$$- \log\Big(\int_{\mathcal{X}} K_h(x, x_2) \left[(1-\varepsilon)\, Z\big(p, \mathcal{T}_t\big)^{-1} e^{-\mathcal{E}(x_2,p)/\mathcal{T}_t} + \varepsilon\, \rho(x_2) \right] d\mu(x_2) \Big) \Bigg]_+, \tag{28}$$

A.12 Mercer's theorem

By Mercer's theorem Mercer (1909), if $K_{\mathcal{T}_t}$ is a continuous, symmetric, positive-definite on a compact measure space $(\mathcal{V}, \mu)$, then there exists a unique RKHS $\mathcal{H}$ which is associated with a reproducing kernel $K_{\mathcal{T}_t}$. In the present context of discrete graph, $\mathcal{V}$ is finite which satisfies the criterion. This theorem ensures that there exists a feature map

$$\Phi : \mathcal{V} \to \mathcal{H}, \tag{29}$$

which admits an orthonormal eigen decomposition. We have leveraged it in Eq. (13).

A.13 Graph Maps

This construction is separate from the modality feature maps $\Phi_M(x^{(M)})$ and prompt embeddings $\Psi_M(p)$ that live in modality RKHS $\mathcal{H}_M$ used in the energy landscape as noted in Section 4.3. Here, Υ is defined on the node set, with $\mathsf{v}, \mathfrak{v}$ being the graph nodes, induced by a single graph RKHS $\mathcal{H}_{\text{graph}}$ or just $\mathcal{H}$ for notational simplicity. Therefore, $\Phi_M : \mathcal{X}_M \to \mathcal{H}_M$ and $\Psi_M : \mathcal{P} \to \mathcal{H}_M$ play complementary roles with $\Upsilon : \mathcal{V} \to \mathcal{H}$ in the context of graph theory (i.e., modality & prompt embeddings vs. graph embeddings).

A.14 Why Time-Varying Eigenpairs?

The eigenpairs of the multimodal Laplacian $\mathcal{L}^{\text{multi}}_{\mathcal{T}_t}$, as presented in Eq. (14) are:

- $\Lambda = \text{diag}\Big(\lambda_1(t), \ldots, \lambda_{|\mathcal{V}|}(t)\Big)$ with $\lambda_i(t) \in \mathbb{R}^+$ being the time-varying eigenvalues at node i (that acts like a frequency-dependent penalty or diffusion coefficient),
- $U = \Big[u_1(t), \ldots, u_{|\mathcal{V}|}(t)\Big]$ is the orthonormal eigenvector matrix with $u_i(t) \in \mathbb{R}^{|\mathcal{V}|}$ being the time-varying eigenfunctions.

Note: We assume $G_{\mathcal{T}_t}$ is connected for each fixed t, so that $\lambda_1(t) = 0$ and $\lambda_2(t) > 0$ hold true; when not connected, all occurrences of $u_1(t)$ and $\lambda_2(t)$ below should be read as the orthogonal complement of the full nullspace and the first strictly positive eigenvalue, respectively.

Eigenvalues $\lambda_i(t)$ contract or expand based on evolving inter-node (semantic) affinities, while eigenvectors $u_i(t)$ adjust the directions of these semantic modes. Including $\mathcal{T}_t$ explicitly allows us to

control hallucination sensitivity: as lower temperatures $\mathcal{T}_t \downarrow 0$ emphasize stable low-energy modes, reducing hallucinations leading to more desired outputs and vice versa. In a nutshell, the time variation of $\{(\lambda_i(t), u_i(t))\}$ arises from the temperature schedule $\mathcal{T}_t$, which changes the affinities on the graph edges and hence the spectrum of $\mathcal{L}_{\mathcal{T}_t}^{\text{multi}}$.

B Extended Proofs

In this section, we provide detailed proofs for Theorems 1 and 2.

B.1 Proof of Theorem 1

Proof. **Step 0 (setup and measurability).** By assumption, $\rho > 0$ μ-a.e. with $\int_{\mathcal{X}} \rho \, d\mu = 1$, and $K_h : \mathcal{X} \times \mathcal{X} \to (0, \infty)$ is a μ-Markov kernel with $\int_{\mathcal{X}} K_h(x_1, x_2) \, d\mu(x_2) = 1$ for all $x_1 \in \mathcal{X}$. Define

$$(T_h q)(x_1) \;:=\; \int_{\mathcal{X}} K_h(x_1, x_2)\, q(x_2)\, d\mu(x_2), \qquad q \in L^1(\mu),\ x_1 \in \mathcal{X}. \tag{30}$$

Let the ε–smoothed model be

$$\tilde{f}_{p,\varepsilon}(x_2) \;:=\; (1-\varepsilon)\, f_p(x_2) \;+\; \varepsilon\, \rho(x_2), \qquad \varepsilon \in (0,1), \tag{31}$$

and its $\mathcal{K}$–restricted renormalization be

$$\tilde{f}^{\mathcal{K}}_{p,\varepsilon}(x_2) \;:=\; \frac{\mathbf{1}_{\{x_2 \in \mathcal{K}\}} \tilde{f}_{p,\varepsilon}(x_2)}{\displaystyle\int_{\mathcal{K}} \tilde{f}_{p,\varepsilon}(x_2)\, d\mu(x_2)} \;=\; \frac{\mathbf{1}_{\{x_2 \in \mathcal{K}\}} \tilde{f}_{p,\varepsilon}(x_2)}{\mathsf{Z}_\varepsilon}, \quad \mathsf{Z}_\varepsilon \in (0,1]. \tag{32}$$

Measurability of $\Pi_{\mathcal{K}} : \mathcal{X} \to \mathcal{K}$ (with $\Pi_{\mathcal{K}}(x) = x$ for $x \in \mathcal{K}$) ensures $(T_h \tilde{f}^{\mathcal{K}}_{p,\varepsilon}) \circ \Pi_{\mathcal{K}}$ is measurable; thus Eq. (6) is meaningful pointwise.

Step 1 (strict positivity $\Rightarrow$ finiteness). From Eq. (31) and Eq. (30), for any $x_1 \in \mathcal{X}$,

$$\begin{aligned}(T_h \tilde{f}_{p,\varepsilon})(x_1) &= \int_{\mathcal{X}} K_h(x_1, x_2) \Big((1-\varepsilon) f_p(x_2) + \varepsilon \rho(x_2) \Big)\, d\mu(x_2) \\ &\geq \varepsilon \int_{\mathcal{X}} K_h(x_1, x_2) \rho(x_2)\, d\mu(x_2) \;=\; \varepsilon\, (T_h \rho)(x_1) \;>\; 0,\end{aligned} \tag{33}$$

since $\rho > 0$ μ-a.e. and $K_h > 0$. Similarly, by Eq. (32),

$$(T_h \tilde{f}^{\mathcal{K}}_{p,\varepsilon})(x_1) = \frac{1}{\mathsf{Z}_\varepsilon} \int_{\mathcal{K}} K_h(x_1, x_2)\, \tilde{f}_{p,\varepsilon}(x_2)\, d\mu(x_2) \;\geq\; 0, \tag{34}$$

and $(T_h \tilde{f}^{\mathcal{K}}_{p,\varepsilon})(x_1) > 0$ whenever $\mu\big(\{x_2 \in \mathcal{K} : K_h(x_1, x_2) > 0\}\big) > 0$, which holds for all x_1 if $K_h > 0$ everywhere. Hence, both logarithms in Eq. (6) are finite; $d^{(\varepsilon,h)}_{\text{sem}}$ is well-defined.

