MEASURING ALIGNMENT COLLAPSE: A TRAINING-FREE FRAMEWORK FOR OBJECT HALLUCINATION IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Object-existence hallucination is a major failure mode of vision-language models, yet it is usually treated as a detection problem: design a scorer and report its AUROC. We argue for a different framing, measurement: what should be recorded from a cross-modal alignment matrix so that grounding failure becomes decidable? We cast object-existence hallucination as *cross-modal alignment collapse*: the patch–token similarity matrix of a frozen CLIP encoder fails to concentrate its evidence on the queried object. Reading that collapse requires two complementary families. The singular-value spectrum is a maximal invariant under patch- and token-orthogonal transformations; it exhausts every position-free reading, but is blind to *which* token columns carry the evidence. The Patch–Object Alignment Score (POAS) supplies the positional information the spectrum cannot. Conversely, POAS is inert to background spectral complexity. Together, and in the idealised model for which the theory is stated, the two families give a block-diagonal response, so the pair identifies both underlying parameters while either family alone leaves one unidentified. We instantiate this as Alignment Collapse Measurement (ACM), a training-free measurement that maps an (image, object query) pair to a deterministic 168-dimensional descriptor (90 spectral, 78 POAS) using a frozen CLIP-ViT-L/14. ACM's coordinates are constructed without any label; a classifier is a separate downstream consumer, trained on a small labeled set and kept apart from the measurement. On POPE, ACM achieves AUROC 0.9591 ± 0.0023 against 0.7823 for the strongest training-free baseline (paired bootstrap $p < 10^{-4}$), and 0.8986 on a discriminative CHAIR construction. Ablations align with the theory: removing the spectral family costs 1.97 AUROC points; removing POAS costs 0.82. This work targets only object-existence hallucination; attribute and relation hallucinations remain open.

1 INTRODUCTION

Vision-language models (VLMs) that answer open-ended questions about images have made rapid progress, driven largely by contrastively pre-trained encoders that align vision and language in a shared space (Radford et al., 2021; Dosovitskiy et al., 2021; Zhai et al., 2023). Yet they still exhibit a fundamental failure: object-existence hallucination. Asked "Is there a fire hydrant in the image?", a VLM answers yes often enough that the failure has a name (Li et al., 2023). The problem matters wherever a fabricated object changes a downstream decision, in clinical reporting, say, or behind the wheel, and it has attracted a steady line of detectors (Gunjal et al., 2024), polling benchmarks (Li et al., 2023; Rohrbach et al., 2018), and wider diagnostic suites (Wang et al., 2023; Guan et al., 2024).

The literature attacks hallucination from two directions. Generation-time methods intervene during decoding—contrastive decoding (Leng et al., 2024), attention retrospection (Huang et al., 2024), and feedback- or description-guided decoding (Zhou et al., 2024; Zhang et al., 2025; Ghosh et al., 2025)—and mitigate the failure without touching the encoder, but they require access to the audited model and its decoding loop. Training-free detectors instead read the alignment of a frozen encoder (Gunjal et al., 2024; Yuksekgonul et al., 2023; Oh & Hwang, 2025; Park & Li, 2025), or inspect the audited model's attention or representations at inference time (Jiang et al., 2025b;a; Zhang et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

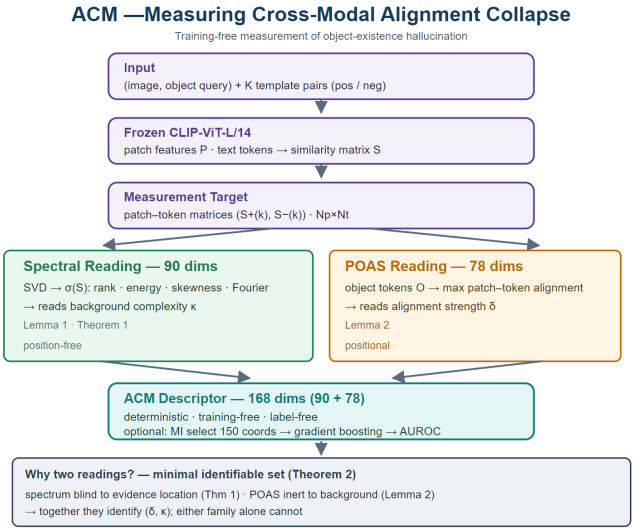

Figure 1: The ACM pipeline. A frozen CLIP-ViT-L/14 encodes each of $K = 5$ paired templates into the patch–token similarity matrix $\mathbf{S}$. The 90-dimensional spectral family reads background complexity and the 78-dimensional POAS family reads object–patch alignment; their concatenation is the 168-dimensional descriptor. The classifier sits downstream.

2026). Nearly every such detector reads the same object: it encodes the image and a prompt naming the queried object with CLIP (Radford et al., 2021), forms the patch–token similarity matrix $\mathbf{S} \in \mathbb{R}^{256\times 77}$, and collapses that matrix to a single number: a global cosine, a positive-minus-negated contrast (Yuksekgonul et al., 2023), a noun-level alignment (Oh & Hwang, 2025), or a weighted global–local fusion (Park & Li, 2025). The collapse is the problem: the structure that survives in $\mathbf{S}$ is exactly what separates hard cases from easy ones.

This paper starts from a different question: what a faithful *measurement* of grounding failure looks like, and how one can tell that a proposed measurement is the right one. A detector is judged by a benchmark number, which can rise for reasons unrelated to the underlying construct; a measurement is judged by whether it responds to its target, is inert to nuisances, and combines without redundancy.

We cast object-existence hallucination as *cross-modal alignment collapse*: the similarity matrix fails to concentrate its evidence on the queried object. Under this reading the measurement target is a matrix, and the question of what to record about it has a sharp answer, which we develop in Section 3.

The answer falls into two parts. The first is the singular-value spectrum. Since $\mathbf{S} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ and orthogonal transformations leave $\mathbf{\Sigma}$ fixed, the spectrum is a maximal invariant under the group that renames patches and tokens, so every position-free reading of the alignment is a function of it. It is nevertheless incomplete: a block of evidence on the object's own token columns and an identical block on function-word columns give matrices with the same singular values and opposite labels, so any spectral reading has power equal to chance on that pair. The second part is a positional reading, the Patch–Object Alignment Score (POAS), which asks where the object noun's evidence lands on the patch grid. Incompleteness runs both ways: POAS is exactly inert to background spectral complexity, and on real data the spectral family separates the POPE classes through its response to template variation (its cross-template dispersion reaches $|d| \approx 0.46$–0.65) rather than through the level of the spectrum alone (Table 1).

Together the two give something stronger than "two features beat one". In the idealised model the response Jacobian of (POAS, effective rank) is block-diagonal and non-singular: POAS responds to alignment strength δ but not to background complexity κ, while effective rank responds to both. The pair *identifies* the parameters $\theta = (\delta, \kappa)$ while either family alone leaves one unidentified (Theorem 2), making spectral and POAS a minimal identifiable measurement set, not just a good feature pair.

We instantiate them as ACM, a training-free measurement mapping one (image, object query) pair to a deterministic 168-dimensional descriptor (90 spectral, 78 POAS) from one frozen CLIP-ViT-L/14 pass per template, without labels. A classifier is an optional downstream consumer; Section 4 reports them separately, and Figure 1 lays out the pipeline.

Contributions.

- A formulation of object-existence hallucination as measurable cross-modal alignment collapse, with the measurement model of Section 3.1 under which claims about measurements become provable, and an optimality result: the spectrum is complete among position-free readings and incomplete in general, POAS is inert to background complexity, and under the proposed idealised measurement model the two form a minimal identifiable set (Section 3.3).
- A controlled synthetic verification that requires neither model weights nor a dataset (Section 4.1), including the predicted block-diagonal response and the chance-level power of spectral readings on the column-mismatch pair.
- ACM, a 168-dimensional training-free measurement, validated on POPE (AUROC 0.9591 ± 0.0023) and a discriminative CHAIR construction (0.8986), plus ablations.

