CALIBRATED VISION-LANGUAGE MODELS MUST HALLUCINATE

Anonymous authors
Paper under double-blind review

ABSTRACT

Calibrated language models must hallucinate rare facts at a rate set by the missing mass of the training sample, yet this result is open domain and does not apply when a source document determines the answer. An image looks like such a source, so a naive transplant would predict a floor of zero. This paper shows when that closed domain caveat fails for object hallucination in vision language models. A vision language model does not observe the image directly but rather a feature produced by an encoder, a connector, and softmax attention, which we model as a bounded evidence channel with retention η. For any generatively calibrated captioner, the object hallucination rate is at least the visual monofact rate times the chance the image is unread, minus miscalibration. The bound recovers the text theorem at $\eta = 0$, vanishes at perfect vision, and is matched by an explicit generator. Forecast calibration on yes or no existence questions does not imply the same necessity on captions, and instruction tuned decoders occupy the abstention vertex of a linear coverage hallucination frontier, so low false Yes on a binary probe need not imply reliable free form description. Softmax attention converts visual token count into η, so adding unrelated visual tokens raises the generative floor. The identities are attained in a Zipf visual fact world and are checked on frozen Qwen3-VL models with official POPE questions and CHAIR style captions, confirming the theory within one percent relative error.

1 INTRODUCTION

Object hallucination, naming an object absent from the image, remains the central safety relevant failure of large vision language models (Rohrbach et al., 2018; Li et al., 2023). A long line of mitigation methods, from contrastive decoding to mutual information calibrated decoding, reduces measured error on standard meters (Leng et al., 2024; Huang et al., 2024; Zhou et al., 2024; Fang et al., 2026). The same literature documents recurring failure modes, including cooccurrence priors, language prior takeover, attention scatter, and a Yes bias under binary existence scoring (Li et al., 2023; Zhou et al., 2024; Huang et al., 2024; Fang et al., 2026; Park & Li, 2026; Seo et al., 2026; Wang et al., 2026; Fazli et al., 2026; Li et al., 2026). These observations are usually treated as separate engineering facts, evaluated on their own without a shared ceiling. None of them answers a more basic question: *How low can the hallucination rate of a calibrated vision language model go?*

For text the question has a sharp answer. Kalai & Vempala (2024) prove that a generatively calibrated language model hallucinates arbitrary facts at a rate close to the fraction of training facts that appear once, the Good Turing missing mass estimate. Kalai et al. (2026) recast the same phenomenon as an *Is It Valid* classification error and show that binary leaderboard scoring keeps a guessing incentive alive after alignment. A parallel line of work reaches related conclusions from other angles, from computability arguments to formal limits that separate consistency from breadth in language generation (Xu et al., 2024; Banerjee et al., 2025; Kleinberg & Mullainathan, 2024; Kalavasis et al., 2025; Das & Srivastava, 2026). All of these results are open domain. Kalai and Vempala explicitly exclude closed domain hallucination, fabrication relative to a given source document, because verification against the source permits filtering. An image looks like such a source. If the encoder recovered I, false objects could be filtered and the floor would vanish. A naive transplant of their bound would predict a floor of zero, since I would determine y.

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.

The transplant fails because the model does not observe I. It observes a feature $Z = \Phi(I)$ produced by an encoder, a connector, and softmax attention over N_v visual tokens. Attention mass on the relevant patches behaves as an erasure probability η. With probability η the visual fact is revealed, and with probability $1 - \eta$ the model falls back to the open domain regime. Hallucination then requires both failures to occur simultaneously. The fact must be a training singleton, and the current image must be unread. This conjunction yields a multiplicative lower bound on the hallucination rate:

$$\mathrm{hallu}(g) \;\geq\; (1-\eta)\,\widehat{\mathrm{MF}}_v - \mathrm{Mis} - o(1), \tag{1}$$

where Mis denotes miscalibration, defined formally in Section 3. At $\eta = 0$ one recovers the text bound. At $\eta = 1$ the bound is 0, in exact agreement with the closed domain caveat. For $\eta \in (0, 1)$, the image acts as a second view of the fact, an information structure that text alone does not have.

The product is a statistical floor rather than a pair of fitted heuristics. An explicit generator $g_\star$ matches equation 1 on the erasure channel, so the order is tight once the model is generatively calibrated and still answers on the unread slice. A Visual *Is It Valid* reduction yields the same order from a Bayes error on Z, which shows the floor is not an artefact of one calibration definition. Softmax competition converts N_v into η, so the floor rises when unrelated visual tokens are added. Retention is random across images, and the typical image then sits above the annealed substitution $\eta \approx M/(M + N_v)$. This gives a monotone see more invent more prediction that links the mass budget of attention to the hallucination rate.

Calibration, low hallucination, and high coverage cannot hold together. Abstention is not a hallucination, so coverage α and hallucination trace a straight Pareto line through $(\eta, 0)$ and $(1, (1-\eta)\widehat{\mathrm{MF}}_v)$. Binary polling based object probing evaluation (POPE) scoring has no strictly optimal abstention point and therefore selects the full coverage vertex (Li et al., 2023; Kalai et al., 2026). Evidence gated abstention selects the opposite vertex, and conformal filtering realizes it in practice (Li et al., 2025). Instruction tuned yes or no decoders behave as if false positives were far more costly than misses, which places them at that same conservative vertex. Forecast calibration on existence logits does not force generative calibration on captions, so a model can be nearly perfect on POPE while still emitting ungrounded object mentions in free form description (Xuan et al., 2025; Xiao et al., 2026; Liu et al., 2026). Low false Yes on a binary probe is therefore not a certificate of caption safety. Caption hallucination assessment with image relevance (CHAIR) style caption metrics are the generative functional to which equation 1 applies (Rohrbach et al., 2018).

We develop a self contained theory of calibrated hallucination under a bounded visual channel and test the resulting identities in a controlled Zipf world and on frozen open vision language models. The main contributions of this work are summarized as follows.

- A tight multiplicative floor for generatively calibrated captioners, interpolating the open domain text theorem and the closed domain caveat, with a matching generator and a Visual *Is It Valid* companion of the same order.
- A softmax lemma that turns visual token count into retention, a monotone see more invent more prediction, and a typical versus annealed refinement of the floor.
- A linear coverage hallucination frontier, together with scoring theorems that place binary POPE and instruction tuned policies at opposite vertices, and a separation between forecast calibration and generative calibration that identifies CHAIR as the quantity governed by the floor.