Step 2 (g-independence). By inspection of Eq. (6), only $(f_p, \rho, K_h, \Pi_{\mathcal{K}}, \mu)$ appear; the ground-truth g is absent. Thus the statistic is independent of g.

Step 3 (behavior on $\mathcal{K}$). We fix $x \in \mathcal{K}$. Then $\Pi_{\mathcal{K}}(x) = x$, and

$$\frac{(T_h \tilde{f}^{\mathcal{K}}_{p,\varepsilon})(x_1)}{(T_h \tilde{f}_{p,\varepsilon})(x_1)} = \frac{\int_{\mathcal{K}} K_h(x_1, x_2) \tilde{f}_{p,\varepsilon}(x_2)\, d\mu(x_2)}{\mathsf{Z}_\varepsilon \int_{\mathcal{X}} K_h(x_1, x_2) \tilde{f}_{p,\varepsilon}(x_2)\, d\mu(x_2)} \;=\; \frac{\mathsf{A}_x}{\mathsf{Z}_\varepsilon \big(\mathsf{A}_x + \mathsf{B}_x\big)}, \tag{35}$$

where

$$\mathsf{A}_x \;:=\; \int_{\mathcal{K}} K_h(x_1, x_2) \tilde{f}_{p,\varepsilon}(x_2)\, d\mu(x_2), \qquad \mathsf{B}_x \;:=\; \int_{\mathcal{X} \setminus \mathcal{K}} K_h(x_1, x_2) \tilde{f}_{p,\varepsilon}(x_2)\, d\mu(x_2) \;\geq 0. \tag{36}$$

If

$$\mathsf{B}_x \;\geq\; \big(\mathsf{Z}_\varepsilon^{-1} - 1\big)\, \mathsf{A}_x, \tag{37}$$

then the right-hand side of Eq. (35) is ≤ 1, so the inner logarithm in Eq. (6) is ≤ 0 and the $[\cdot]^+$-clipping yields $d^{(\varepsilon,h)}_{\text{sem}}(x; \mathcal{K}, \mathcal{X}) = 0$. Even when Eq. (37) fails, the clipped score never becomes negative, so no spurious negative penalties occur on $\mathcal{K}$.

Step 4 (behavior off $\mathcal{K}$). We fix $x \notin \mathcal{K}$. Then $\Pi_{\mathcal{K}}(x) \in \mathcal{K}$ and

$$\frac{(T_h \tilde{f}^{\mathcal{K}}_{p,\varepsilon})(\Pi_{\mathcal{K}}(x_1))}{(T_h \tilde{f}_{p,\varepsilon})(x_1)} = \frac{\int_{\mathcal{K}} K_h(\Pi_{\mathcal{K}}(x_1), x_2) \tilde{f}_{p,\varepsilon}(x_2)\, d\mu(x_2)}{\mathsf{Z}_\varepsilon \int_{\mathcal{X}} K_h(x_1, x_2) \tilde{f}_{p,\varepsilon}(x_2)\, d\mu(x_2)}. \tag{38}$$

We assume the following localization/consistency condition holds for some $\mathrm{coeff} > 0$:

$$\int_{\mathcal{K}} K_h(\Pi_{\mathcal{K}}(x_1), x_2) \tilde{f}_{p,\varepsilon}(x_2)\, d\mu(x_2) \;\geq\; (1 + \mathrm{coeff})\, \mathsf{Z}_\varepsilon \int_{\mathcal{X}} K_h(x_1, x_2) \tilde{f}_{p,\varepsilon}(x_2)\, d\mu(x_2), \quad \forall x \notin \mathcal{K}. \tag{39}$$

Then the ratio in Eq. (38) exceeds 1, the inner log in Eq. (6) is strictly positive, and thus

$$x \notin \mathcal{K} \text{ and Eq. (39)} \implies d^{(\varepsilon,h)}_{\mathrm{sem}}(x; \mathcal{K}, \mathcal{X}) \;>\; 0. \tag{40}$$

Therefore a strictly positive, finite penalty is assigned to implausible outputs under the mild consistency assumption in Eq. (39).

Step 5 (conclusion for hallucination tracking). From Step 1, Eq. (6) is finite and well-defined; from Step 2 it is reference-free (independent of g). Step 3 shows the score vanishes on $\mathcal{K}$ under Eq. (37) and never assigns negative values there; Step 4 shows it is strictly positive off $\mathcal{K}$ under Eq. (39). Hence Eq. (6) furnishes a pointwise, KL-calibrated signal separating plausible from implausible outputs in the smoothed sense determined by (ε, h, K_h), enabling stable hallucination tracking across prompts and model versions without access to g. □

B.2 Proof of Lemma 1

Proof. Since $\mathcal{H}_M$ is separable, Bochner measurability of Φ_M and Ψ_M is equivalent to strong (Borel) measurability; see, e.g., (Diestel & John J. Uhl, 1977, Ch. II). Thus

$$\Phi_M^{-1}(U) \in \mathcal{F}_{\mathcal{X}_M} \quad \text{and} \quad \Psi_M^{-1}(V) \in \mathcal{F}_{\mathcal{P}} \quad \text{for all open } U, V \subset \mathcal{H}_M. \tag{41}$$

We define the product map

$$\Upsilon :\; \mathcal{X}_M \times \mathcal{P} \to \mathcal{H}_M \times \mathcal{H}_M, \qquad \Upsilon(x, p) := \big(\Phi_M(x), \Psi_M(p)\big). \tag{42}$$

Let $\mathcal{B}(\mathcal{H}_M \times \mathcal{H}_M)$ denote the product Borel σ-algebra. For any open rectangles $U \times V$ with $U, V \subset \mathcal{H}_M$ open,

$$\Upsilon^{-1}(U \times V) = \big\{(x, p) : \Phi_M(x) \in U,\; \Psi_M(p) \in V\big\} = \Phi_M^{-1}(U) \times \Psi_M^{-1}(V) \in \mathcal{F}_{\mathcal{X}_M} \otimes \mathcal{F}_{\mathcal{P}} \tag{43}$$

by Eq. (41). Since the family of open rectangles generates $\mathcal{B}(\mathcal{H}_M \times \mathcal{H}_M)$ and $\mathcal{F}_{\mathcal{X}_M} \otimes \mathcal{F}_{\mathcal{P}}$ is a σ-algebra, a monotone class/π–λ argument implies that

$$\Upsilon \text{ is } \big(\mathcal{F}_{\mathcal{X}_M} \otimes \mathcal{F}_{\mathcal{P}}\big)\text{–}\mathcal{B}(\mathcal{H}_M \times \mathcal{H}_M) \text{ measurable.} \tag{44}$$