2 Related Work

Object-existence hallucination has been benchmarked, mitigated, and scored, but the scores it has attracted collapse the alignment matrix to a single point; what we ask instead is *which* readings the matrix requires. POPE (Li et al., 2023) turned the failure into a binary polling task over MSCOCO with random, popular and adversarial splits; CHAIR (Rohrbach et al., 2018) scores captions against instance annotations; AMBER (Wang et al., 2023) extends polling to attributes and relations; HallusionBench (Guan et al., 2024) diagnoses the entanglement of language hallucination and visual illusion. Training and generation-time mitigation (Zhou et al., 2024; Huang et al., 2024) target generation itself, while detector-style scores read a frozen encoder's alignment directly (Yuksekgonul et al., 2023; Oh & Hwang, 2025; Park & Li, 2025); we work in the second family but do not frame the output as a detector. Closest in spirit are methods that inspect attention or representations at inference time (Jiang et al., 2025b;a; Zhang et al., 2026). They share our premise that the failure is readable before generation, but they require the audited model's internals; ACM reads only the frozen encoder's outputs. **On the scoring side**, the canonical score is the CLIP cosine (Radford et al., 2021; Dosovitskiy et al., 2021); NegCLIP (Yuksekgonul et al., 2023) contrasts positive and negated prompts, Fine-grained CLIPScore (F-CLIPScore) (Oh & Hwang, 2025) aligns object-noun tokens with patches, GLSim (Park & Li, 2025) fuses global and local similarity, and contrastive decoding (Leng et al., 2024; Zhang et al., 2025; Ghosh et al., 2025) corrects generation-time distributions. Each is a scalar or shallow weighted sum (a single point in the space of readings of $\mathbf{S}$). **The nearest comparator to our positional reading** is F-CLIPScore: both locate the object-noun tokens and extract a per-patch alignment profile, but they choose different aggregations, which encode different measurement assumptions. F-CLIPScore averages the profile (a position-free summary of grounding mass); POAS (Section 3.2) takes its *maximum*, the read that stays responsive when a small, confusable object concentrates its evidence on a few patches—where a mean is diluted by background. Both are instantiated; F-CLIPScore is used as an external baseline (Section 4). **On the measurement side**, we import a sufficient-statistic lens from information theory (Cover & Thomas, 1991): a measurement should respond to the target, stay inert to nuisances, and combine without redundancy. Hallucination resembles out-of-distribution detection (Yang et al., 2024; Liu et al., 2020; Lu et al., 2025; Zhu et al., 2025; Yang et al., 2025) but is cross-modal misalignment, so its natural object of study is the alignment matrix; spectral graph theory (Chung, 1997; Horn & Johnson, 2012) and effective rank (Roy & Vetterli, 2007) supply the tools for treating $\mathbf{S}$ as such an object. Modern contrastive encoders refine CLIP's geometry (Zhai et al., 2023), and the protocol is agnostic to which frozen encoder supplies the matrix; what the hallucination literature lacks is an argument for which readings are necessary, a gap Sections 3.1–3.3 fill.

3 METHOD

We treat object-existence hallucination as a measurable quantity: a deterministic, training-free descriptor of a frozen CLIP's patch–token alignment, constructed with no labels, whose concrete coordinates are defined next and whose necessity is argued for after that.

3.1 THE MEASUREMENT TARGET

Notation. An input image $\mathcal{I}$ and a queried object noun o (from "Is there a *car* in the image?") are encoded by a frozen CLIP-ViT-L/14. The vision tower gives $N_p = 256$ patch features $\mathbf{P} \in \mathbb{R}^{N_p \times d}$ together with a global feature $\mathbf{g} \in \mathbb{R}^d$; the text tower gives $N_t = 77$ token features $\mathbf{T}_k \in \mathbb{R}^{N_t \times d}$ for each of $K = 5$ prompt templates. Tildes denote row-wise ℓ_2 normalisation. For template k the cross-modal similarity matrix is

$$\mathbf{S}^{(k)} = \tilde{\mathbf{P}}\, \tilde{\mathbf{T}}_k^\top / \tau \;\in\; \mathbb{R}^{N_p \times N_t}, \tag{1}$$

with τ the CLIP temperature, adjusted per batch by the geometric mean of the feature standard deviations and clamped to $[\tau/2, 2\tau]$. Templates come in positive/negative pairs ("a photo of a {obj}" / "a photo without a {obj}"), giving $\mathbf{S}_+^{(k)}$ and $\mathbf{S}_-^{(k)}$. The measurement target is the pair $(\mathbf{S}_+^{(k)}, \mathbf{S}_-^{(k)})_{k=1}^K$; the templates are replicates of the same object, and nothing in Sections 3.1–3.3 is affected by K.

Two structural assumptions. Two assumptions turn the target into an object with parameters, so that statements about measurements become statements about those parameters; each is a standard idealisation whose sole role is to make the target non-degenerate.

The first is the expected consequence of contrastive pre-training: that an aligned patch and a token naming an object present in it have, in expectation, higher similarity than an unaligned pair. Concretely, for CLIP's encoders f_v, f_t there exist constants $c_+ > c_- \geq 0$ such that $\mathbb{E}[f_v(p)^\top f_t(w)] \geq c_+$ when the token w names an object present in patch p, and $\leq c_-$ otherwise.

The second turns the alignment matrix into a generative object. For a given sample, the similarity matrix takes the form

$$\mathbf{S} \;=\; c_-\,\mathbf{J} \;+\; \delta\,\mathbf{M} \;+\; \mathbf{\Xi}, \qquad \delta \;=\; c_+ - c_-, \tag{2}$$

where $\mathbf{J}$ is the all-ones matrix, $\mathbf{M} \in \{0,1\}^{N_p \times N_t}$ indicates the patches and token columns that carry the object's evidence, and $\mathbf{\Xi}$ has i.i.d. zero-mean sub-Gaussian entries with variance ν^2. The object is present exactly when $\mathbf{M} \neq \mathbf{0}$.

Eq. 2 defines a measurement model. It specifies a ground truth for $\mathbf{M}$ so that we can ask which functions of $\mathbf{S}$ determine which properties of $\mathbf{M}$, and it provides parameters to perturb one at a time. Section 4.1 evaluates every theoretical claim on synthetic matrices drawn from it, and Section 4 checks the predictions that survive to real CLIP features.

We call a sample *cross-modal-alignment-collapsed* when $\mathbf{M} = \mathbf{0}$: no patch–token block carries evidence for the queried object, and the matrix reduces to a flat baseline plus noise. An affirmative statement about the object under collapse is exactly an object-existence hallucination—the phenomenon the measurement is designed to expose.

Empirical support for the target. Table 1 reports the four spectral summaries over all 9,000 POPE pairs at two levels of aggregation. Read at the level of the alignment (the descriptor averaged over the template prompts) no summary separates the classes on its own: every $|d| \leq 0.16$, so the average spectrum of a grounded pair is, by itself, indistinguishable from a hallucinated one. Read as a *response* (the descriptor's standard deviation across the five positive templates) three of the four summaries do separate the classes with $|d|$ between 0.46 and 0.65: hallucinated pairs show larger cross-template dispersion in effective rank, spectral entropy and dominance ratio. The level alone is therefore only weakly informative; it is read jointly with the response, which is why the descriptor retains both readings and why the multi-template construction of Section 3.2 exists. We read this as evidence that the alignment carries a parameter, the spectral response to template variation, that tracks existence, and any complete measurement must report it alongside the level.

Table 1: Spectral descriptors on POPE (9,000 pairs), split by label. *Level* is the descriptor averaged over the five positive templates; *dispersion* is its standard deviation across those templates. Cohen's d is the standardised mean difference. The marginal level separates the classes weakly ($|d| \leq 0.16$). The cross-template response separates them for three of the four descriptors ($|d| \approx 0.46$–0.65).

Descriptor	Level		d	Dispersion		d
	Grounded	Hallucinated		Grounded	Hallucinated	
$\mathrm{effRank}(\mathbf{S})$	3.41 ± 0.45	3.46 ± 0.46	-0.11	0.286 ± 0.091	0.329 ± 0.092	-0.46
$H(\mathbf{S})$, spectral entropy	1.58 ± 0.10	1.60 ± 0.10	-0.16	0.100 ± 0.016	0.109 ± 0.012	-0.65
$\mathrm{domRatio}(\mathbf{S})$	0.494 ± 0.051	0.490 ± 0.050	$+0.08$	0.026 ± 0.007	0.029 ± 0.007	-0.49
$\Delta\sigma$, spectral gap	0.632 ± 0.088	0.623 ± 0.087	$+0.10$	0.032 ± 0.017	0.032 ± 0.018	-0.03

3.2 The measurement families

The descriptor is a deterministic map from one (image, object query) pair to $\mathbb{R}^{168}$: 90 spectral coordinates and 78 POAS coordinates. Every coordinate is a per-sample function of the alignment matrices of that pair, so measuring a sample is independent of which other samples were collected alongside it. Section S1 lists all 168 coordinates by name and Section S2 gives the pseudocode.

Spectral family: the position-free reading. Write the SVD of the similarity matrix as $\mathbf{S} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ with singular values $\sigma_1 \geq \cdots \geq \sigma_r$, $r = \min(N_p, N_t)$. With $p_i = \sigma_i / \sum_j \sigma_j$, the core descriptors are

$$\mathrm{effRank}(\mathbf{S}) = \frac{(\sum_i \sigma_i)^2}{\sum_i \sigma_i^2}, \qquad H(\mathbf{S}) = -\sum_i p_i \log p_i, \qquad \mathrm{domRatio}(\mathbf{S}) = \frac{\sigma_1}{\sum_i \sigma_i}, \qquad \Delta\sigma = \frac{\sigma_1 - \sigma_2}{\sigma_1}, \tag{3}$$

together with the inverse condition number, the top-5 cumulative energy and a tanh-compressed spectral skewness. Three global agreement terms (cosine, Euclidean and normalised Euclidean distance between image and text global features) join this family because they are computed from the pair as a whole. A Fourier block reads $\mathbf{S}$ as a two-dimensional signal, partitioning its power spectrum by radial frequency. Each of these descriptors is aggregated over the K positive templates and the K negative templates into five per-sample recordings: a positive-template mean, a positive-template standard deviation, a negative-template mean, a negative-template standard deviation, and their difference, giving $10 \times 5 + 8 \times 5 = 90$ coordinates.