2 Related Work

Hallucination in language models. Kalai & Vempala (2024) introduce generative calibration as a coarsening of the factoid distribution and prove that a calibrated language model hallucinates arbitrary facts at a rate bounded by the monofact rate, the Good Turing estimate of missing mass. Kalai et al. (2026) recast this phenomenon as an *Is It Valid* classification error and show that binary leaderboard scoring preserves a guessing incentive after alignment. Their remark that training may include multimodal inputs does not introduce a visual channel or evidence parameter, which is the gap this paper addresses. Das & Srivastava (2026) identify text hallucination with an innovation rate, and a visual innovation rate should subtract the mass that the current image has already revealed, which is the factor $1-\eta$ in our bound. Beyond calibration, computability arguments show that some hallucinating input exists for every computable model (Xu et al., 2024; Banerjee et al., 2025), and

consistency versus breadth trade offs in the Gold Angluin model provide a non probabilistic cousin of our coverage versus hallucination line (Kleinberg & Mullainathan, 2024; Kalavasis et al., 2025). Mechanized safety and liveness proofs for asynchronous Byzantine consensus make the same split between coverage and correctness (Qiu et al., 2026a;b). Liu et al. (2025) prove that even Bayes optimal estimators can δ hallucinate under multimodal targets, with text to image experiments that approach but do not enter the calibrated visual channel regime studied here.

Object hallucination in vision-language models. CHAIR, POPE, and AMBER are the standard meters of object hallucination (Rohrbach et al., 2018; Li et al., 2023; Wang et al., 2023), and a large body of empirical work documents recurring failure modes including cooccurrence priors, language prior dominance, attention scatter, and knowledge entanglement across mixed sources (Li et al., 2023; Zhou et al., 2024; Leng et al., 2024; Huang et al., 2024; Fang et al., 2026; Wang et al., 2026; Fazli et al., 2026; Li et al., 2026; Qian et al., 2026; Zhu et al., 2026). Two findings are direct evidence for the erasure channel we model. Tong et al. (2024) construct CLIP blind pairs, images that CLIP embeds as near duplicates despite being visually distinct, which is an empirical $G = 0$ event. Lin et al. (2024) show that a blind language model ignoring the image can outperform vision language models on several retrieval benchmarks, which is the $\eta = 0$ slice of equation 1. Composed image retrieval likewise fails when Φ does not retain the intended visual edit (Tang et al., 2026). On the mitigation side, contrastive decoding, over trust penalties, post hoc rewriting, and mutual information calibrated decoding all reduce measured error but do not produce a calibration floor (Leng et al., 2024; Huang et al., 2024; Zhou et al., 2024; Fang et al., 2026). Fang et al. (2026) note that these methods are typically grounded in empirical findings and lack convincing theoretical foundations. Spectral energy bounds and attention calibration quantify or reduce hallucination but likewise stop short of a lower bound (Sarkar & Das, 2025; Park & Li, 2026; Seo et al., 2026; Ye et al., 2025). ConfLVLM gives a finite sample upper bound on residual risk by filtering claims (Li et al., 2025), which occupies the abstention vertex of our triangle and is complementary to a lower bound on calibrated generators. The mass law that produces η is self contained in Lemma 2, and a needle query can remain solvable after free form coverage of many objects has already risen, because top one retrieval and mass budget are different functionals of the same logits.

Calibration and proper scoring. Strictly proper scoring rules are characterized by Gneiting & Raftery (2007), and binary scoring with an abstention option is not strictly proper, which is why POPE cannot elicit an honest posterior. The coverage hallucination frontier in this paper inherits the same scoring geometry, with the error penalty c selecting a point on the line and the scoring rule fixing the slope. Missing mass is the Good Turing estimator (McAllester & Schapire, 2000), and Fano's inequality converts a capacity constraint into a Bayes error floor, which is the relaxation we record in Appendix E. A growing line of work studies confidence calibration in vision language models. Verbalized calibration analyzes how reported confidences relate to visual evidence (Xuan et al., 2025), decoupled confidence calibration separates visual and reasoning uncertainties (Xiao et al., 2026), and verified atomic confidence estimation probes object level confidence at a finer granularity (Liu et al., 2026). These methods estimate or improve calibration but do not derive a lower bound on hallucination, which is the gap our floor fills. Internal probes can predict factual errors from activations that are not the reported confidence (Kadavath et al., 2022), and Appendix D explains why a detector that only reads a calibrated forecast cannot separate true from false statements inside one confidence bin.

3 SETUP

Let $I \in \mathcal{I}$ denote an image with finite object set $O(I) \subseteq \mathcal{O} = \{o_1, \ldots, o_K\}$. A visual statement $s \in \mathcal{S}$ has truth value $y(s, I) \in \{0, 1\}$, and the running case is existence, so $y(s_o, I) = \mathbf{1}\{o \in O(I)\}$. Abstention $\perp$ is scored as non hallucination by setting $y(\perp, I) = 1$. A document is one extracted factoid from an image and text training sample, following the one per document idealization of Kalai & Vempala (2024). The case of many facts per document is deferred to Appendix F.3. The language distribution p is the law of (I, s), and training consists of n i.i.d. documents without labeling error. Caption noise in the wild can only raise empirical hallucination, so a lower bound remains valid. The visual monofact rate is the Good Turing singleton frequency

$$\widehat{\mathrm{MF}}_v \;:=\; n^{-1}\,\#\{\text{training factoids appearing once}\}. \tag{2}$$

Let U denote the unobserved factoids and $p(U)$ the missing mass. Then $|p(U) - \widehat{\mathrm{MF}}_v| = \tilde{O}(n^{-1/2})$ with high probability for every p (McAllester & Schapire, 2000). Denote the set of true visual facts as F and the remainder as $H = \mathcal{S} \setminus F$.

Assumption 1 (Sparsity and regularity)**.** There is $s \geq 0$ with $|F| \leq e^{-s}|H|$. There is $r \geq 1$ such that no unobserved factoid is more than r times as likely, given the training sample, as the average unobserved factoid (Kalai & Vempala, 2024).

Sparsity says that false visual statements vastly outnumber true ones. Regularity rules out a single prototype object that absorbs almost all of the missing mass. These two conditions are inherited from the text setting of Kalai & Vempala (2024) and are needed only to control the tail of the unobserved pool. POPE popular negatives deliberately violate a small r regime and are predicted to hallucinate more, which matches the published ranking of POPE splits (Li et al., 2023).