Let's consider the inner-product map

$$\mathrm{ip} :\; \mathcal{H}_M \times \mathcal{H}_M \to \mathbb{R}, \qquad \mathrm{ip}(u, v) := \langle u, v \rangle_{\mathcal{H}_M}. \tag{45}$$

Continuity of ip follows from the Cauchy–Schwarz and triangle inequalities: for all $u_1, u_2, v_1, v_2 \in \mathcal{H}_M$,

$$\begin{aligned} \big|\mathrm{ip}(u_1, v_1) - \mathrm{ip}(u_2, v_2)\big| &= \big|\langle u_1 - u_2,\, v_1\rangle + \langle u_2,\, v_1 - v_2\rangle\big| \\ &\leq \|u_1 - u_2\|\, \|v_1\| + \|u_2\|\, \|v_1 - v_2\|, \end{aligned} \tag{46}$$

which shows that ip is continuous and hence Borel measurable with respect to $\mathcal{B}(\mathcal{H}_M \times \mathcal{H}_M)$.

The composition

$$(x, p) \;\longmapsto\; \mathrm{ip}\big(\Upsilon(x, p)\big) \;=\; \big\langle \Phi_M(x),\, \Psi_M(p) \big\rangle_{\mathcal{H}_M} \tag{47}$$

is therefore measurable from $\big(\mathcal{X}_M \times \mathcal{P},\; \mathcal{F}_{\mathcal{X}_M} \otimes \mathcal{F}_{\mathcal{P}}\big)$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ by Eq. (44) and the Borel measurability of ip in Eq. (45)–(46). This yields the claimed joint measurability on $\mathcal{F}_{\mathcal{X}_M} \otimes \mathcal{F}_{\mathcal{P}}$. □

B.3 Proof of Theorem 2

Proof. **Step 1: Well-posedness and nonnegativity of the block quadratic form.** Let $m := |\mathcal{M}| \geq 2$ be fixed. For each $M \in \mathcal{M}$, set

$$v_M(x,p) \;:=\; A_M^{1/2}\, \mathsf{r}_M(x,p) \;\in \mathcal{H}_M, \qquad \mathsf{r}_M(x,p) = \Phi_M(x^{(M)}) - \Psi_M(p). \tag{48}$$

By boundedness and self-adjoint PSD of A_M, $A_M^{1/2}$ is bounded and self-adjoint PSD, and v_M is well-defined. Write the first two terms of Eq. (10) as

$$\sum_M \|v_M\|^2_{\mathcal{H}_M} \;+\; \frac{2}{m-1} \sum_{M<M'} \big\langle v_M, R_{MM'} v_{M'} \big\rangle_{\mathcal{H}_M}. \tag{49}$$

Since $R_{MM'} : \mathcal{H}_{M'} \to \mathcal{H}_M$ is a symmetric contraction with $\|R_{MM'}\| \leq 1$ and $R_{M'M} = R_{MM'}^*$, the Cauchy–Schwarz inequality and the operator norm bound yield

$$\big|\langle v_M, R_{MM'} v_{M'} \rangle\big| \;\leq\; \|R_{MM'}\|\, \|v_M\|\, \|v_{M'}\| \;\leq\; \|v_M\|\, \|v_{M'}\|. \tag{50}$$

Therefore,

$$\begin{aligned} \sum_M \|v_M\|^2 + \frac{2}{m-1} \sum_{M<M'} \langle v_M, R_{MM'} v_{M'} \rangle &\geq \sum_M \|v_M\|^2 - \frac{2}{m-1} \sum_{M<M'} \|v_M\|\, \|v_{M'}\| \\ &= \frac{m}{m-1} \sum_M \|v_M\|^2 \;-\; \frac{1}{m-1} \Big(\sum_M \|v_M\| \Big)^2, \end{aligned} \tag{51}$$

where the identity $\sum_{M<M'} ab = \frac{1}{2}\big[(\sum_M a)^2 - \sum_M a^2\big]$ has been used with $a = \|v_M\|$. By the Cauchy–Schwarz inequality,

$$\Big(\sum_M \|v_M\| \Big)^2 \;\leq\; m \sum_M \|v_M\|^2. \tag{52}$$

Substituting Eq. (52) into Eq. (51) gives

$$\sum_M \|v_M\|^2 + \frac{2}{m-1} \sum_{M<M'} \langle v_M, R_{MM'} v_{M'} \rangle \;\geq\; 0. \tag{53}$$

Hence the block quadratic form in Eq. (49) is nonnegative for all (x,p).

Step 2: Nonnegativity of the joint tensor term. By construction,

$$\mathcal{E}_{\mathcal{M}}(x,p) = \Big\| \bigotimes_{M\in\mathcal{M}} \Phi_M(x^{(M)}) - \bigotimes_{M\in\mathcal{M}} \Psi_M(p) \Big\|^2_{\otimes \mathcal{H}_M} \;\geq\; 0, \tag{54}$$

since it is the square of a norm in the tensor-product RKHS $\otimes_M \mathcal{H}_M$.

Step 3: Measurability. Bochner measurability of Φ_M and Ψ_M into the separable Hilbert space $\mathcal{H}_M$ (refer to Lemma 1) implies that $(x,p) \mapsto \mathsf{r}_M(x,p)$ is $\mathcal{F}_{\mathcal{X}} \otimes \mathcal{F}_{\mathcal{P}}$–measurable for each M, because subtraction is continuous. Since $A_M^{1/2}$ is bounded linear, $(x,p) \mapsto v_M(x,p) = A_M^{1/2} \mathsf{r}_M(x,p)$ is measurable, and so are $(x,p) \mapsto \|v_M(x,p)\|^2$ and $(x,p) \mapsto \langle v_M(x,p), R_{MM'} v_{M'}(x,p) \rangle$; inner products are continuous (hence Borel–measurable), and composition with measurable maps preserves measurability. For the joint tensor term, bilinearity and continuity of the finite tensor product map $(u_M)_M \mapsto \bigotimes_M u_M$ in separable Hilbert spaces imply Bochner measurability of $(x,p) \mapsto \bigotimes_M \Phi_M(x^{(M)})$ and $(x,p) \mapsto \bigotimes_M \Psi_M(p)$; the norm $\|\cdot\|_{\otimes \mathcal{H}_M}$ is continuous, hence $(x,p) \mapsto \mathcal{E}_{\mathcal{M}}(x,p)$ is measurable. Combining these facts shows that $(x,p) \mapsto \mathcal{E}(x,p)$ in Eq. (10) is $\mathcal{F}_{\mathcal{X}} \otimes \mathcal{F}_{\mathcal{P}}$–measurable.