POAS family: the positional reading. This family asks a different question of the same matrix: *where* does the object noun's evidence land? We locate the token indices $\mathcal{O}$ covered by the object noun using the tokenizer's offset mapping, falling back to a subsequence search over token ids when offsets are unavailable. Averaging the corresponding token embeddings and renormalising gives an object direction $\mathbf{t}_o$, and the per-patch alignment profile is $\mathbf{s}_o = \tilde{\mathbf{P}}\mathbf{t}_o \in \mathbb{R}^{N_p}$. The Patch–Object Alignment Score is its maximum,

$$\mathrm{POAS}(\mathcal{I}, o) \;=\; \max_p \; \frac{1}{|\mathcal{O}|} \sum_{t \in \mathcal{O}} \mathbf{S}_{p,t}, \tag{4}$$

with the top-k patch concentration $\mathrm{TopK}_k = \frac{1}{k} \sum_{p \in \mathcal{P}_k} \mathbf{s}_o[p]$ as a companion. The block also carries a patch–token pooling profile that does not use $\mathcal{O}$ at all, and an attention-flow profile recording how unevenly alignment strength is spread over patches (Gini, skewness, kurtosis, drive intensity). Each POAS descriptor appears three times, once per template polarity and once as their difference, giving $26 \times 3 = 78$ coordinates. In the released code and in the coordinate list of Section S1 the prefix `poas_` names this object-free pool, whereas the scalar of Eq. 4 is reported as `obj_max_sim`; throughout the paper, POAS always denotes the object-direction maximum.

Fine-grained CLIPScore (Oh & Hwang, 2025) is the closest prior—it too locates the object token and computes patch-level alignment, but pools to a mean over patches, whereas POAS records the *maximum*, the more sensitive read when grounding evidence sits on a few patches. F-CLIPScore is used as an external baseline (Section 4).

Each descriptor is computed under both the positive and the negative template, and the difference is retained as an explicit coordinate. Contextual mixing between the object noun and its function words affects both templates, so the difference is specific to the object. This difference is also the coordinate that distinguishes which columns carry the evidence, as formalised in Section 3.3.

The instrument is parameter-free and label-free, and it queries only a frozen encoder shared across CLIP-backed models; once the similarity matrices of a sample are available, computing all 168 coordinates costs roughly two orders of magnitude less than a single CLIP-ViT-L/14 forward pass.

3.3 Which readings of the alignment matrix are necessary?

Section 3.2 proposed two families. We argue that two is the right number: the spectrum is complete among position-free readings but provably incomplete in general; POAS is inert to a nuisance parameter; and the pair identifies the two parameters of the model while either half does not. Each claim is stated formally below and verified on synthetic matrices in Section 4.1; detailed proof steps are given in the appendix (Section S6).

The spectrum is complete among position-free readings. Let $G = O(N_p) \times O(N_t)$ act on $\mathbb{R}^{N_p \times N_t}$ by $(\mathbf{U}_0, \mathbf{V}_0) \cdot \mathbf{S} = \mathbf{U}_0 \mathbf{S} \mathbf{V}_0^\top$. A reading f is *position free* if it is invariant under this action, i.e. if it does not depend on how patches and tokens are indexed.

A reading f is position free if and only if $f(\mathbf{S}) = \varphi(\sigma(\mathbf{S}))$ for some function φ, where $\sigma(\mathbf{S})$ is the vector of singular values: the polar decomposition $\mathbf{S} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ composes with the acting group to absorb $\mathbf{U}$ and $\mathbf{V}$ into the orthogonal factors, leaving $\mathbf{\Sigma}$, and every function of the singular values is itself invariant (Horn & Johnson, 2012, §7.3). The spectral family of Section 3.2 is therefore *complete* within this natural but restricted class.

The spectrum is incomplete in general. Completeness within the position-free class is a restricted result, and the limitation is structural rather than numerical.

Theorem 1 (Spectral blindness to column mismatch)**.** *Under the measurement model of Section 3.1, there exist samples with identical singular values and opposite ground-truth labels, so any reading that depends on* $\mathbf{S}$ *only through* $\sigma(\mathbf{S})$ *has power equal to its test level.*

Column relabelling is orthogonal, so the elevated similarity can sit on the object's own token columns or on columns that do not name it, with the spectrum unchanged; the proof is in Section S6.

Column mismatch is relevant to POPE's adversarial split, where contextual mixing between an object noun and its function words is common. Location information is lost under position-free readings, so POAS retains the positive-minus-negative difference as an explicit coordinate.

POAS is inert to background complexity. This property is complementary to the spectral blindness of Theorem 1, and the two families together identify both parameters of the measurement model. Under the measurement model of Section 3.1, POAS decomposes exactly as

$$\mathrm{POAS}(\mathbf{S}) = c_- + \max_p \left[\delta \bar{M}_p + \bar{\Xi}_p\right], \qquad \bar{M}_p = \tfrac{1}{|\mathcal{O}|} \sum_{t \in \mathcal{O}} M_{p,t},$$

POAS therefore responds to alignment strength and is invariant to background complexity: it reads only the object-token columns, so an off-column perturbation $\mathbf{E}$ does not affect it, and once δ exceeds the noise maximum the row maximum has slope $\partial \mathrm{POAS}/\partial\delta = 1$. Table 1 shows the spectral family separating the classes on POPE through its cross-template response—a signal we would discard by reading level alone.

Minimal identifiability of the pair. Write the measurement model with two parameters $\theta = (\delta, \kappa)$: alignment strength of the object block, and the number of background blocks contributing comparable singular values. The two families resolve these independently, and the combination is minimal.

Theorem 2 (Local identifiability)**.** *Let* $\mathbf{m}(\mathbf{S}) = (\mathrm{POAS}(\mathbf{S}), \mathrm{effRank}(\mathbf{S}))$ *and* $\mathbf{J}_\theta = \partial \mathbf{m}/\partial\theta$*. Under the measurement model of Section 3.1, at any* θ *with* $\delta > \max_p \bar{\Xi}_p$ *and* $\kappa \geq 1$*,*

$$\mathbf{J}_\theta = \begin{pmatrix} 1 & 0 \\ \partial_\delta \mathrm{effRank} & \partial_\kappa \mathrm{effRank} \end{pmatrix}, \qquad \partial_\kappa \mathrm{effRank} > 0,$$

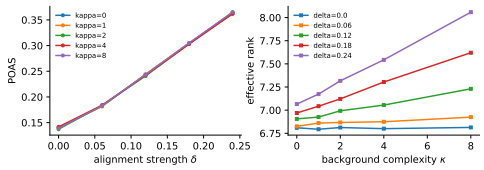

Figure 2: Estimated response of POAS (left) and effective rank (right) over the (δ, κ) grid. POAS tracks δ only; effective rank tracks both, matching the block-diagonal Jacobian predicted by Theorem 2.

so $\det \mathbf{J}_\theta = \partial_\kappa \mathrm{effRank} \neq 0$ *and* $\theta \mapsto \mathbf{m}$ *is locally invertible. Either family alone provides a single equation for two parameters, leaving one parameter unidentified.*

The block-diagonal form is exactly what the maximal-invariant argument and the POAS decomposition predict: POAS feeds the $(\delta, \cdot)$ corner only, while effective rank responds to both. The identification is *local*: $\det \mathbf{J}_\theta \neq 0$ gives neighbourhood invertibility under the measurement model of Section 3.1. The result is a local identifiability statement.

Among readings of $\mathbf{S}$, the pair (spectral, POAS) is therefore a minimal identifiable measurement set for (δ, κ): it identifies both parameters, and dropping either family leaves one unidentified.

Identifiability is established within the measurement model of Section 3.1. Real token embeddings exhibit dependence, so $\partial_\kappa \mathrm{POAS}$ is small but non-zero (Section S5); the empirical consequence is that the spectral response to template variation separates the POPE classes with $|d|$ between 0.46 and 0.65 (Table 1). Identifiability concerns the measurement; downstream AUROC is evaluated separately in Section 4.3.

4 Experiments

Section 3.3 constrains what the measurement should do; this section examines whether it does. Section 4.1 verifies the claims on matrices with known ground truth, where no classifier is needed. On real data we keep the instrument and its consumer separate: ACM yields $\mathbb{R}^{168}$ per sample with no labels, and every benchmark number below comes from attaching the same small classifier to the coordinates under test; a gradient-boosted head fed by a mutual-information selection of 150 coordinates.