A large vision language model (LVLM) is a conditional generator $g(\cdot \mid I)$, optionally reporting a confidence $\hat{p}(s \mid I) \in [0, 1]$. The generator does not observe I directly but rather a feature $Z = \Phi(I, \xi)$ produced by an encoder, a connector, and softmax attention. Hallucination and coverage are

$$\mathrm{hallu}(g) := \Pr_{I, s \sim g(\cdot|I)}\big[y(s, I) = 0\big], \qquad \alpha(g) := \Pr[s \neq \perp]. \tag{3}$$

Abstention is a coverage cost rather than a hallucination, and this separation is what makes the Pareto line of Section 5 linear.

Assumption 2 (Erasure)**.** There is $\eta \in [0, 1]$ such that a Bernoulli gate G of mean η, independent of y given revealability, satisfies the following. If $G = 1$ then y is a measurable function of Z. If $G = 0$ then $Z \perp y$.

Assumption 3 (Capacity)**.** The mutual information satisfies $I(Z; y) \leq C$. Erasure is the special case $C \leq \eta\, H(y)$ of Shannon entropy $H(y)$.

Erasure is a first order model of softmax sparsity. Attention either hits the relevant patches or it does not, and Lemma 2 derives η from the logits. CLIP blind pairs (Tong et al., 2024) are an empirical $G = 0$ event, and a posterior that takes values only in $\{0, 1, \pi\}$ is equivalent to hard erasure. Soft channels such as a binary symmetric channel replace the product $(1 - \eta)\widehat{\mathrm{MF}}_v$ by a Fano transform of the same capacity, which is recorded in Appendix E.

Proposition 1 (Product form)**.** *The following are equivalent. Hard erasure of retention η. The posterior $P(y \mid Z)$ taking values in $\{0, 1, \pi\}$ almost surely. Either condition yields the product floor of Theorem 1. A binary symmetric channel whose crossover is not 0 or $1/2$ does not collapse the posterior to that set.*

Proposition 1 is the reason the paper works with erasure rather than with a generic capacity constraint. The factor $1 - \eta$ is then an identity on the unread slice rather than an inverse entropy transform of C. The Fano and binary symmetric curves remain available as relaxations when the encoder always leaks a weak signal.

Definition 1 (Average and generative calibration)**.** (A) $\mathbb{E}[y \mid \hat{p} = q] = q$, and $\mathrm{Mis}_A = \mathrm{ECE}_b$ is the b bin expected calibration error (ECE). (G) g is a coarsening of p in the sense of Kalai & Vempala (2024), and $\mathrm{Mis}_G = \|p^{\mathcal{V}_b(g)} - g\|_{\mathrm{TV}}$ is the total variation (TV) gap.

Grounded and event algebra variants sit in Appendix B. The main bound is proved for (A) and (G) and therefore holds for the stronger two. A single threshold gap $\delta = |g(\mathcal{A}) - p(\mathcal{A})|$, denoting $\mathcal{A}$ as $\{g > t\}$, is the derivative of cross entropy under a one parameter rescaling of mass on $\mathcal{A}$ (Kalai et al., 2026), and local optimization in a class closed under that rescaling drives δ to zero. Forecast calibration is (A) on an existence bit, while generative calibration is (G) on the occupancy measure of emitted objects. The two coincide only under extra structure that instruction tuned yes or no heads need not satisfy.

A numerical snapshot makes the scale concrete. Fix $K = 300$ objects, $n = 1600$ training documents, retention $\eta = 0.4$, and an observed monofact rate $\widehat{\mathrm{MF}}_v = 0.046$. The product floor is then $(1 - \eta)\widehat{\mathrm{MF}}_v = 0.028$. A calibrated generator that always answers when the image is unread must hallucinate at least 2.8% of the time, even if the visible branch is perfect. Figure 2 returns to this operating point, and a three object worked example is in Appendix H.

4 THE EVIDENCE FLOOR

Let η denote visual retention and $\widehat{\mathrm{MF}}_v$ the visual monofact rate. Hallucination requires two failures, since the fact must be missing from training and the current image must be unread. The argument first restricts to the unread slice, where the text theorem of Kalai & Vempala (2024) applies verbatim, and then mixes with the visible slice, whose error can be dropped for a lower bound.

Lemma 1 (Kalai and Vempala on the erasure slice)**.** *Under Assumptions 2 and 1, every g that is generatively calibrated in the sense (G) satisfies, with high probability over the training sample,*

$$\mathbb{E}[\mathrm{hallu}(g) \mid G = 0] \;\geq\; \widehat{\mathrm{MF}}_v - \mathrm{Mis}_G - re^{-s} - \tilde{O}(n^{-1/2}). \tag{4}$$

The (A) variant replaces Mis_G *by* $\mathrm{ECE}_b + O(1/b)$.

Given $G = 0$, Z is independent of y, so unobserved visual facts remain unpredictable. Regularity and sparsity pass to the slice because the gate is independent of the fact labels, and the Good Turing statistic is a property of training counts rather than of the test time gate.

Theorem 1 (Main lower bound)**.** *Under the hypotheses of Lemma 1,*

$$\mathrm{hallu}(g) \;\geq\; (1-\eta)\big(\widehat{\mathrm{MF}}_v - \tau\big) - \mathrm{Mis} - re^{-s}, \tag{5}$$

with $\tau = \tilde{O}(n^{-1/2})$. *Equivalently* $\mathrm{hallu} \geq (1-\eta)\widehat{\mathrm{MF}}_v - \mathrm{Mis} - o(1)$.

Decompose $\mathrm{hallu} = \eta\,\mathbb{E}[\mathrm{hallu} \mid G = 1] + (1-\eta)\,\mathbb{E}[\mathrm{hallu} \mid G = 0]$ and drop the visible term. On $\{G = 0\}$, $Z \perp y$, so unobserved visual facts are as unpredictable as in the text theorem, and Lemma 1 supplies the slice bound $\widehat{\mathrm{MF}}_v - \mathrm{Mis} - o(1)$. Multiplying by $1 - \eta$ is the claim. Miscalibration may leak mass off the slice, and total variation or binned ECE controls the leak. The matching generator below plays the text construction on $\{G = 0\}$ and a true fact on $\{G = 1\}$, so the product is attained rather than an artefact of a slack inequality.