Step 4: Finiteness of the partition function. Since $\mathcal{E}(x,p) \geq 0$ by Steps 1–2, for any $\mathcal{T}_t > 0$,

$$0 \;\leq\; Z(p, \mathcal{T}_t) \;=\; \int_{\mathcal{X}} \exp\big(-\mathcal{E}(x,p)/\mathcal{T}_t\big)\, d\mu(x) \;\leq\; \int_{\mathcal{X}} 1\, d\mu(x). \tag{55}$$

Hence, whenever $\mu(\mathcal{X}) < \infty$, $Z(p, \mathcal{T}_t) \leq \mu(\mathcal{X}) < \infty$. In the case $\mu(\mathcal{X}) = \infty$, a standard integrability condition suffices: assume there exists a measurable, coercive lower bound $\phi : \mathcal{X} \to [0, \infty)$

with $\mathcal{E}(x,p) \geq \phi(x)$ for all x and $\int_{\mathcal{X}} e^{-\phi(x)/\mathcal{T}_t}\, d\mu(x) < \infty$ (e.g., $\phi(x) = c\|x\|^2$ under Lebesgue measure on $\mathbb{R}^d$). Then

$$Z(p,\mathcal{T}_t) \;\leq\; \int_{\mathcal{X}} e^{-\phi(x)/\mathcal{T}_t}\, d\mu(x) \;<\; \infty. \tag{56}$$

Under either case, $Z(p,\mathcal{T}_t)$ is finite, so f_p in Eq. (9) is well-defined.

Step 5: Canonical instances and summary. Equation (10) is a finite sum of measurable, nonnegative terms, hence measurable and nonnegative. The block quadratic part is nonnegative by Eq. (53), and the joint tensor term is nonnegative by Eq. (54). The partition function is finite under Eq. (55) or Eq. (56). Therefore, $\mathcal{E}$ is a valid energy and the Boltzmann density f_p in Eq. (9) is a proper probability density. This completes the proof. □

C Supplementary Results

In this section, we provide further empirical details complementing the main results of ours.

C.1 Derivation of Full Energy Functional

Setup and identities. By Eq. (13), the diffusion kernel is $K_{\mathcal{T}_t} = \exp(-\tau\, \mathcal{L}^{\text{multi}}_{\mathcal{T}_t})$, and $\Upsilon : \mathcal{V} \to \mathcal{H}$ is a feature map with $\langle \Upsilon(\mathsf{v}), \Upsilon(\mathfrak{v})\rangle_{\mathcal{H}} = K_{\mathcal{T}_t}(\mathsf{v},\mathfrak{v})$. Let $\{(\lambda_i(t), u_i(t))\}_{i=1}^{|\mathcal{V}|}$ be the eigenpairs of $\mathcal{L}^{\text{multi}}_{\mathcal{T}_t}$ as in Eq. (14). For any nodes $\mathsf{v},\mathfrak{v} \in \mathcal{V}$ and any graph signal $s \in \mathbb{R}^{|\mathcal{V}|}$, the two standard spectral identities used throughout are:

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

which are exactly the two statements in Eq. (15).

From operator energies to graph-kernel distances. Recall the total energy decomposition from Eq. (10):

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

where $\mathsf{r}_M = \Phi_M(x^{(M)}) - \Psi_M(p)$. By the interconnection note after Eq. (13), fix, for each modality M, two designated nodes $(\mathsf{v}_x^{(M)}, \mathfrak{v}_p) \in \mathcal{V}$ that represent the output and prompt anchors used to evaluate the modality-M discrepancy in the graph-RKHS. The bounded PSD operators A_M define a (possibly weighted) inner product on $\mathcal{H}_M$; absorbing this metric into the graph-kernel geometry (as described in the appendix note referenced there), each $\langle \mathsf{r}_M, A_M \mathsf{r}_M\rangle$ can be written as a nonnegative multiple of the squared distance between the corresponding graph features:

$$\langle \mathsf{r}_M, A_M \mathsf{r}_M\rangle_{\mathcal{H}_M} \;=\; \alpha_M \big\|\Upsilon(\mathsf{v}_x^{(M)};\mathcal{T}_t) - \Upsilon(\mathfrak{v}_p;\mathcal{T}_t)\big\|^2_{\mathcal{H}}, \qquad \alpha_M \in \mathbb{R}_{\geq 0}.$$

Likewise, using the polarization identity and the symmetric contraction structure $B_{MM'} = A_M^{1/2} R_{MM'} A_{M'}^{1/2}$, the cross term is representable as a signed combination of graph-kernel distances between the same anchors; collecting the prefactors into $\beta_{MM'} \in \mathbb{R}_{\geq 0}$ (as in the main text where $\text{coeff}_{\text{cross}_{MM'}} = \beta_{MM'}$), we may write

$$\big\langle A_M^{1/2}\mathsf{r}_M,\; R_{MM'}\, A_{M'}^{1/2}\mathsf{r}_{M'}\big\rangle \;=\; \beta_{MM'}\, \Xi_{MM'}(x,p;\mathcal{T}_t),$$

where $\Xi_{MM'}(\cdot)$ is a bilinear form built from the same pairwise graph-feature differences (its explicit expansion into distance terms follows from polarization and is omitted here for compactness). Finally, the joint term $\mathcal{E}_{\mathcal{M}}(x,p) = \Big\| \bigotimes_{M\in\mathcal{M}} \Phi_M(x^{(M)}) - \bigotimes_{M\in\mathcal{M}} \Psi_M(p)\Big\|^2_{\otimes_M \mathcal{H}_M}$ is nonnegative and measurable; by the same graph-kernel identification used for the intra/cross parts (applied to the joint anchor selection explained in the appendix note you referenced), it too can be expressed as a quadratic form in graph signals supported on $\{\mathsf{v}_x^{(M)}, \mathfrak{v}_p\}_{M\in\mathcal{M}}$ and thus admits the same spectral expansion pattern with a nonnegative coefficient $\gamma_{\mathcal{M}}$.

Modal spectral expansions. Define, for each modality M, the basic signed indicator $s_M(x,p) := \delta_{\mathsf{v}_x^{(M)}} - \delta_{\mathfrak{v}_p} \in \mathbb{R}^{|\mathcal{V}|}$. Then, by the first identity in Eq. (15),

$$\left\|\Upsilon(\mathsf{v}_x^{(M)};\mathcal{T}_t) - \Upsilon(\mathfrak{v}_p;\mathcal{T}_t)\right\|_{\mathcal{H}}^2 = \sum_{i=1}^{|\mathcal{V}|} e^{-\tau\lambda_i(t)} \left|\langle u_i(t), s_M(x,p)\rangle\right|^2.$$

Hence each intra-modal contribution expands as

$$\alpha_M \left\|\Upsilon(\mathsf{v}_x^{(M)};\mathcal{T}_t) - \Upsilon(\mathfrak{v}_p;\mathcal{T}_t)\right\|_{\mathcal{H}}^2 = \sum_{i=1}^{|\mathcal{V}|} \alpha_M\, e^{-\tau\lambda_i(t)} \left|\langle u_i(t), s_M(x,p)\rangle\right|^2,$$

which gives the per-mode terms

$$\mathsf{E}_i^{(\mathrm{intra}_M)}(x,p,t) \;:=\; e^{-\tau\lambda_i(t)} \left|\langle u_i(t), s_M(x,p)\rangle\right|^2 \quad \text{with coefficient } \mathrm{coeff}_{\mathrm{intra}_M} = \alpha_M.$$