4.1 Controlled synthetic verification

These claims can be checked on synthetic matrices with known ground truth, without model weights. We draw $\mathbf{S}$ from Eq. 2 with $N_p = 256$, $N_t = 77$, $c_- = 0.10$, $\nu = 0.02$, an object block of 12 patches by 2 object-token columns, and κ background blocks of the same footprint on disjoint rows and columns.

Table 2 reproduces each prediction and Figure 2 shows the response surfaces: invariance holds to floating-point precision while POAS does not; the mismatch pair puts effective rank at chance (0.514) while POAS separates perfectly; and the estimated Jacobian matches the block-diagonal form of Theorem 2. The first slope differs from 1 by at most the noise maximum and approaches 1 as $\nu \to 0$.

Protocol. AUROC is the primary metric; we also report AP, F1 at the F1-optimal threshold, and accuracy. On CHAIR we report the false-positive rate on grounded objects. All results are 5-fold stratified cross-validation at seed 42, mean $\pm$ sample standard deviation over folds; differences use a paired bootstrap with $B = 10{,}000$ resamples over pooled out-of-fold predictions, two-sided. Section S7 lists every hyperparameter.

Table 2: Controlled checks on synthetic matrices, 400 replicates per condition. (a) Orthogonal invariance: the spectrum is unchanged to machine precision while POAS moves, as the maximal-invariant argument predicts. (b) Column mismatch: effective rank sits at chance, POAS separates perfectly. (c) Response Jacobian: the estimated block matches the block-diagonal form predicted by Theorem 2.

Claim	Quantity	Value
Orthogonal invariance	$\max_i \lvert\sigma_i(\mathbf{S}) - \sigma_i(\mathbf{U}_0\mathbf{S}\mathbf{V}_0^\top)\rvert$	5.3×10^{-15}
	POAS before $\rightarrow$ after rotation	$0.298 \rightarrow 0.080$
Column mismatch	AUROC, effective rank alone	0.514
	AUROC, POAS alone	1.000
Response Jacobian	$\partial\mathrm{POAS}/\partial\delta$ (predicted 1)	0.949
	$\partial\mathrm{POAS}/\partial\kappa$ (predicted 0)	-3×10^{-5}
	$\partial\mathrm{effRank}/\partial\delta$	1.092
	$\partial\mathrm{effRank}/\partial\kappa$ (predicted > 0)	0.041
	$\det \mathbf{J}_\theta$ (scaled)	1.61
	condition number of $\mathbf{J}_\theta$	94

Table 3: POPE, 9,000 pairs, 5-fold cross-validation, backbone matched (Section 4). The 95% bootstrap interval for ACM is $[0.9553, 0.9630]$; paired bootstrap against every baseline gives $p < 10^{-4}$. A linear consumer on the same 168 coordinates is reported in the last row.

Reading of $\mathbf{S}$	AUROC	AP	F1	Accuracy
CLIP-ZeroShot (Radford et al., 2021)	0.7823 ± 0.0035	0.7516	0.7464	0.6972
NegCLIP (Yuksekgonul et al., 2023)	0.6512 ± 0.0042	0.6165	0.6898	0.5844
F-CLIPScore (Oh & Hwang, 2025)	0.7781 ± 0.0076	0.7453	0.7472	0.6919
GLSim (Park & Li, 2025)	0.7685 ± 0.0054	0.7352	0.7430	0.6803
ACM (spectral + POAS, 168-d)	$\mathbf{0.9591 \pm 0.0023}$	**0.9627**	**0.9046**	**0.9050**
ACM (168-d, linear consumer)	0.8648 ± 0.0117	–	–	–

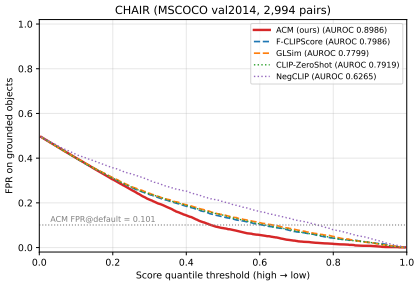

Figure 3: False-positive rate on grounded objects versus threshold on CHAIR. ACM (orange) has a lower false-positive rate than the baselines across the operating range.

4.2 Discrimination on POPE

Table 3 compares the measurement against the four scalar readings it generalises; the gap over the strongest is 17.7 AUROC points, and a linear consumer on the same 168 coordinates still reaches 0.8648 (last row). NegCLIP reaches AUROC 0.6512. Its statistic—the positive-minus-negated-logit difference—is sensitive to the magnitude equalisation induced by contrastive pre-training.

Random-feature and shuffled-label controls yield AUROC 0.5109 and 0.5248 ± 0.0158, respectively.

Residual errors show a structure: adversarial pairs (confusable, genuinely co-occurring objects) are the hardest: per-split AUROC 0.9315 versus 0.9703 and 0.9756 on the random and popular splits (Section S3), so mistakes concentrate where even correct evidence localisation leaves only a near-margin difference.

Table 4: Leave-one-family-out and leave-one-block-out on POPE; "retained" is the dimensionality left after removal. Lower rows vary the number of templates K. AUROC $\pm$ sample standard deviation over folds; Δ is the change from the full measurement.

Configuration	Retained	AUROC	Δ
Full ACM ($K = 5$)	168	0.9591 ± 0.0023	—
without POAS family	90	0.9509 ± 0.0015	-0.82
without spectral family	78	0.9394 ± 0.0039	-1.97
without spectral core & Fourier	93	0.9521 ± 0.0041	-0.70
without spectral core only	133	0.9559 ± 0.0033	-0.32
without Fourier only	128	0.9573 ± 0.0029	-0.18
without CLIP-similarity block	153	0.9547 ± 0.0022	-0.44
without object-token block	144	0.9558 ± 0.0025	-0.33
Single template ($K = 1$)	168	0.9555	-0.36
$K = 3$	168	0.9593	$+0.02$
$K = 7$	168	0.9607	$+0.16$
Random 168 features	168	0.5109	-44.82
Shuffled labels	168	0.5248 ± 0.0158	-43.43

4.3 Incremental validity of each family

Table 4 reports leave-one-family-out and leave-one-block-out results. Removing the spectral family changes AUROC by -1.97 points; removing POAS changes it by -0.82 points. The spectral family tracks the label through its cross-template response rather than its level, and multi-template aggregation is within 0.02 AUROC points of the $K = 5$ setting already at $K = 3$, while $K = 7$ adds a further $+0.16$ points; we keep $K = 5$ as the default.

4.4 Transfer to another image source

On CHAIR (MSCOCO val2014, 2,994 pairs over 500 images) we build the same discriminative existence task with the same pairwise AUROC used on POPE. This protocol is a discriminative existence task over CHAIR annotations, rather than the generative caption-scoring protocol of Rohrbach et al. (2018); the AUROC below is therefore specific to this task. Within this task the measurement reaches AUROC 0.8986 ± 0.0087 against 0.7986 for the strongest baseline (F-CLIPScore), a false-positive rate of 0.101 on grounded objects, and the ordering holds at every threshold (Figure 3). It is monotone in the object's relative area among grounded objects (Spearman -0.40, $p < 10^{-57}$; Section S3), indicating that it responds to grounding strength.

5 Conclusion

Because object-existence hallucination is a property of a cross-modal alignment matrix, what to record about it has a sharp answer: the spectrum exhausts every position-free reading and is blind to column mismatch; POAS reports alignment strength and is inert to background complexity; the pair identifies both parameters, while either half leaves one unidentified. Spectral plus POAS is therefore a minimal measurement set, and Section 4.1 reproduces each prediction.

Two properties bound the scope of the measurement. Identifiability is local, within the idealised model only: κ has no clean discrete counterpart in real CLIP's continuous SVD spectrum, so its role transfers qualitatively, not as a recoverable count. The coordinates are training- and label-free, and the proofs idealise CLIP's token geometry. The consumer output is a discriminant score rather than a calibrated probability estimate. Calibration is a downstream decision, separate from the measurement itself. Attributes, relations, out-of-vocabulary objects, non-English prompts and out-of-distribution imagery lie beyond; the main text uses a single backbone, with cross-backbone transfer verified in the supplement.

AI USE STATEMENT

Large language models were used only for language editing and stylistic refinement. Scientific content, experimental design, data analysis, and results are the authors' own.

ETHICS STATEMENT

This work studies the measurement of a failure mode in publicly available vision-language models. It involves no human subjects, no personally identifiable information, and no new data collection: all datasets used (POPE, MSCOCO, and CHAIR annotations) are publicly released under their respective licences. The measurement operates on externally observable encoder outputs of a frozen model and does not access model internals, training data, or user content. The output is a measurement rather than a definitive judgement of model truthfulness.