Theorem 2 (Matching generator)**.** *There exists a generator $g_\star$ with* $\mathrm{Mis}_G = O(e^{-s} + n^{-1/2})$ *and*

$$\mathrm{hallu}(g_\star) \;\leq\; (1-\eta)\widehat{\mathrm{MF}}_v + O(e^{-s} + n^{-1/2}). \tag{6}$$

The construction emits a true fact when $G = 1$ and spreads the Good Turing mass uniformly on the unobserved pool when $G = 0$, so $(1-\eta)\widehat{\mathrm{MF}}_v$ is tight. A sum would double count the visible slice, and a minimum would vanish at either perfect vision or an exhausted training sample. Only the product interpolates both the open domain endpoint $\eta = 0$ and the closed domain endpoint $\eta = 1$. An encoder that fails even when $G = 1$ can only raise empirical hallucination above the floor. Temperature sampling, nucleus sampling, and random decoding seeds are mixtures of generators, and hallucination is linear in g, so no such mixture beats equation 5 without raising η or abstaining.

Corollary 1 (Interpolation)**.** $\eta = 0$ *recovers the open domain bound of Kalai & Vempala (2024), and* $\eta = 1$ *recovers their closed domain caveat. For* $\eta \in (0, 1)$ *a hallucination is a fact that is missing from training and unread in the current image.*

Das & Srivastava (2026) identify text hallucination with an innovation rate. Visually a statement can be unseen in training and still true of I, so let $\mathrm{innov}_v := \Pr(s \notin \mathrm{train} \wedge G = 0)$ denote the unread slice innovation. Only this quantity is a candidate hallucination, as recorded in Appendix F.4. Any post processing $Z' = \Psi(Z)$ obeys $\eta' \leq \eta$ by the data processing inequality, so no contrastive merge or probe of Z can drive the floor below $(1 - \eta_{\mathrm{Bayes}})\widehat{\mathrm{MF}}_v$. The coarsening proof uses generative calibration, while the next reduction uses only a classifier error, so the same order is not an artefact of definition (G).

Fix a prompt class and partition responses into valid $\mathcal{V}_c$, erroneous $\mathcal{E}_c$, and $\perp$. Visual *Is It Valid* draws a labelled example from a balanced mix of p and the uniform law $\mathrm{Unif}(\mathcal{E}_c)$ and uses the generator as a classifier by thresholding $g(s \mid I)$.

Theorem 3 (Visual Is It Valid)**.** *For any p without labeling error and any g, denoting the error of this test as* $\mathrm{err}_{\mathrm{iiv}}$*,*

$$\mathrm{hallu}(g) \;\geq\; 2\,\mathrm{err}_{\mathrm{iiv}} - \frac{\max_c |\mathcal{V}_c|}{\min_c |\mathcal{E}_c|} - \delta. \tag{7}$$

The algebra is that of Kalai et al. (2026), where Z enters only as an additional input to the classifier. Extra information can only lower $\mathrm{err}_{\mathrm{iiv}}$ and therefore relax the hallucination bound. Under Assumption 2 the Bayes error of this test is $(1-\eta)\varepsilon_0$, where ε_0 is the no vision error. For balanced existence questions $\varepsilon_0 = 1/2$, and for arbitrary facts ε_0 is of order $\widehat{\mathrm{MF}}_v$. Under Assumption 3 alone, Fano's inequality with binary entropy h_2 gives a Bayes error $h_2^{-1}((h_2(\rho) - C)_+)$ for binary y of bias ρ. Binary Fano is not tight on a two point alphabet, as shown in Figure 1. Erasure attains $P_e = (1-\eta)\min(\rho, 1-\rho)$, while a binary symmetric channel of the same capacity meets the Fano inverse. The paper uses erasure because $1-\eta$ is then an identity rather than an inverse entropy transform. The two proofs, generative coarsening and *Is It Valid*, show that the floor is not an artefact of one calibration definition.

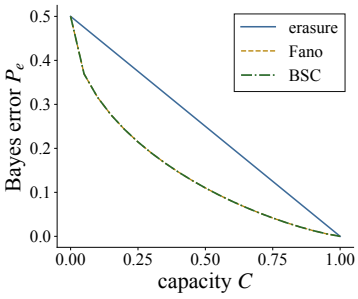


Figure 1: Erasure, Fano, and a binary symmetric channel. Erasure supplies the factor $1-\eta$.

Noise tokens compete with signal tokens in the softmax. As N_v grows, the share of attention mass on the patches that carry the answer falls, so $1-\eta$ rises and the floor of Theorem 1 rises with it.

Lemma 2 (Attention retention)**.** *Let m signal logits be $\Delta + \sigma_S \xi_i$ and N_v noise logits be $\sigma_V \zeta_j$, with ξ, ζ i.i.d. standard normal. Let $\eta = A/(A+B)$ denote the softmax mass on the signal set, and let $\hat{A} = Ae^{-\sigma_V^2/2}$. Then $\eta/(\hat{A}/(\hat{A}+N_v)) \to_p 1$. In the self averaging regime $\hat{A} \to_p M := m\exp(\Delta + \frac{1}{2}(\sigma_S^2 - \sigma_V^2))$, hence $\eta(N_v) \approx M/(M+N_v)$, with half mass budget $N_v^\star = M$.*

Theorem 4 (See more, invent more)**.** *In the self averaging regime of Lemma 2,*

$$\mathrm{hallu}(N_v) \;\geq\; \frac{N_v}{M+N_v}\,\widehat{\mathrm{MF}}_v - \mathrm{Mis} - o(1). \tag{8}$$

The right hand side is strictly increasing in N_v, crosses $\widehat{\mathrm{MF}}_v/2$ at $N_v = M$, and tends to $\widehat{\mathrm{MF}}_v$ as $N_v \to \infty$.

Retention $\eta(I)$ is random across images. The annealed substitution uses $\mathbb{E}\eta$, while the typical image uses the median. Whenever the median lies below the mean, which holds for right skewed log normal signal mass, the typical image floor lies above the annealed curve. Let Φ_{N} denote the standard normal cdf. A retrieval hit probability $\Phi_{\mathrm{N}}(\Delta/\sigma - \sqrt{2\log N_v})$ typically collapses at an exponentially larger N_v than M. There is therefore a window in which POPE existence of one named object still succeeds while free form CHAIR, which needs many objects, has already risen with the mass budget floor. Raising resolution can help only insofar as it raises effective η, for instance by enlarging the signal budget M. Adding independent visual noise raises N_v in Lemma 2, while upsampling a correlated field typically lengthens a spatial block and need not add a new noise population. Last layer attention mass is a convenient proxy for η rather than η itself, and the next lemma records what that distinction costs.