For the cross-modal part, set $s_{MM'}(x,p) := s_M(x,p)$ and $s'_{MM'}(x,p) := s_{M'}(x,p)$. Using the polarization identity in the RKHS generated by $K_{\mathcal{T}_t}$ and the same eigenbasis $\{u_i(t)\}$, one obtains a spectral expansion that is bilinear in the modal projections:

$$\Xi_{MM'}(x,p;\mathcal{T}_t) = \sum_{i=1}^{|\mathcal{V}|} e^{-\tau\lambda_i(t)} \langle u_i(t), s_{MM'}(x,p)\rangle \langle u_i(t), s'_{MM'}(x,p)\rangle,$$

so that

$$\frac{2}{|\mathcal{M}|-1} \sum_{M\neq M'} \beta_{MM'}\, \Xi_{MM'}(x,p;\mathcal{T}_t) = \sum_{i=1}^{|\mathcal{V}|} \frac{2}{|\mathcal{M}|-1} \sum_{M\neq M'} \beta_{MM'}\, e^{-\tau\lambda_i(t)} \langle u_i(t), s_M(x,p)\rangle \langle u_i(t), s_{M'}(x,p)\rangle.$$

Thus the per-mode cross-modal contributions are

$$\mathsf{E}_i^{(\mathrm{cross}_{MM'})}(x,p,t) \;:=\; e^{-\tau\lambda_i(t)} \langle u_i(t), s_M(x,p)\rangle \langle u_i(t), s_{M'}(x,p)\rangle \quad \text{with coefficient } \mathrm{coeff}_{\mathrm{cross}_{MM'}} = \beta_{MM'}.$$

For the joint term, denote by $s_{\mathcal{M}}(x,p) \in \mathbb{R}^{|\mathcal{V}|}$ the graph signal associated (as per the appendix link you gave) to the joint interaction in $\mathcal{E}_{\mathcal{M}}(x,p)$. Since this term is a quadratic form in the same graph-kernel geometry, it has the spectral expansion

$$\mathcal{E}_{\mathcal{M}}(x,p) = \gamma_{\mathcal{M}} \sum_{i=1}^{|\mathcal{V}|} e^{-\tau\lambda_i(t)} \left|\langle u_i(t), s_{\mathcal{M}}(x,p)\rangle\right|^2,$$

whence

$$\mathsf{E}_i^{(\mathrm{joint}_{\mathcal{M}})}(x,p,t) \;:=\; e^{-\tau\lambda_i(t)} \left|\langle u_i(t), s_{\mathcal{M}}(x,p)\rangle\right|^2 \quad \text{with coefficient } \mathrm{coeff}_{\mathrm{joint}_{\mathcal{M}}} = \gamma_{\mathcal{M}}.$$

Summing all components. By construction of the multimodal Laplacian as a nonnegative combination of the intra/cross/joint blocks and the definitions of the interaction coefficients in $\mathcal{L}_{\mathcal{T}_t}^{\mathrm{multi}} = \sum_* \mathrm{coeff}_*\, \mathcal{L}_{\mathcal{T}_t}^{(*)}$, the total energy $\mathcal{E}(x,p;\mathcal{T}_t)$ is the sum of the three families above. Collecting the per-mode pieces yields

$$\mathcal{E}(x,p;\mathcal{T}_t) = \sum_* \sum_{i=1}^{|\mathcal{V}|} \mathrm{coeff}_*\, \mathsf{E}_i^{(*)}(x,p,t),$$

where the index $* \in \{\mathrm{intra}_M, \mathrm{cross}_{MM'}, \mathrm{joint}_{\mathcal{M}}\}$, and each $\mathsf{E}_i^{(*)}$ depends only on $\lambda_i(t)$, $u_i(t)$, and the fixed graph signals determined by (x,p) as detailed above. This is the claimed spectral form:

$$\mathcal{E}(x,p;\mathcal{T}_t) = \sum_* \sum_{i=1}^{|\mathcal{V}|} \mathrm{coeff}_*\, \mathsf{E}_i^{(*)}(x,p,t). \tag{57}$$

Now choosing $\pi_{\mathcal{K}} \in \Delta(\mathcal{K})$, where $\Delta(\mathcal{K})$ is the probability simplex on $\mathcal{K}$, satisfies

$$\sum_{\mathsf{v}\in\mathcal{K}} \pi_{\mathcal{K}}(\mathsf{v}) \left(\mathcal{D}_{\mathcal{T}_t}^{\mathrm{multi}}\right)_{\mathsf{vv}} = \left(\mathcal{D}_{\mathcal{T}_t}^{\mathrm{multi}}\right)_{\mathsf{v}_x\mathsf{v}_x}, \quad c_{x,\mathcal{K}}^{\mathrm{raw}}(t) = \mathcal{D}_{\mathcal{T}_t}^{\mathrm{multi}\,1/2}\left(\delta_{\mathsf{v}_x} - \pi_{\mathcal{K}}\right) \in \mathbb{R}^{|\mathcal{V}|}, \tag{58}$$

where $c_{x,\mathcal{K}}^{\mathrm{raw}}(t)$ is the raw contrast vector. Projecting away the leading mode gives $c_{x,\mathcal{K}}(t) = \left(\mathbf{I} - u_1(t)u_1(t)^\top\right) c_{x,\mathcal{K}}^{\mathrm{raw}}(t)$ that ensures $c_{x,\mathcal{K}}(t) \perp u_1(t)$ without assuming a specific null-space structure of the assembled hypergraph.

Why the bounds in Eq. (18) hold, and how to choose $m(t), M(t)$ (non-vacuous). By Eq. (57), the full energy is a nonnegative linear combination of blockwise spectral terms. For the degree–matched contrast $c_{x,\mathcal{K}}(t) \perp u_1(t)$, the energy difference admits the decomposition

$$\mathcal{E}(x,p;\mathcal{T}_t) - \mathcal{E}_{\mathcal{K}}(x,p;\mathcal{T}_t) = \sum_{i=2}^{|\mathcal{V}|} \zeta_i(t,\tau)\, \big|\langle u_i(t), c_{x,\mathcal{K}}(t)\rangle\big|^2, \qquad \zeta_i(t,\tau) = \sum_{*} \theta_*\, \varphi_*^{(i)}(t,\tau), \tag{59}$$

where $* \in \{\mathrm{intra}_M, \mathrm{cross}_{MM'}, \mathrm{joint}_{\mathcal{M}}\}$ indexes the blocks, $\theta_* \in \{\alpha_M, \beta_{MM'}, \gamma_{\mathcal{M}}\}$ are the nonnegative coefficients from Eq. (57), and

$$\varphi_*^{(i)}(t,\tau) := \big\langle u_i(t),\, \mathfrak{D}_*(t,\tau)\, u_i(t)\big\rangle, \qquad \mathfrak{D}_*(t,\tau) \succeq 0,$$

are block response factors evaluated on the same eigenmodes $\{u_i(t)\}_{i\ge 2}$ of $\mathcal{L}_{\mathcal{T}_t}^{\mathrm{multi}}$. For normalized hypergraph constructions (Eq. (24)–(25)) and diffusion-type couplings (Section 4.1), the block responses satisfy the Loewner sandwich

$$e^{-2\tau\,\mathcal{L}_{\mathcal{T}_t}^{\mathrm{multi}}} \preceq \mathfrak{D}_*(t,\tau) \preceq \mathbf{I} \quad\Longrightarrow\quad e^{-2\tau\,\lambda_i(t)} \le \varphi_*^{(i)}(t,\tau) \le 1, \;\; i \ge 2. \tag{60}$$