REPRODUCIBILITY STATEMENT

The measurement is deterministic and training-free, and all experimental conditions are fixed and reported. The evaluation protocol is described in Section 4; the appendix gives the measurement pseudocode; Section S7 records the hyperparameter settings; and the released code provides the full coordinate list. Anonymous source code is provided as supplementary material and will be made publicly available in a GitHub repository upon acceptance.

REFERENCES

Fan RK Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.

Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. Wiley, 1991.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Utkarsh Tyagi, Oriol Nieto, Zeyu Jin, and Dinesh Manocha. Visual description grounding reduces hallucinations and boosts reasoning in LVLMs. In *International Conference on Learning Representations*, 2025.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024.

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *AAAI Conference on Artificial Intelligence*, 2024.

Roger A Horn and Charles R Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2012.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024.

Nick Jiang, Anish Kachinthaya, Suzie Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations. In *International Conference on Learning Representations*, 2025a.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Conference on Computer Vision and Pattern Recognition*, pp. 25004–25014, 2025b.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Empirical Methods in Natural Language Processing*, 2023.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, 2020.

Shuo Lu, Yingsheng Wang, Lijun Sheng, Lingxiao He, Aihua Zheng, and Jian Liang. Out-of-distribution detection: A task-oriented survey of recent advances. *ACM Computing Surveys*, 58 (2):1–39, 2025. doi: 10.1145/3760390.

Hongseok Oh and Wonseok Hwang. Do vision encoders truly explain object hallucination?: Mitigating object hallucination via simple fine-grained CLIPScore. *Transactions on Machine Learning Research*, 2025.

Seongheon Park and Yixuan Li. GLSim: Detecting object hallucinations in LVLMs via global-local similarity. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Empirical Methods in Natural Language Processing*, 2018.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *European Signal Processing Conference (EUSIPCO)*, pp. 606–610, 2007.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. AMBER: An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

Hermann Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479, 1912.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132:5635–5662, 2024.

Yifeng Yang, Lin Zhu, Zewen Sun, Hengyu Liu, Qinying Gu, and Nanyang Ye. OODD: Test-time out-of-distribution detection with dynamic dictionary. In *Conference on Computer Vision and Pattern Recognition*, pp. 30630–30639, 2025.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *International Conference on Computer Vision*, 2023.

Ce Zhang, Zifu Wan, Zhehan Kan, Martin Q Ma, Simon Stepputtis, Deva Ramanan, Russ Salakhutdinov, Louis-Philippe Morency, Katia Sycara, and Yaqi Xie. Self-correcting decoding with generative feedback for mitigating hallucinations in large vision-language models. In *International Conference on Learning Representations*, 2025.

Feiran Zhang, Yixin Wu, Zhenghua Wang, Xiaohua Wang, Changze Lv, Xuanjing Huang, and Xiaoqing Zheng. VIB-Probe: Detecting and mitigating hallucinations in vision-language models via variational information bottleneck. In *Annual Meeting of the Association for Computational Linguistics*, pp. 23509–23521, 2026.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *International Conference on Learning Representations*, 2024.

Lin Zhu, Yifeng Yang, Xinbing Wang, Qinying Gu, and Nanyang Ye. δEnergy: Optimizing energy change during vision-language alignment improves both OOD detection and OOD generalization. In *Advances in Neural Information Processing Systems*, 2025.

SUPPLEMENTARY MATERIAL

CONTENTS OF THE SUPPLEMENTARY MATERIAL

S1 Full descriptor enumeration **14**

S2 Measurement pseudocode **17**

S3 Additional experimental results **17**

S3.1 CHAIR results and threshold behaviour 17

S3.2 Calibration of the consumer output 18

S3.3 Discriminative evidence for the κ sensor 18

S3.4 Monotonicity in a continuous proxy of alignment strength 18

S3.5 Residual errors and cross-backbone transfer 19

S4 Spectral motivating analysis **19**

S5 Synthetic controlled study **20**

S6 Formal statements and proofs **20**

S7 Hyperparameters **22**

S1 FULL DESCRIPTOR ENUMERATION

The measurement is a deterministic map from one (image, object-query) pair to $\mathbb{R}^{168}$, assembled from two families as in Section 4 of the main paper. Below every coordinate is listed by index, name and family. Spectral coordinates appear with five aggregations each (positive-template mean, positive-template standard deviation, negative-template mean, negative-template standard deviation, and their difference); Fourier coordinates appear with five aggregations (positive mean, negative mean, difference, positive std, negative std); POAS-family coordinates appear with three aggregations (positive, negative, difference). The names are produced by the released code (`acm/measurement.py`) and match the descriptor order exactly.

Table S1: All 168 coordinates of the ACM descriptor, listed in the exact order produced by the released code (`acm/measurement.py`). Every spectral and Fourier coordinate carries five aggregations; every POAS coordinate carries three.

Index	Coordinate name	Family
1	pos_effective_rank_normalized_mean	spectral
2	pos_effective_rank_normalized_std	spectral
3	neg_effective_rank_normalized_mean	spectral
4	neg_effective_rank_normalized_std	spectral
5	diff_effective_rank_normalized	spectral
6	pos_spectral_entropy_normalized_mean	spectral
7	pos_spectral_entropy_normalized_std	spectral
8	neg_spectral_entropy_normalized_mean	spectral
9	neg_spectral_entropy_normalized_std	spectral
10	diff_spectral_entropy_normalized	spectral
11	pos_dominance_ratio_mean	spectral
12	pos_dominance_ratio_std	spectral
13	neg_dominance_ratio_mean	spectral
14	neg_dominance_ratio_std	spectral
15	diff_dominance_ratio	spectral
16	pos_spectral_gap_normalized_mean	spectral
17	pos_spectral_gap_normalized_std	spectral
18	neg_spectral_gap_normalized_mean	spectral
19	neg_spectral_gap_normalized_std	spectral
20	diff_spectral_gap_normalized	spectral
21	pos_condition_number_inv_mean	spectral
22	pos_condition_number_inv_std	spectral
23	neg_condition_number_inv_mean	spectral
24	neg_condition_number_inv_std	spectral
25	diff_condition_number_inv	spectral
26	pos_cumulative_energy_top5_mean	spectral
27	pos_cumulative_energy_top5_std	spectral
28	neg_cumulative_energy_top5_mean	spectral
29	neg_cumulative_energy_top5_std	spectral
30	diff_cumulative_energy_top5	spectral
31	pos_spectral_skewness_mean	spectral
32	pos_spectral_skewness_std	spectral
33	neg_spectral_skewness_mean	spectral
34	neg_spectral_skewness_std	spectral
35	diff_spectral_skewness	spectral
36	pos_cosine_similarity_mean	spectral
37	pos_cosine_similarity_std	spectral
38	neg_cosine_similarity_mean	spectral
39	neg_cosine_similarity_std	spectral
40	diff_cosine_similarity	spectral
41	pos_euclidean_distance_mean	spectral

Table S1: (continued) All 168 coordinates of the ACM descriptor.

Index	Coordinate name	Family
42	pos_euclidean_distance_std	spectral
43	neg_euclidean_distance_mean	spectral
44	neg_euclidean_distance_std	spectral
45	diff_euclidean_distance	spectral
46	pos_normalized_distance_mean	spectral
47	pos_normalized_distance_std	spectral
48	neg_normalized_distance_mean	spectral
49	neg_normalized_distance_std	spectral
50	diff_normalized_distance	spectral
51	pos_fourier_low_ratio	spectral
52	neg_fourier_low_ratio	spectral
53	diff_fourier_low_ratio	spectral
54	pos_fourier_low_ratio_std	spectral
55	neg_fourier_low_ratio_std	spectral
56	pos_fourier_high_ratio	spectral
57	neg_fourier_high_ratio	spectral
58	diff_fourier_high_ratio	spectral
59	pos_fourier_high_ratio_std	spectral
60	neg_fourier_high_ratio_std	spectral
61	pos_fourier_lh_ratio	spectral
62	neg_fourier_lh_ratio	spectral
63	diff_fourier_lh_ratio	spectral
64	pos_fourier_lh_ratio_std	spectral
65	neg_fourier_lh_ratio_std	spectral
66	pos_fourier_entropy	spectral
67	neg_fourier_entropy	spectral
68	diff_fourier_entropy	spectral
69	pos_fourier_entropy_std	spectral
70	neg_fourier_entropy_std	spectral
71	pos_fourier_peak_energy	spectral
72	neg_fourier_peak_energy	spectral
73	diff_fourier_peak_energy	spectral
74	pos_fourier_peak_energy_std	spectral
75	neg_fourier_peak_energy_std	spectral
76	pos_fourier_cumulative_25	spectral
77	neg_fourier_cumulative_25	spectral
78	diff_fourier_cumulative_25	spectral
79	pos_fourier_cumulative_25_std	spectral
80	neg_fourier_cumulative_25_std	spectral
81	pos_fourier_cumulative_50	spectral
82	neg_fourier_cumulative_50	spectral
83	diff_fourier_cumulative_50	spectral
84	pos_fourier_cumulative_50_std	spectral
85	neg_fourier_cumulative_50_std	spectral
86	pos_fourier_cumulative_75	spectral
87	neg_fourier_cumulative_75	spectral
88	diff_fourier_cumulative_75	spectral
89	pos_fourier_cumulative_75_std	spectral
90	neg_fourier_cumulative_75_std	spectral
91	pos_poas_max	POAS
92	neg_poas_max	POAS
93	diff_poas_max	POAS
94	pos_poas_mean	POAS
95	neg_poas_mean	POAS
96	diff_poas_mean	POAS

Table S1: (continued) All 168 coordinates of the ACM descriptor.