Lemma 3 (Attention proxies)**.** *On a single image the needle span is identically one. A top quantile concentration statistic is bounded away from zero under uniform attention. There exist channels with $I(Z;y) = 0$ and arbitrarily peaked attention. Substituting a positively biased proxy $\hat{\eta}$ for η reports a floor that lies strictly below $(1-\eta)\widehat{\mathrm{MF}}_v$.*

Identifiable instruments include multi image needle dilution, patch masks, and CLIP blind pairs. Extra images raise N_v while holding the needle fixed, so a rise in CHAIR cannot be attributed to a change of the query. A gate that thresholds a single image concentration statistic therefore cuts coverage while leaving generative hallucination nearly unchanged, which is the pattern later read from frozen captions. Existence, attributes, and relations differ in how many independent evidence pieces they need. Under i.i.d. AND gates the same bound holds with $\eta \leftarrow \eta^m$, so relation floors exceed existence floors when the monofact scale is shared, as formalized in Appendix D.

5 Coverage, Scoring, and Two Calibrations

The floor of Theorem 1 was derived under full coverage on the unread slice. Abstention is not a hallucination, so a calibrated generator can trade coverage for safety by refusing to answer. Let $\alpha_0 = \Pr[s \neq \perp \mid G = 0]$ be the erasure time coverage, and let α be overall coverage.

Theorem 5 (Coverage hallucination line)**.** *For calibrated g,*

$$\mathrm{hallu}(g) \;\geq\; (1-\eta)\alpha_0\,\widehat{\mathrm{MF}}_v - \mathrm{Mis} - o(1) = \max(0, \alpha-\eta)\,\widehat{\mathrm{MF}}_v - \mathrm{Mis} - o(1), \tag{9}$$

where the identity uses $\alpha = \eta + (1-\eta)\alpha_0$. The Pareto frontier is the line segment joining $(\alpha, \mathrm{hallu}) = (\eta, 0)$ and $(1, (1-\eta)\widehat{\mathrm{MF}}_v)$.

Three vertices organize the feasible set. Honest gating ($\alpha_0 = 0$) gives $\mathrm{hallu} = 0$ at coverage at most η. Full coverage gives $\mathrm{hallu} \geq (1-\eta)\widehat{\mathrm{MF}}_v$, and zero hallucination at full coverage requires $\mathrm{Mis} \geq (1-\eta)\widehat{\mathrm{MF}}_v - o(1)$. Calibration, low hallucination, and high coverage are not jointly attainable when $\eta < 1$ and $\widehat{\mathrm{MF}}_v > 0$. ConfLVLM occupies the first vertex by filtering (Li et al., 2025), and binary POPE occupies the second by scoring. The triangle is a scoring diagram, where changing the threshold or the abstention rule moves the operating point along the line, and changing the encoder, and thereby η, moves the line itself.

A POPE item is answered Yes, No, or I don't know (IDK). Let $q = \Pr(y = 1 \mid Z)$ denote the posterior belief. A correct hard answer pays $+1$, an error pays $-c$, and IDK pays 0.

Theorem 6 (Incentive gap)**.** *If $c = 0$, IDK is never strictly optimal. If $c > 0$, the optimal policy answers Yes iff $q > c/(1+c)$, No iff $q < 1/(1+c)$, and abstains otherwise. The $c = 0$ rule is not a strictly proper scoring rule (Gneiting & Raftery, 2007). On the erasure slice q equals the language prior π, so $c = 0$ forces a guess on every imbalanced object, with expected gain $\mathbb{E}[\max(\pi, 1-\pi)]$ over honest IDK.*

Corollary 2. *A calibrated model that is optimal for the $c = 0$ score has $\alpha_0 = 1$ and therefore sits on the full coverage ray of Theorem 5. Changing c moves the operating point along the Pareto line without a weight update.*

A strictly proper score elicits q, and a downstream threshold converts q into a hard action. The $c = 0$ convention collapses those layers, never eliciting q. Instruction tuned policies use an asymmetric c, since safety training and refusal make a fabricated Yes costlier than a missed object.

Theorem 7 (Asymmetric Instruct costs)**.** *If a fabricated Yes costs c_{fp} and a missed object costs c_{fn}, the optimal Yes threshold is $c_{\mathrm{fp}}/(c_{\mathrm{fp}} + c_{\mathrm{fn}})$. As the ratio diverges the decoder occupies the gating vertex of Theorem 5 on the erasure slice.*

On that slice $q = \pi(o) < 1$ for almost every object, so erasure time coverage vanishes. Raising an exogenous symmetric c in the decision rule is then predicted to lower both coverage and false Yes, which is a training free intervention and the quantity later varied on frozen POPE logits.

Let $\mathrm{Gate}(\eta_\star)$ denote the policy that emits $\perp$ when an estimator $\hat{\eta}$ falls below $\eta_\star$. If $\hat{\eta}$ recovers G, the gate attains $(\alpha, \mathrm{hallu}) = (\eta, 0)$. Among decoders that see only Z, no rule beats $\mathrm{hallu} \geq (1-\eta_{\mathrm{ans}})\alpha_{\mathrm{ans}}\widehat{\mathrm{MF}}_v - \mathrm{Mis}$, denoting η_{ans} as $\mathbb{E}[\eta \mid \text{answer}]$. Contrastive decoding can raise η to some $\eta' \leq \eta_{\mathrm{Bayes}}$ and cannot drive the floor to zero without abstention. VCD, OPERA, and CMI-VLD modify $g(\cdot \mid I)$ while remaining close to $\alpha_0 = 1$ (Leng et al., 2024; Huang et al., 2024; Fang et al., 2026), compatible with large empirical gains and a strictly positive floor.

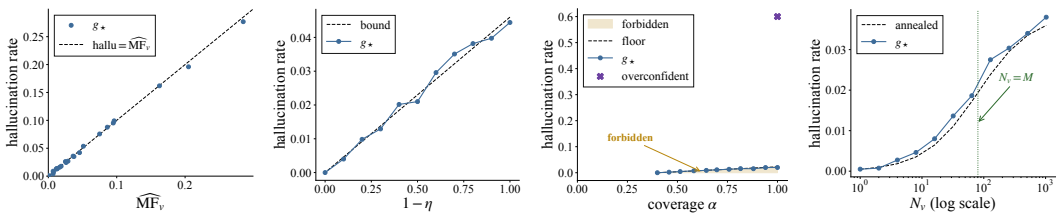
Forecast calibration on an existence bit does not imply generative calibration on captions. A model can report a calibrated probability on every yes or no probe and still emit ungrounded objects in free form description, so the two notions are genuinely different. The next theorem records both the separation and the floor that generative calibration restores.