The left inequality follows from monotonicity of the matrix exponential and the fact that each block smoother is at least as contractive as the global diffusion on $u_1^{\perp}$; the right inequality follows from $\mathfrak{D}_*(t,\tau) \preceq \mathbf{I}$. Plugging Eq. (60) into Eq. (59) yields

$$\sum_{*} \theta_*\, e^{-2\tau\,\lambda_i(t)} \le \zeta_i(t,\tau) \le \sum_{*} \theta_*, \qquad i \ge 2.$$

Refined (spectral) empirical bounds. Define, for each block $*$,

$$\kappa_*^{\max}(t) := \big\|\mathfrak{D}_*(t,0)\big\|_{\mathrm{op}} \le 1, \qquad \kappa_*^{\min}(t) := \lambda_{\min}\big(\mathfrak{D}_*(t,0)\big|_{u_1(t)^{\perp}}\big) \in [0,1], \tag{61}$$

where both quantities are directly estimable from the spectrum of the effective adjacency in Eq. (24)–(25) (restricted to $u_1^{\perp}$). Then, using $e^{-2\tau\mathcal{L}} \preceq \mathfrak{D}_*(t,\tau) \preceq \mathfrak{D}_*(t,0)$ and the Courant–Fischer characterization on $u_1^{\perp}$,

$$\Big(\sum_{*} \theta_*\, \kappa_*^{\min}(t)\Big)\, e^{-2\tau\,\lambda_i(t)} \le \zeta_i(t,\tau) \le \sum_{*} \theta_*\, \kappa_*^{\max}(t), \qquad i \ge 2, \tag{62}$$

so one can take

$$m(t) := \sum_{*} \theta_*\, \kappa_*^{\min}(t), \qquad M(t) := \sum_{*} \theta_*\, \kappa_*^{\max}(t). \tag{63}$$

In practice, $\kappa_*^{\max}(t)$ equals the top eigenvalue of the block response on $u_1^{\perp}$ (often close to 1), while $\kappa_*^{\min}(t)$ equals the blockwise algebraic connectivity surrogate (the smallest nonzero eigenvalue on $u_1^{\perp}$). Estimating (63) from the spectra of $W_{\mathcal{T}_t}^{(*),\mathrm{eff}}$ or the corresponding normalized block Laplacians yields tight, *data-driven* $m(t), M(t)$ for Eq. (18).

Below is the block decomposition of the multimodal Laplacian:

$$\mathcal{L}_{\mathcal{T}_t}^{\mathrm{multi}} = \begin{bmatrix} \mathcal{L}_{\mathrm{intra}}^{(T)} & \mathcal{L}_{\mathrm{cross}}^{(TV)} & \mathcal{L}_{\mathrm{cross}}^{(TA)} \\ \mathcal{L}_{\mathrm{cross}}^{(VT)} & \mathcal{L}_{\mathrm{intra}}^{(V)} & \mathcal{L}_{\mathrm{cross}}^{(VA)} \\ \mathcal{L}_{\mathrm{cross}}^{(AT)} & \mathcal{L}_{\mathrm{cross}}^{(AV)} & \mathcal{L}_{\mathrm{intra}}^{(A)} \end{bmatrix} + \mathcal{L}_{\mathrm{joint}}^{(\mathcal{M})}. \tag{64}$$

The corresponding eigenvalue problem for the i-th mode becomes:

$$\mathcal{L}_{\mathcal{T}_t}^{\mathrm{multi}}\; u_i(t) = \lambda_i(t)\; u_i(t), \tag{65}$$

with eigenvalues $\lambda_i(t)$ encoding the "cost" of semantic diffusion along each mode i.

C.2 Derivations of hallucination bounds and temperature annealing

We derive the operator-tight lower/upper bounds, noted in Eq. (19) in Section 5.3, for $\mathcal{E}_{\mathrm{hall}}^{\mathrm{multi}}(x,p,\cdot)$ using the block-weighted, temperature–modulated Laplacian spectrum in Eq. (14), the spectral

energy form in Eq. (57), and the hallucination component in Eq. (11). By Eq. (14) and the Courant–Fischer principle, the quadratic in Section 5.3 satisfies the two-sided spectral envelope

$$e^{-2\tau\,\lambda_{\max}(t)}\,\|c_{x,\mathcal{K}}(t)\|^2 \;\le\; \mathbb{D}_\tau(x;\mathcal{T}_t) \;\le\; e^{-2\tau\,\lambda_2(t)}\,\|c_{x,\mathcal{K}}(t)\|^2, \tag{66}$$

with $\lambda_{\max}(t) = \lambda_{|\mathcal{V}|}(t)$.

Next, we relate the full energy to $\mathbb{D}_\tau$. Under Theorem 2 and the block assembly in Eqs. (24)–(25), there exist finite scale factors $m(t), M(t) \in (0,\infty)$, determined only by the operator norms of the intra-/cross-/joint blocks (i.e., by $\{A_M\}$, $\{R_{MM'}\}$ with $\|R_{MM'}\| \le 1$, the interaction weights $\alpha_M, \beta_{MM'}, \gamma_{\mathcal{M}}$, and the temperature–modulated hyperedge weights inducing $\mathcal{L}^{\text{multi}}_{\mathcal{T}_t}$), such that

$$m(t)\,\mathbb{D}_\tau(x;\mathcal{T}_t) \;\le\; \mathcal{E}(x,p;\mathcal{T}_t) \;\le\; M(t)\,\mathbb{D}_0(x;\mathcal{T}_t), \qquad \tau \ge 0, \tag{67}$$

where $\mathbb{D}_0$ corresponds to $\tau = 0$. The left inequality follows from bounding each spectral contribution $\mathsf{E}_i^{(*)}(x,p,t)$ below by a nonnegative multiple of $\big|\langle u_i(t), c_{x,\mathcal{K}}(t)\rangle\big|^2$ using the PSD structure of A_M and the contraction bound on $R_{MM'}$, while the right inequality follows from operator-norm upper bounds on the same spectral blocks; full details are supplied in Appendix C.2.