Index	Coordinate name	Family
97	pos_poas_std	POAS
98	neg_poas_std	POAS
99	diff_poas_std	POAS
100	pos_poas_top5_mean	POAS
101	neg_poas_top5_mean	POAS
102	diff_poas_top5_mean	POAS
103	pos_poas_top10_mean	POAS
104	neg_poas_top10_mean	POAS
105	diff_poas_top10_mean	POAS
106	pos_poas_saliency	POAS
107	neg_poas_saliency	POAS
108	diff_poas_saliency	POAS
109	pos_poas_entropy_mean	POAS
110	neg_poas_entropy_mean	POAS
111	diff_poas_entropy_mean	POAS
112	pos_poas_entropy_std	POAS
113	neg_poas_entropy_std	POAS
114	diff_poas_entropy_std	POAS
115	pos_poas_entropy_min	POAS
116	neg_poas_entropy_min	POAS
117	diff_poas_entropy_min	POAS
118	pos_obj_max_sim	POAS
119	neg_obj_max_sim	POAS
120	diff_obj_max_sim	POAS
121	pos_obj_mean_sim	POAS
122	neg_obj_mean_sim	POAS
123	diff_obj_mean_sim	POAS
124	pos_obj_std_sim	POAS
125	neg_obj_std_sim	POAS
126	diff_obj_std_sim	POAS
127	pos_obj_top1_sim	POAS
128	neg_obj_top1_sim	POAS
129	diff_obj_top1_sim	POAS
130	pos_obj_top5_mean	POAS
131	neg_obj_top5_mean	POAS
132	diff_obj_top5_mean	POAS
133	pos_obj_top10_mean	POAS
134	neg_obj_top10_mean	POAS
135	diff_obj_top10_mean	POAS
136	pos_obj_contrast	POAS
137	neg_obj_contrast	POAS
138	diff_obj_contrast	POAS
139	pos_obj_above_thresh_ratio	POAS
140	neg_obj_above_thresh_ratio	POAS
141	diff_obj_above_thresh_ratio	POAS
142	pos_attn_flow_max	POAS
143	neg_attn_flow_max	POAS
144	diff_attn_flow_max	POAS
145	pos_attn_flow_mean	POAS
146	neg_attn_flow_mean	POAS
147	diff_attn_flow_mean	POAS
148	pos_attn_flow_std	POAS
149	neg_attn_flow_std	POAS
150	diff_attn_flow_std	POAS
151	pos_attn_flow_high_ratio	POAS

Algorithm 1 The ACM measurement for one (image, object-query) pair.

Require: image $\mathcal{I}$, object noun o, frozen CLIP-ViT-L/14, K template pairs

build positive/negative prompt templates $T_+^{(k)}, T_-^{(k)}$ for $k = 1 \dots K$
for $k = 1$ **to** K **do**
  $\mathbf{P}, \mathbf{g} \leftarrow$ patch and global image features of $\mathcal{I}$
  $\mathbf{T}_k, \mathbf{t}_k \leftarrow$ token and global text features of $T_\pm^{(k)}$
  $\mathbf{S}^{(k)} \leftarrow \tilde{\mathbf{P}}\tilde{\mathbf{T}}_k^\top / \tau$ {adaptive τ per batch}
  $\sigma^{(k)} \leftarrow \mathrm{svd}(\mathbf{S}^{(k)})$ {spectral family}
  $\mathcal{O} \leftarrow$ token indices of o via offset mapping
  $\mathbf{s}_o \leftarrow$ per-patch alignment to object direction $\mathbf{t}_o$
  record spectral, POAS, global, Fourier, attention-flow descriptors
end for
aggregate each descriptor over K: mean_+, mean_-, std_+, std_-, $(\mathrm{mean}_+ - \mathrm{mean}_-)$
return 168-dimensional vector

Table S1: (continued) All 168 coordinates of the ACM descriptor.

Index	Coordinate name	Family
152	`neg_attn_flow_high_ratio`	POAS
153	`diff_attn_flow_high_ratio`	POAS
154	`pos_attn_flow_skewness`	POAS
155	`neg_attn_flow_skewness`	POAS
156	`diff_attn_flow_skewness`	POAS
157	`pos_hallucination_drive_intensity`	POAS
158	`neg_hallucination_drive_intensity`	POAS
159	`diff_hallucination_drive_intensity`	POAS
160	`pos_hallucination_drive_ratio`	POAS
161	`neg_hallucination_drive_ratio`	POAS
162	`diff_hallucination_drive_ratio`	POAS
163	`pos_attn_flow_gini`	POAS
164	`neg_attn_flow_gini`	POAS
165	`diff_attn_flow_gini`	POAS
166	`pos_attn_flow_kurtosis`	POAS
167	`neg_attn_flow_kurtosis`	POAS
168	`diff_attn_flow_kurtosis`	POAS

S2 MEASUREMENT PSEUDOCODE

Algorithm 1 summarises the deterministic measurement. No learned parameter, label, or access to the audited model is used; only a frozen CLIP encoder is queried.

S3 ADDITIONAL EXPERIMENTAL RESULTS

S3.1 CHAIR RESULTS AND THRESHOLD BEHAVIOUR

CHAIR (MSCOCO val2014, 2,994 (image, object) pairs over 500 images) is used here as an image source for the same discriminative existence task as POPE, rather than the generative caption-scoring protocol of Rohrbach et al. (2018) The AUROC below is therefore specific to this discriminative existence task. On this task the measurement reaches AUROC 0.8986 ± 0.0087 against 0.7986 for the best baseline, with a false-positive rate of 0.101 on grounded objects at the default operating point. Figure S1 shows the false-positive rate on grounded objects as a function of the operating threshold; the measurement has a lower false-positive rate than the baselines across the operating range. Per-split AUROC on POPE is 0.9315 (adversarial), 0.9703 (random) and 0.9756 (popular),

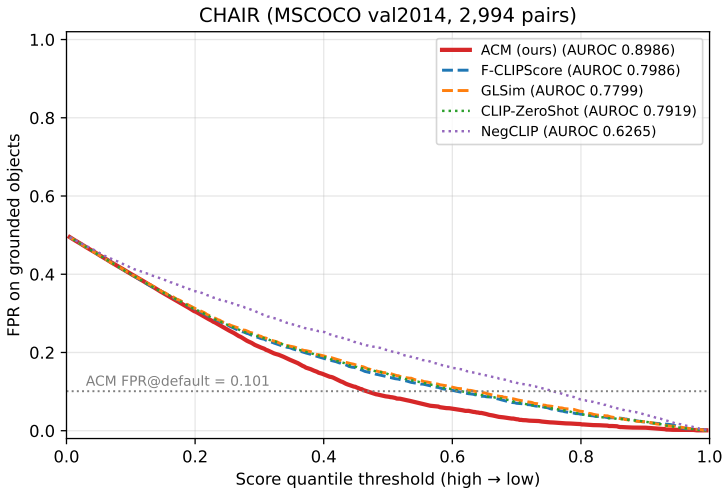

Figure S1: False-positive rate on grounded objects versus threshold on CHAIR. Lower is better; ACM (orange) stays below every baseline across the whole operating range.

from the pooled out-of-fold predictions of the full model; per-category results on CHAIR and the failure-case breakdown are produced by the released `scripts/`.

S3.2 Calibration of the consumer output

The measurement itself is a deterministic descriptor, not a probability; any calibration enters through the downstream consumer. Figure S2 shows the reliability curve of the gradient-boosted head's pooled out-of-fold output on POPE (Brier score 0.082). The head is over-confident in both tails—it predicts 0.86 where the observed frequency is 0.98—so the consumer output should not be read as a calibrated probability. This is a property of the head, not of the measurement: the raw coordinates are used exactly as extracted, and a user who needs calibrated scores can fit a sigmoid or isotonic map on a validation split, as with any classifier.

S3.3 Discriminative evidence for the κ sensor

Theorem 2 makes effective rank the sensor for the background-complexity parameter κ. On real POPE features the effective-rank block alone (five coordinates, pos/neg means, standard deviations and their difference) reaches AUROC 0.8125 ± 0.0068 under the standard protocol, and a shuffled-label control gives 0.50, so the reading the theory singles out is informative on real CLIP features rather than only in the idealised model. This complements the marginal analysis of Table S2: the sensor's information is multivariate (no single coordinate exceeds $|d| \approx 0.2$), which is why the theory treats it as a response to template variation rather than a level.