Theorem 8 (Two calibrations)**.** *There exist models with vanishing ECE on existence forecasts and arbitrarily small false Yes under a high decision threshold, whose occupancy measure on emitted caption objects is not a coarsening of p. Conversely, if the occupancy measure is generatively calibrated, then a uniformly drawn mention o satisfies*

$$\Pr(o \notin O(I)) \;\geq\; (1-\eta)\widehat{\mathrm{MF}}_v - \mathrm{Mis}_G - O(e^{-s} + n^{-1/2}). \tag{10}$$

The left side is the mention level CHAIR rate. If the expected number of mentions is m_S, the sentence level rate is at least $1 - (1 - \mathrm{CHAIR}_i)^{m_S}$.

The first claim reports $\hat{p} = y$ when visible and $\hat{p} = \pi$ when erased, answers Yes only at a high threshold, and samples caption objects from π on the unread slice. The second claim is Theorem 1



(a) No vision. (b) Product floor. (c) Coverage line. (d) Token dilution.

Figure 2: The matching generator $g_\star$ attains the calibrated floor in a Zipf visual fact world.

applied to occupancy. POPE existence is a forecast decision with $m_S = 1$ and an option to say No. CHAIR is a generative functional with $m_S \geq 1$ and little penalty for a superfluous noun. Hallucination is the act of emitting a false object, not the act of assigning it a small probability, so the two meters can move in opposite directions without contradicting equation 1. Under average calibration, any detector that reads only $\hat{p}$ has equal true and false positive rates inside one confidence bin, so within bin AUROC (area under the ROC curve) is $1/2$, as proved in Appendix D. Detection needs η, a perturbation divergence, or semantic entropy (Kadavath et al., 2022; Leng et al., 2024).

6 Experiments

We test the theory in two complementary regimes. The first is a controlled Zipf visual fact world, where objects follow $p_k \propto k^{-\beta}$, training yields an empirical monofact rate $\widehat{\mathrm{MF}}_v$, and test time erasure implements Assumption 2 with the matching generator $g_\star$. The second regime uses frozen Qwen3-VL Instruct models (Bai et al., 2025), where we read first token yes or no logits, last layer attention mass as a snapshot of η, official POPE splits on COCO val2014, and one sentence captions scored with a CHAIR style lexicon against COCO labels. Protocol details and additional grids are in Appendix I.

6.1 Zipf Visual Fact World

The product floor is tight in the Zipf world, and the typical floor sits above the annealed one. At $\eta = 0$, ordinary least squares of hallu on $\widehat{\mathrm{MF}}_v$ through the origin has slope 0.979 and mean absolute gap 0.0018 across a (β, n) grid, as shown in Figure 2a. Sweeping η at fixed training sample with $\widehat{\mathrm{MF}}_v = 0.046$, $g_\star$ tracks $(1-\eta)\widehat{\mathrm{MF}}_v$ with relative slope error 0.42% and $\mathrm{ECE} = 0$, as shown in Figure 2b. An overconfident policy that always names an object on the unread slice lies far above that line, not a calibrated coarsening. Inserting sampled softmax $\eta(N_v)$ into $g_\star$ produces a monotone curve against $N_v \in [1, 1024]$ that follows the annealed bound and crosses half mass near $M \approx 80$, as shown in Figure 2d. A needle retrieval probability remains above one half past that crossing, the POPE versus CHAIR window predicted by Theorem 4.

When signal logits fluctuate, the median image retains less mass than the annealed mean. The median hallucination rate then tracks $(1-\eta_{\mathrm{typ}})\widehat{\mathrm{MF}}_v$ and lies above the annealed curve by $7 \cdot 10^{-3}$ on average, as shown in Figure 3. Good Turing itself has mean $|\widehat{\mathrm{MF}}_v - p(U)| = 0.0095$. At $\eta = 0.4$, varying erasure time coverage traces the Pareto line, and none of the calibrated points enter the forbidden region, as shown in Figure 2c. Raising the error penalty c walks coverage along that line once $c/(1+c)$ enters the bulk of the Zipf prior. Five independent seeds give a relative slope 0.975 ± 0.018 of hallucination on the product, confirming tightness. Caption noise raises empirical hallucination strictly above the clean floor, the direction required of a lower bound under labeling error. AND arity increases the floor as $1-\eta^m$ and OR reverses it, matching Theorem 10. A gate that flips G with probability ϵ traces $(1-\eta)\epsilon\,\widehat{\mathrm{MF}}_v$. These identities are collected in Appendix J.

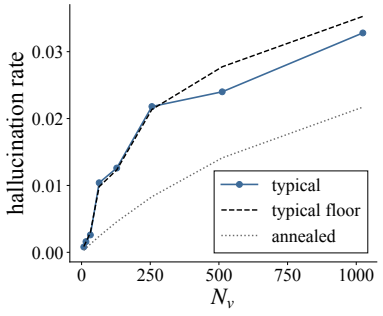
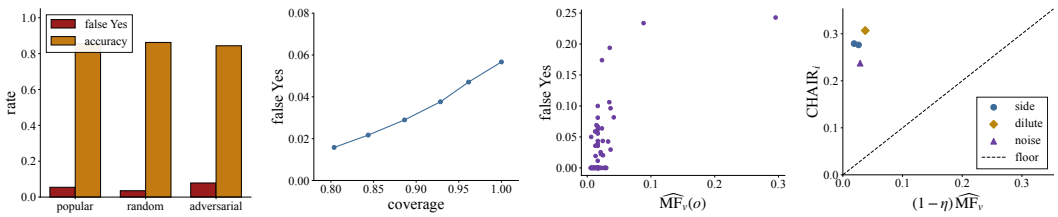


Figure 3: Median $g_\star$ tracks the typical floor above the annealed substitution.



(a) Official POPE. (b) Error penalty. (c) Monofact axis. (d) Caption floor.

Figure 4: Frozen Qwen3-VL-8B Instruct on official POPE and CHAIR style captions. Panel (b) walks the Pareto line and panel (d) plots CHAIR against the attention estimated floor.

6.2 Frozen Vision-Language Models

A frozen decoder follows the scoring line and the two calibration split. On official POPE with Qwen3-VL-8B Instruct, the popular, random, and adversarial splits rank as in Table 1 and Figure 4a, with false Yes 0.055, 0.035, and 0.079. The ranking matches the regularity prediction of Assumption 1 and Theorem 6. Binned ECE on the 8910 items is 0.113. Raising an exogenous error penalty from $c = 1$ to $c = 49$ lowers coverage from 1.00 to 0.80 and false Yes from 0.057 to 0.016, as shown in Figure 4b. This is the Pareto walk of Corollary 2 by changing only the decision threshold. Across 73 objects, false Yes on negatives correlates 0.64 with a caption estimated $\widehat{\mathrm{MF}}_v(o)$, as shown in Figure 4c, so the order of Theorem 1 is visible even when the absolute floor is a proxy.