Combining Eqs. (66) and (67) yields the Courant–Fischer sandwich for the full energy:

$$m(t)\,e^{-2\tau\,\lambda_{\max}(t)}\,\|c_{x,\mathcal{K}}(t)\|^2 \;\le\; \mathcal{E}(x,p;\mathcal{T}_t) \;\le\; M(t)\,e^{-2\cdot 0\cdot\lambda_2(t)}\,\|c_{x,\mathcal{K}}(t)\|^2 \;=\; M(t)\,\|c_{x,\mathcal{K}}(t)\|^2. \tag{68}$$

Since the hallucination energy is the positive part of the difference in Eq. (11), we obtain, for $x \notin \mathcal{K}$. When $\mathcal{E}_{\mathcal{K}}(x,p;\mathcal{T}_t)$ is implemented as the same operator restricted to $\mathcal{K}$, the same spectral envelope applies to it, hence the difference inherits a sandwich with the same eigenvalue pair $\{\lambda_2(t), \lambda_{\max}(t)\}$ and scales $\{m(t), M(t)\}$.

A calibrated lower bound of the form advocated by Kalai & Vempala (2024) is matched empirically by choosing a time-indexed temperature profile and interaction scales so that $m(t)\,e^{-2\tau\,\lambda_{\max}(t)} = \Theta(t)$ for a prescribed calibration function $\Theta(t) > 0$; for instance,

$$\mathcal{T}_t \text{ and } \tau(t) \text{ chosen so that} \quad \Theta(t) \;=\; m(t)\,e^{-2\tau(t)\,\lambda_{\max}(t)}, \tag{69}$$

which yields the explicit calibrated bound

$$\mathcal{E}^{\text{multi}}_{\text{hall}}(x,p,\cdot) \;\ge\; \Big(\Theta(t)\,\|c_{x,\mathcal{K}}(t)\|^2 \;-\; \mathcal{E}_{\mathcal{K}}(x,p;\mathcal{T}_t)\Big)_+, \qquad x \notin \mathcal{K}. \tag{70}$$

In particular, for $\mathcal{E}_{\mathcal{K}}$ treated as a fixed baseline (e.g., a distributional or quantile baseline computed on $\mathcal{K}$), Eq. (70) reproduces the calibrated-margin–times–distance structure and can be tuned to overlay the empirical lower bound in calibrated models by setting $\Theta(t)$ to the target slope. The upper envelope in Eq. (19) is simultaneously controlled by $M(t)$ and the spectral gap $\lambda_2(t)$ via Eq. (66), and both $\{\lambda_i(t)\}$ and $\{m(t), M(t)\}$ are tunable through the time-indexed temperature profile $\mathcal{T}_t$ and the block weights inside $W^{(*)}_{\mathcal{T}_t}$ that define $\mathcal{L}^{\text{multi}}_{\mathcal{T}_t}$.

D Experimental Setup

As noted in Section 6.2, below are the essential details about our experiments followed by a full-pager algorithm box.

D.1 Metrics and evaluation

Primary. AUROC/AUPRC for hallucination detection using $d^{(\varepsilon,h)}_{\text{sem}}$ (instance-level, aggregated per dataset/model). **Baselines.** Entropy, max-probability, and margin from $\mathcal{K}$-posteriors. **Secondary.** CF bounds for $\mathcal{E}^{\text{multi}}_{\text{hall}}$ and their *temperature/ε* surfaces; decay with increasing τ (nonincreasing, sandwiched between $e^{-2\tau\lambda_{\max}}$ and $e^{-2\tau\lambda_2}$); Good–Turing–calibrated lower envelope (strictly >0). **Observed.** Our score is *best* across all three datasets: COCO **0.86/0.84**, VQAv2 **0.84/0.81**, AudioCaps **0.80/0.77** (Table 1a). CF planes are tight and monotone with lower $\mathcal{T}_t$ and higher τ, matching theory (Fig. 3); AudioCaps–BLIP is blank by design (as expected!).

D.2 Protocol and design

For each prompt p, we form an admissible set $\mathcal{K}$ of candidate answers (dataset-provided or programmatically generated) and use the selector $\Pi_{\mathcal{K}}$ as `soft_nearest` (nearest-point with convex projection fallback). We sweep a grid over temperature $\mathcal{T}_t$ and smoothing mass ε; *plots show* $Z_{\text{mid}} = \frac{1}{2}(Z_{\text{lo}} + Z_{\text{hi}})$ bounded by per-panel CF lower/upper planes. When plotting, we *aggregate across* diffusion time τ and kernel bandwidth h by the median.

Defaults. $\varepsilon = 0.01$, $h = 0.4$, $\tau = 0.25$, fixed $\mathcal{T}_t$ per run unless stated, logits sharpening $\tau_{\text{logits}} \in [0.01, 0.05]$. Each run logs the full YAML config.

D.3 Inference and compute

Experiments run on `Databricks` (A100) with private checkpoints (gated tokens). Datasets stream from the Hub with synthetic fallback when a split is unavailable. Diffusion kernels use sparse Chebyshev/Lanczos; hypergraphs are CSR; eigen-modes via iterative solvers. *Throughput (ex/s):* `CLIP+Whisper+T5` **420** (fastest), `SigLIP+Whisper+T5` 400, `BLIP+CLIP+Whisper` 360 (Table 1b). Seeds and env versions are pinned in run reports.

Takeaways. $d_{\text{sem}}^{(\varepsilon,h)}$ consistently outperforms entropy/margin baselines (Table 1a). Spectrally, `SigLIP+Whisper+T5` achieves the *lowest median energy* across datasets (COCO **1.92**, VQAv2 **1.99**, AudioCaps **2.08**), while `CLIP+Whisper+T5` is *fastest* (420 ex/s), exposing a clean accuracy–efficiency trade-off (Table 1b).

Algorithm 2: KL-SMOOTHED MULTIMODAL HALLUCINATION — *Extended version of Alg. 1*

Input: Prompt $p \in \mathcal{P}$; sampler for f_p (model generations); admissible set $\mathcal{K}$; base measure μ; kernel K_h (bandwidth h); smoothing mass $\varepsilon \in (0,1)$; baseline density ρ; incidence matrices $\{\mathcal{I}^{(*)}\}$ and block selectors $E^{(*)}$; interaction weights $\{\omega_*\}$; time horizon $t = 0, \dots, T$; temperature profile $\mathcal{T}_t$; diffusion schedule $\tau(t)$.

Output: Node scores $d_{\mathrm{sem}}^{(\varepsilon,h)}(x \,|\, p)$; hyperedge weights $w_{\mathcal{T}_t}(e)$; effective adjacency $W_{\mathcal{T}_t}$; block/multi Laplacians $\{\mathcal{L}_{\mathcal{T}_t}^{(*)}\}$, $\mathcal{L}_{\mathcal{T}_t}^{\mathrm{multi}}$; spectra $\{\lambda_i(t), u_i(t)\}$; contrasts $c_{x,\mathcal{K}}(t)$; hallucination energy bounds for $\mathcal{E}_{\mathrm{hall}}^{\mathrm{multi}}(x, p, \cdot)$.

Phase I: per-prompt semantic score (Eq. (6)).