S3.4 Monotonicity in a continuous proxy of alignment strength

Identifiability concerns the response of the descriptor to the model parameters, for which real data has no ground truth; a partial, testable counterpart is monotonicity in a continuous proxy of alignment strength. On the CHAIR pairs, a grounded object's relative area in the image is such a proxy: a larger object carries stronger visual evidence, so a faithful measurement should assign it a lower hallucination score. Over the 1,497 grounded objects the pooled out-of-fold score falls monotonically with relative area (Spearman -0.40, $p < 10^{-57}$): the mean score is 0.395, 0.282, 0.178, 0.122 across area quartiles, and the false-positive rate at the default threshold falls from 0.37 to 0.08. The descriptor therefore tracks how strongly the object is visually grounded, not only which side of the label it falls on. This is the real-data counterpart of the identifiability argument in the main text: the measurement responds continuously to the quantity it claims to measure.

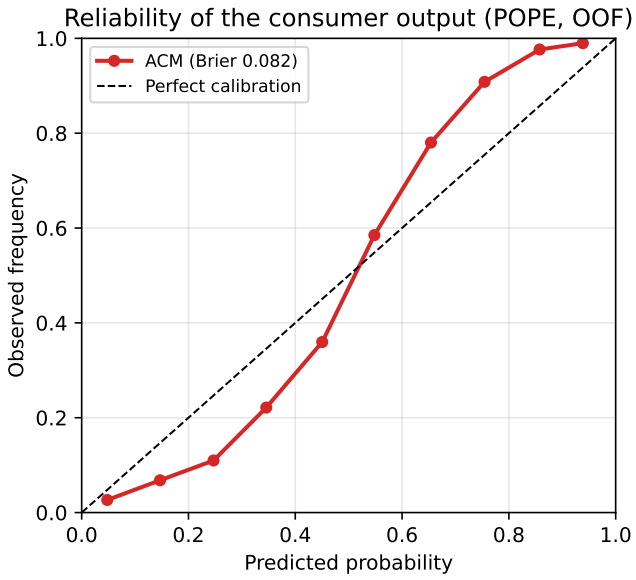

Figure S2: Reliability of the gradient-boosted head's pooled out-of-fold output on POPE (Brier 0.082). The head is over-confident in both tails; the consumer output is a discriminant score, not a calibrated probability.

S3.5 Residual errors and cross-backbone transfer

At the AUROC 0.9591 reported in the main text, about 5% of POPE pairs are still misclassified, and the errors are structured rather than scattered. Stratifying the pooled out-of-fold predictions by split reproduces the main-text ordering—adversarial 0.9315 versus random 0.9703 and popular 0.9756—and shows the failure is mostly false alarms in the adversarial split (248 false positives versus 132 false negatives there, against $131/124$ and $106/119$ on the random and popular splits). The ten object classes with the most errors are common, confusable, often co-occurring categories (person 98, car 64, chair 62, dining table 56, truck 52, ...): exactly the cases where correct evidence localisation leaves only a small alignment margin. Object size predicts the residual on the same samples: among grounded objects, the false-negative rate is 0.117 for small instances (COCO area $< 32^2$) versus 0.060 for larger ones, so small, partially occluded objects are about twice as likely to be missed. The residual errors are therefore concentrated on near-margin, adversarially chosen, sparsely grounded evidence, consistent with the measurement model.

The measurement protocol can be applied to any frozen encoder that supplies $\mathbf{S}$. Extracting the same 168 coordinates from DINOv2-Base and DINOv3-Base (ViT-B/16) and attaching the same classifier head gives POPE AUROC 0.8560 ± 0.0080 and 0.8589 ± 0.0058 respectively, against a random-feature control near chance; both still exceed every training-free baseline on POPE reported in the main text. The spectral and POAS families therefore transfer across visual backbones rather than being specific to CLIP; CLIP-ViT-L/14 remains the primary backbone throughout the main text.

S4 Spectral motivating analysis

Figure S3 shows the four spectral summaries on all 9,000 POPE pairs, split by label, at two levels of aggregation. The marginal *level* of each descriptor does not separate the classes; the *dispersion* across the five positive templates does, with hallucinated pairs showing larger cross-template dispersion in effective rank, spectral entropy and dominance ratio. Table S2 gives the statistics; the collapse signature is therefore not in the spectrum's level but in how the spectrum responds to the prompt.

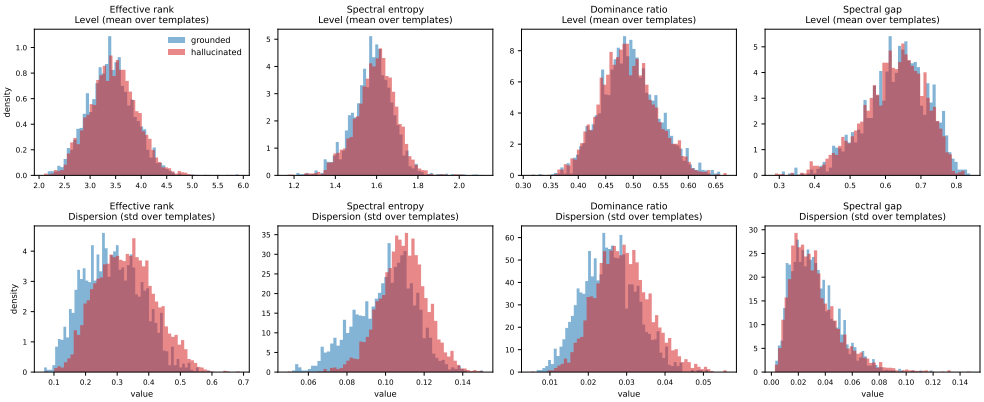

Figure S3: Spectral signatures of collapse on POPE (9,000 samples, CLIP-ViT-L/14). Top row: level of each spectral summary; bottom row: dispersion across the five positive templates. Hallucinated pairs show larger dispersion in effective rank, spectral entropy and dominance ratio.

Table S2: Spectral descriptors on POPE, split by label: marginal level and cross-template dispersion. Cohen's d is the standardised mean difference. The level does not separate the classes; the dispersion does, for three of the four descriptors.

Descriptor	Level		d	Dispersion		d
	Grounded	Hallucinated		Grounded	Hallucinated	
$\mathrm{effRank}(\mathbf{S})$	3.41 ± 0.45	3.46 ± 0.46	-0.11	0.286 ± 0.091	0.329 ± 0.092	-0.46
$H(\mathbf{S})$, spectral entropy	1.58 ± 0.10	1.60 ± 0.10	-0.16	0.100 ± 0.016	0.109 ± 0.012	-0.65
$\mathrm{domRatio}(\mathbf{S})$	0.494 ± 0.051	0.490 ± 0.050	$+0.08$	0.026 ± 0.007	0.029 ± 0.007	-0.49
$\Delta\sigma$, spectral gap	0.632 ± 0.088	0.623 ± 0.087	$+0.10$	0.032 ± 0.017	0.032 ± 0.018	-0.03

S5 Synthetic controlled study

The controlled checks of Table 2 in the main paper are reproduced by `scripts/run_theory_checks.py`; the response surfaces are shown here in Figure S4 over the (δ, κ) grid. POAS depends only on δ, while effective rank rises with both parameters, confirming the predicted block-diagonal Jacobian structure. For the column-mismatch check, the two members of the pair are drawn with independent noise realisations, so their spectra agree only up to the noise floor. Theorem 1 assumes a shared noise draw followed by a column permutation.

S6 Formal statements and proofs

This section restates the formal claims of Section 3.3 and proves them, so that the section is self-contained. The statements reuse the numbering shown here and use the measurement model Eq. 2: $\mathbf{S} = c_{-}\mathbf{J} + \delta\mathbf{M} + \mathbf{\Xi}$, where $\mathbf{J}$ is the all-ones matrix, $\mathbf{M} \in \{0,1\}^{N_p \times N_t}$ marks the patches and token columns that carry the object's evidence, and $\mathbf{\Xi}$ has i.i.d. zero-mean sub-Gaussian entries with variance ν^2; the object is present exactly when $\mathbf{M} \neq \mathbf{0}$.

Let $G = O(N_p) \times O(N_t)$ act on $\mathbb{R}^{N_p \times N_t}$ by $(\mathbf{U}_0, \mathbf{V}_0) \cdot \mathbf{S} = \mathbf{U}_0 \mathbf{S} \mathbf{V}_0^\top$. A reading f is *position free* if it is invariant under this action.