Table 1: Official POPE on Qwen3-VL-8B Instruct.

split	n	acc.	F1	false Yes
popular	3000	0.856	0.841	0.055
random	2910	0.862	0.851	0.035
adversarial	3000	0.844	0.831	0.079

The same frozen 8B model, asked for a one sentence description of 200 COCO validation images, yields mention level CHAIR 0.276 at side 224 and 0.279 at side 448, as shown in Figure 4d. This is several times the POPE false Yes, as Theorem 8 predicts. The CHAIR points lie above the proxy floor $(1-\eta)\widehat{\mathrm{MF}}_v$ computed from last layer concentration and a $n = 2000$ Good Turing estimate, and no bucket falls below that floor. Changing input resolution from 224 to 448 nearly doubles the visual token count but leaves CHAIR unchanged, so dilution comes from unrelated tokens, not pixel count. Concatenating four independent distractor images and prompting the model to describe only the first image lowers needle mass from 0.59 to 0.42 and raises CHAIR from 0.276 to 0.307. An attention threshold gate on single image concentration reduces coverage sharply while barely moving CHAIR, the proxy bias of Lemma 3, so identifiable dilution uses extra images.

Across the frozen Instruct family, the same axes appear as qualitative comparisons rather than fitted slopes. Existence questions with an image remain near zero false Yes at 2B, 4B, and 8B, while the matched blind text slice rises to 0.75, 0.25, and 0.42. The imaged condition is the conservative vertex of Theorem 7, and the blind condition is the $\eta = 0$ slice of Theorem 1 where guessing returns. CLIP blind pairs are empirical $G = 0$ events (Tong et al., 2024), and blind language models that ignore the image are the $\eta = 0$ slice (Lin et al., 2024). Language prior takeover is that slice under a long tailed π (Leng et al., 2024; Fang et al., 2026). POPE ranking and Yes incentives are Assumption 1 and Theorem 6 at $c = 0$ (Li et al., 2023). Contrastive decoding and over trust filters occupy interior points of the coverage line. Figure 4 matches the POPE ranking and adds the generative counterpart of Theorem 8, with a caption floor several times the existence false Yes.

7 Conclusion

Closed domain visual hallucination is statistically necessary for a calibrated vision language model. The core bound factors the unread probability from the monofact rate, interpolating the text floor and the closed domain caveat, with a matching generator and a Visual *Is It Valid* companion certifying tightness. A softmax lemma converts visual token count into retention, yielding a see more invent more prediction and a typical versus annealed refinement. Coverage and hallucination trace a linear Pareto frontier, with binary POPE and instruction tuned policies at opposite vertices, and a formal separation between forecast and generative calibration that identifies CHAIR as the quantity governed by the floor. The identities are tight in a Zipf visual fact world and on frozen Qwen3-VL Instruct, with slopes matching the theory within one percent.

AI USE STATEMENT

In this work, we used generative AI tools for providing feedback on research methodology or experiments, and assisting with translation. We have not used generative AI tools for formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, implementing methods, supporting qualitative and thematic data analysis, interpreting results. Besides, generating synthetic data sets, helping develop theoretical models or conceptual frameworks, cleaning and reformatting dataset are not applicable to this work. Additionally, we used generative AI tools for modifying scientific figures or images, editing a research paper to improve readability. We have reviewed all AI assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Sourav Banerjee, Ayushi Agarwal, and Saloni Singla. Llms will always hallucinate, and we need to live with this. In *Intelligent Systems Conference*, pp. 624–648. Springer, 2025.

Nishant P Das and Piyush Srivastava. Innovation: An almost characterization of hallucination. In *Proceedings of the 43rd International Conference on Machine Learning*, pp. 1–15, 2026.

Hao Fang, Changle Zhou, Jiawei Kong, Kuofeng Gao, Bin Chen, Tao Liang, Guojun Ma, and Shu-Tao Xia. Grounding language with vision: A conditional mutual information calibrated decoding strategy for reducing hallucinations in lvlms. In *Advances in Neural Information Processing Systems*, volume 38, pp. 105233–105258, 2026.

Mehrdad Fazli, Bowen Wei, Ahmet Sari, and Ziwei Zhu. Caac: Confidence-aware attention calibration to reduce hallucinations in large vision-language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1242–1251. IEEE, 2026.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13418–13427. IEEE, 2024.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

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

Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. Evaluating large language models for accuracy incentivizes hallucinations. *Nature*, pp. 1–3, 2026.

Alkis Kalavasis, Anay Mehrotra, and Grigoris Velegkas. On the limits of language generation: Trade-offs between hallucination and mode-collapse. In *Proceedings of the 57th annual ACM Symposium on Theory of Computing*, pp. 1732–1743, 2025.

Jon Kleinberg and Sendhil Mullainathan. Language generation in the limit. In *Advances in Neural Information Processing Systems*, volume 37, pp. 66058–66079, 2024.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882. IEEE, 2024.

Jiaming Li, Jiacheng Zhang, Zequn Jie, Lin Ma, Ming Li, Xiaonan Luo, and Guanbin Li. Cross-modal attention calibration for lvlm hallucination mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40186–40196, 2026.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

Zhuohang Li, Chao Yan, Nicholas J Jackson, Wendi Cui, Bo Li, Jiaxin Zhang, and Bradley A Malin. Towards statistical factuality guarantee for large vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 11446–11467, 2025.

Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. Revisiting the role of language priors in vision-language models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 1–21, 2024.

Hude Liu, Jerry Yao-Chieh Hu, Jennifer Yuntong Zhang, Zhao Song, and Han Liu. Are hallucinations bad estimations? *arXiv preprint arXiv:2509.21473*, 2025.

Jiarui Liu, Weihao Xuan, Zhijing Jin, and Mona Diab. Taming object hallucinations with verified atomic confidence estimation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5430–5444, 2026.

David A McAllester and Robert E Schapire. On the convergence rate of good-turing estimators. In *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, pp. 1–6, 2000.

Seongheon Park and Sharon Li. Glsim: Detecting object hallucinations in lvlms via global-local similarity. In *Advances in Neural Information Processing Systems*, volume 38, pp. 37747–37776, 2026.