1. Estimate f_p from model samples (density or histogram on $\mathcal{X}$ under μ).

2. Form $\tilde{f}_{p,\varepsilon}(x) = (1-\varepsilon) f_p(x) + \varepsilon \rho(x)$ and $\tilde{f}_{p,\varepsilon}^{\mathcal{K}}(x_2) = \mathbf{1}_{\{x_2 \in \mathcal{K}\}} \tilde{f}_{p,\varepsilon}(x_2) \big/ \int_{\mathcal{K}} \tilde{f}_{p,\varepsilon} d\mu$.

3. Compute $(T_h \tilde{f}_{p,\varepsilon})(x_1) = \int K_h(x_1, x_2) \tilde{f}_{p,\varepsilon}(x_2)\, d\mu(x_2)$ and $(T_h \tilde{f}_{p,\varepsilon}^{\mathcal{K}})(x_1)$; evaluate $d_{\mathrm{sem}}^{(\varepsilon,h)}(x \,|\, p) = \big[\log(T_h \tilde{f}_{p,\varepsilon}^{\mathcal{K}}(\Pi_{\mathcal{K}}(x))) - \log(T_h \tilde{f}_{p,\varepsilon}(x))\big]_+$.

Phase II: hyperedges, weights, and Laplacian blocks (Eqs. (24)–(25), (26)).

4. For each node $\mathsf{v}_a \sim (x_a, p)$, store $\Delta_a := d_{\mathrm{sem}}^{(\varepsilon,h)}(x_a \,|\, p)$.

5. For each hyperedge $e = \{\mathsf{v}_1, \dots, \mathsf{v}_{r(e)}\} \in E^{(*)}$, set $w_{\mathcal{T}_t}(e) = \mathbf{1}_{\{e \in E^{(*)}\}} \exp\big(-\eta_* \frac{\sum_{a<b} |\Delta_a - \Delta_b|}{\sum_a \mathcal{T}_t(\mathsf{v}_a)}\big)$.

6. Build $W_{\mathcal{T}_t}^{(*)} = \mathrm{diag}\{w_{\mathcal{T}_t}(e)\}$, degrees $\mathcal{D}_{\mathsf{v},\mathcal{T}_t}^{(*)}$ and $\mathcal{D}_{e,\mathcal{T}_t}^{(*)}$, effective adjacency $W_{\mathcal{T}_t}^{(*),\mathrm{eff}} = \mathcal{I}^{(*)} W_{\mathcal{T}_t}^{(*)} (\mathcal{D}_{e,\mathcal{T}_t}^{(*)})^{-1} (\mathcal{I}^{(*)})^\top$.

7. Form block Laplacians $\mathcal{L}_{\mathcal{T}_t}^{(*)} = \mathbf{I} - (\mathcal{D}_{\mathsf{v},\mathcal{T}_t}^{(*)})^{-1/2} W_{\mathcal{T}_t}^{(*),\mathrm{eff}} (\mathcal{D}_{\mathsf{v},\mathcal{T}_t}^{(*)})^{-1/2}$ and aggregate $W_{\mathcal{T}_t} = \sum_* \omega_* W_{\mathcal{T}_t}^{(*),\mathrm{eff}}$; assemble $\mathcal{L}_{\mathcal{T}_t}^{\mathrm{multi}}$ accordingly.

Phase III: spectral objects and contrasts (Eqs. (14), (58)).

8. Compute leading spectrum of $\mathcal{L}_{\mathcal{T}_t}^{\mathrm{multi}}$: $\{\lambda_i(t), u_i(t)\}$ (e.g., LOBPCG/power iteration on sparse matrices). Ensure $\lambda_2(t) > 0$ (connectedness).

9. Build degree–matched $\pi_{\mathcal{K}}$ and raw contrast $c_{x,\mathcal{K}}^{\mathrm{raw}}(t) = \mathcal{D}_{\mathcal{T}_t}^{\mathrm{multi}\,1/2}\big(\delta_{\mathsf{v}_x} - \pi_{\mathcal{K}}\big)$; project $c_{x,\mathcal{K}}(t) = (\mathbf{I} - u_1 u_1^\top) c_{x,\mathcal{K}}^{\mathrm{raw}}(t)$.

Phase IV: energies and guarantees (Eqs. (17) & (19)).

10. Evaluate the diffusion quadratic form $Q_\tau(t) = \langle c_{x,\mathcal{K}}(t),\, e^{-2\tau(t)\mathcal{L}_{\mathcal{T}_t}^{\mathrm{multi}}}\, c_{x,\mathcal{K}}(t) \rangle$ via Krylov–exponential or spectral filter.

11. Choose empirical $m(t), M(t)$ from block coefficients/operator norms (bounds discussion) and report

$$m(t)\, e^{-2\tau(t)\lambda_{\max}(t)} \|c_{x,\mathcal{K}}(t)\|^2 \;\le\; \mathcal{E}(x,p;\mathcal{T}_t) - \mathcal{E}_{\mathcal{K}}(x,p;\mathcal{T}_t) \;\le\; M(t)\, e^{-2\tau(t)\lambda_2(t)} \|c_{x,\mathcal{K}}(t)\|^2.$$

12. Set $\mathcal{E}_{\mathrm{hall}}^{\mathrm{multi}}(x,p,\cdot) = \big(\mathcal{E} - \mathcal{E}_{\mathcal{K}}\big)_+ \mathbf{1}_{\{x \notin \mathcal{K}\}}$ and record bounds from Eq. (19).

Phase V: calibration and decay control (Good–Turing, KV embedding, decay).

13. Compute Good–Turing missing-mass $\widehat{m}_{\mathrm{GT}}(t)$ on $\mathcal{X} \setminus \mathcal{K}$; set $\vartheta_{\mathrm{KV}}(t) = \xi\, \widehat{m}_{\mathrm{GT}}(t)$ with $\xi \in (0,1]$.

14. Update $\tau(t)$ to satisfy $m(t)\, e^{-2\tau(t)\lambda_{\max}(t)} \|c_{x,\mathcal{K}}(t)\|^2 \ge \vartheta_{\mathrm{KV}}(t)$ (Eq. (20)); enforce nondecreasing $\tau(t)$.

15. Monitor decay envelope $m(t) e^{-2\tau(t)\lambda_{\max}(t)} \|c\|^2 \le \mathcal{E}_{\mathrm{hall}}^{\mathrm{multi}} \le M(t) e^{-2\tau(t)\lambda_2(t)} \|c\|^2$ and stop when below a target threshold.

Implementation notes (Colab). Sparse matrices for $\mathcal{I}^{(*)}$, $W_{\mathcal{T}_t}^{(*),\mathrm{eff}}$, and $\mathcal{L}_{\mathcal{T}_t}^{\mathrm{multi}}$; row-normalize K_h; stabilize logs via log-sum-exp; estimate $\lambda_2, \lambda_{\max}$ by LOBPCG/power method; compute $e^{-2\tau\mathcal{L}}$ via expm_multiply or truncated Chebyshev; Good–Turing from frequency table on $\mathcal{X} \setminus \mathcal{K}$.