Lemma 1 (Maximal invariant)**.** *$f : \mathbb{R}^{N_p \times N_t} \to \mathbb{R}$ is G-invariant if and only if $f(\mathbf{S}) = \varphi(\sigma(\mathbf{S}))$ for some φ, where $\sigma(\mathbf{S})$ is the vector of singular values.*

Proof. Write $\mathbf{S} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$. If f is G-invariant, taking $\mathbf{U}_0 = \mathbf{U}^\top$ and $\mathbf{V}_0 = \mathbf{V}^\top$ gives $f(\mathbf{S}) = f(\mathbf{\Sigma})$, so f depends on $\mathbf{S}$ only through $\mathbf{\Sigma}$. Conversely singular values are invariant under left and right multiplication by orthogonal matrices (Horn & Johnson, 2012, §7.3), so $\varphi \circ \sigma$ is G-invariant. Hence the spectrum exhausts every position-free reading. □ □

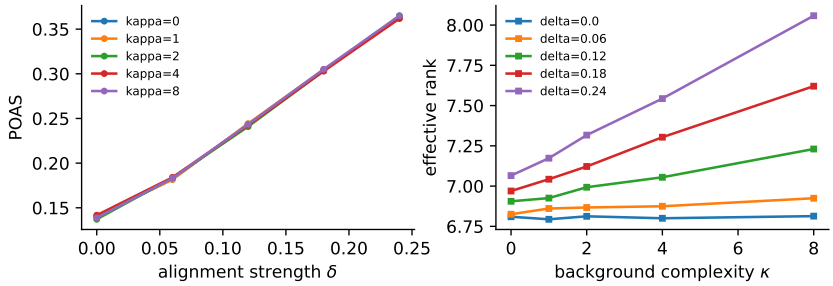

Figure S4: Estimated response of POAS (left) and effective rank (right) over the (δ, κ) grid. POAS tracks δ only; effective rank tracks both, matching the identifiability claim.

Theorem 1 (Spectral blindness to column mismatch)**.** *Under the measurement model Eq. 2, there exist samples with identical singular values and opposite ground-truth labels. Any reading that depends on* $\mathbf{S}$ *only through* $\sigma(\mathbf{S})$ *therefore has power equal to its test level on such a pair.*

Proof. Let $\mathcal{O}$ be the object-token columns, $\mathcal{W}$ a disjoint set of token columns with $|\mathcal{W}| = |\mathcal{O}|$, and $\mathcal{P}_{\text{obj}}$ the patches the object occupies. Set $\mathbf{M}_{\text{a}} = \mathbf{1}_{\mathcal{P}_{\text{obj}}}\mathbf{1}_{\mathcal{O}}^{\top}$ and let $\mathbf{Q}$ be the column permutation exchanging $\mathcal{O}$ with $\mathcal{W}$, so $\mathbf{M}_{\text{m}} := \mathbf{M}_{\text{a}}\mathbf{Q} = \mathbf{1}_{\mathcal{P}_{\text{obj}}}\mathbf{1}_{\mathcal{W}}^{\top}$. Both are $\{0,1\}$ blocks of the same footprint. Take a noise draw $\mathbf{\Xi}$ and set $\mathbf{\Xi}_{\text{m}} = \mathbf{\Xi}\mathbf{Q}$, which has the same distribution because $\mathbf{\Xi}$ is i.i.d. and $\mathbf{Q}$ is orthogonal. Since $\mathbf{J}\mathbf{Q} = \mathbf{J}$,

$$\mathbf{S}_{\text{m}} = c_{-}\mathbf{J} + \delta\mathbf{M}_{\text{a}}\mathbf{Q} + \mathbf{\Xi}\mathbf{Q} = (c_{-}\mathbf{J} + \delta\mathbf{M}_{\text{a}} + \mathbf{\Xi})\mathbf{Q} = \mathbf{S}_{\text{a}}\mathbf{Q},$$

so $\sigma(\mathbf{S}_{\text{m}}) = \sigma(\mathbf{S}_{\text{a}})$ exactly. But $\mathbf{S}_{\text{a}}$ places its evidence block on the object's own token columns, whereas in $\mathbf{S}_{\text{m}}$ the elevated similarity sits on columns that do not name the object, leaving the object-token columns at the baseline c_{-}. By Definition 1 the first is grounded and the second is collapsed, so the labels differ while every spectral reading agrees. □ □

Lemma 2 (POAS response)**.** *Under the measurement model Eq. 2,* $\text{POAS}(\mathbf{S}) = c_{-} + \max_p[\delta\bar{M}_p + \bar{\Xi}_p]$*, with* $\bar{M}_p = |\mathcal{O}|^{-1}\sum_{t\in\mathcal{O}} M_{p,t}$ *and* $\bar{\Xi}_p = |\mathcal{O}|^{-1}\sum_{t\in\mathcal{O}} \Xi_{p,t}$*. Hence* $\partial\text{POAS}/\partial\delta = 1$ *once* δ *exceeds the noise maximum* $\max_p \bar{\Xi}_p$*, and* $\text{POAS}(\mathbf{S}+\mathbf{E}) = \text{POAS}(\mathbf{S})$ *for every* $\mathbf{E}$ *with* $\mathbf{E}_{:,\mathcal{O}} = \mathbf{0}$*.*

Proof. By definition $\text{POAS}(\mathbf{S}) = \max_p |\mathcal{O}|^{-1}\sum_{t\in\mathcal{O}} S_{p,t}$, and under Eq. 2 the summand is $c_{-} + \delta M_{p,t} + \Xi_{p,t}$, giving $\text{POAS}(\mathbf{S}) = c_{-} + \max_p[\delta\bar{M}_p + \bar{\Xi}_p]$. For (i), when $\delta > \max_p \bar{\Xi}_p$ the maximiser lies in the support of $\bar{M}$, so the maximum is $\delta + \max_{p\in\mathcal{P}_{\text{obj}}} \bar{\Xi}_p$ and the derivative in δ is 1. For (ii), $(\mathbf{S}+\mathbf{E})_{:,\mathcal{O}} = \mathbf{S}_{:,\mathcal{O}}$, so the object-token profile and hence the maximum are unchanged. □ □

Theorem 2 (Identifiability)**.** *Let* $\mathbf{m}(\mathbf{S}) = (\text{POAS}(\mathbf{S}), \text{effRank}(\mathbf{S}))$ *and* $\mathbf{J}_\theta = \partial\mathbf{m}/\partial\theta$*. Under the measurement model Eq. 2, at any* θ *with* $\delta > \max_p \bar{\Xi}_p$ *and* $\kappa \geq 1$*,*

$$\mathbf{J}_\theta = \begin{pmatrix} 1 & 0 \\ \partial_\delta\text{effRank} & \partial_\kappa\text{effRank} \end{pmatrix}, \qquad \partial_\kappa\text{effRank} > 0,$$

so $\det\mathbf{J}_\theta = \partial_\kappa\text{effRank} \neq 0$ *and* $\theta \mapsto \mathbf{m}$ *is locally invertible. Either family alone gives one equation in two unknowns and leaves* θ *unidentified.*

Proof. The first row is Lemma 2. For the second, a background block supported on rows and columns disjoint from the object block adds a non-negligible singular value while leaving the existing spectrum essentially unchanged, so effective rank is strictly increasing in κ; monotonicity under the addition of a rank-one component with disjoint support follows from Weyl's interlacing inequalities (Weyl, 1912; Horn & Johnson, 2012). The determinant is then $\partial_\kappa\text{effRank}$ and the inverse function theorem gives local invertibility. A single family gives a 1×2 Jacobian of rank at most $1 < 2$, so one parameter remains unidentified. □ □

Corollary 1 (Minimal identifiability)**.** *Among readings of* $\mathbf{S}$*, the pair (spectral, POAS) is a minimal identifiable measurement set for* (δ, κ)*: it identifies both parameters, and dropping either family leaves one unidentified. If the alternative class contains column-mismatched samples of the form in Theorem 1, any purely spectral measurement has power equal to its level at any sample size.*

Proof. The first claim is Theorem 2: a single family gives a Jacobian of rank at most $1 < 2$. The second is Theorem 1: a spectral reading is a function of $\sigma(\mathbf{S})$, and the two matched samples are indistinguishable by the spectrum yet carry opposite labels. □ □

S7 HYPERPARAMETERS

All experiments use a single frozen `openai/clip-vit-large-patch14` encoder, $K = 5$ positive/negative template pairs (listed in `acm/measurement.py`), batch size 8, and the adaptive temperature of Section 3. The downstream consumer is a gradient-boosted classifier `scikit-learn GradientBoostingClassifier` with mutual-information selection of 150 coordinates, applied inside each cross-validation fold on the training partition only. Hyperparameters are selected on training folds. All random seeds are fixed at 42. Repeating the POPE protocol with selection seeds 0–4 (both the mutual-information selection and the consumer re-seeded) gives AUROC 0.9601 ± 0.0017 (min 0.9572, max 0.9621), so the family-level conclusions are stable to the choice of seed. Reported intervals are 5-fold stratified cross-validation with a paired bootstrap ($B = 10{,}000$).