Jiaye Qian, Ge Zheng, Yuchen Zhu, and Sibei Yang. Intervene-all-paths: Unified mitigation of lvlm hallucinations across alignment formats. In *Advances in Neural Information Processing Systems*, volume 38, pp. 142168–142198, 2026.

Longfei Qiu, Jingqi Xiao, and Zhong Shao. Mechanized safety and liveness proofs for the mysticeti consensus protocol under the lido-dag framework. In *IEEE Symposium on Security and Privacy (SP)*, pp. 1842–1861. IEEE, 2026a.

Longfei Qiu, Jingqi Xiao, Ji-Yong Shin, and Zhong Shao. Suredistrib: Verifying almost-sure termination of composite asynchronous byzantine protocols. *Proceedings of the ACM on Programming Languages*, 10(PLDI):1152–1177, 2026b.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 4035–4045, 2018.

Supratik Sarkar and Swagatam Das. Grounding the ungrounded: A spectral-graph framework for quantifying hallucinations in multimodal llms. *arXiv preprint arXiv:2508.19366*, 2025.

Hoigi Seo, Dong Un Kang, Hyunjin Cho, Joohoon Lee, and Se Young Chun. On epistemic uncertainty of visual tokens for object hallucinations in large vision-language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 130605–130660, 2026.

Yuanmin Tang, Jing Yu, Keke Gai, Gang Xiong, Gaopeng Gou, Meikang Qiu, and Qi Wu. Manipulation intention understanding for zero-shot composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9466–9474, 2026.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, et al. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

Weixing Wang, Zifeng Ding, Jindong Gu, Rui Cao, Christoph Meinel, Gerard de Melo, and Haojin Yang. Image token matters: Mitigating hallucination in discrete tokenizer-based large vision-language models via latent editing. In *Advances in Neural Information Processing Systems*, volume 38, pp. 111306–111335, 2026.

Wenyi Xiao, XU Xinchi, and Leilei Gan. Vl-calibration: Decoupled confidence calibration for large vision-language models reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 44791–44815, 2026.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*, 2024.

Weihao Xuan, Qingcheng Zeng, Heli Qi, Junjue Wang, and Naoto Yokoya. Seeing is believing, but how much? a comprehensive analysis of verbalized calibration in vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1408–1450, 2025.

Xubing Ye, Yukang Gan, Xiaoke Huang, Yixiao Ge, and Yansong Tang. Voco-llama: Towards vision compression with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29836–29846. IEEE, 2025.

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*, pp. 56969–56998, 2024.

Yinlin Zhu, Xunkai Li, Jishuo Jia, Miao Hu, Di Wu, and Meikang Qiu. Towards effective federated graph foundation model via mitigating knowledge entanglement. In *Advances in Neural Information Processing Systems*, volume 38, pp. 64599–64628, 2026.

APPENDIX

A Notation 14

B Deferred statements 14

C Preliminaries 15

D Proofs 16

E Fano, erasure, and the binary symmetric channel 22

F Additional identities 23

G Scoring and reliability 25

H Worked example 26

I Experimental protocol 26

J Zipf world identities 27

K Frozen decoder measurements 31

A NOTATION

Table 2 lists the symbols used in the theorems and in the experiments. The same symbols are used without further definition in the proofs, and a quantity estimated from a finite sample is written with the same letter as its population counterpart.

Table 2: Notation used in the theorems and the experiments.

Symbol	Meaning
$I, O(I)$	image and its object set
$s, y(s, I)$	visual statement and its truth value
$\perp$	abstention (IDK), scored $y = 1$
p, g	world factoid law and generator
$Z = \Phi(I, \xi)$	visual features actually observed
$G \sim \mathrm{Bern}(\eta)$	erasure gate (Assumption 2)
$\widehat{\mathrm{MF}}_v$	visual monofact (Good Turing) rate
$p(U)$	missing mass of unobserved factoids
F, H, U	true facts, hallucinations, and the unobserved pool
r, s	regularity and sparsity (Assumption 1)
$\mathrm{Mis}_G, \mathrm{ECE}_b$	generative miscalibration, and calibration error in b bins
δ	single threshold calibration gap
α, α_0	coverage, and erasure time coverage
N_v, M	visual token count, and annealed signal mass
m	evidence arity (object, attribute, or relation)
c	error penalty in the scoring rule

B DEFERRED STATEMENTS

The statements below are used in the main text and proved in Appendix D. Grounded calibration (C) and event algebra calibration (E) are included so that the main bound remains valid under these stronger definitions as well.

Definition 2 (Grounded and event algebra calibration)**.** Grounded calibration (C) requires $\mathbb{E}[y \mid \hat{p} = q, I] = q$ almost surely. Event algebra calibration (E) requires generative calibration only on the atoms of a visual event algebra, such as object presence or a finer algebra of attributes and relations.

Proposition 2 (Implications)**.** *Grounded calibration (C) implies average calibration (A). If $\hat{p} = g(\cdot \mid I)$, then generative calibration (G) and binned average calibration differ by at most the bin width. Average calibration does not imply grounded calibration. The constant forecast $\hat{p} \equiv \mathbb{E}\rho$ is average calibrated for $y \mid I \sim \mathrm{Bern}(\rho(I))$, and it fails (C) unless ρ is constant.*

Theorem 9 (Bayes error under erasure)**.** *Under Assumption 2, the Bayes error of Visual Is It Valid is $(1 - \eta)\varepsilon_0$, where ε_0 is the error of the same test with no visual input.*

Theorem 10 (Combinatorial dilution)**.** *If m independent pieces of evidence are required and retention uses an AND rule, then Theorem 1 holds with η replaced by η^m. Under an OR rule the effective retention is $1 - (1 - \eta)^m$.*

Theorem 11 (Evidence gating)**.** *If $\hat{\eta}$ recovers G and the policy abstains on $\{G = 0\}$, then $\alpha = \eta$ and $\mathrm{hallu} = 0$. Any decoder that answers on a set $\mathcal{A}$ measurable with respect to Z satisfies $\mathrm{hallu} \geq (1 - \eta_{\mathrm{ans}})\alpha_{\mathrm{ans}}\widehat{\mathrm{MF}}_v - \mathrm{Mis} - o(1)$, where $\eta_{\mathrm{ans}} = \mathbb{E}[\eta \mid \mathcal{A}]$.*

Theorem 12 (Within bin undetectability)**.** *If $\hat{p}$ is average calibrated, then every detector that is a function of $\hat{p}$ alone has equal true and false positive rates on each level set $\{\hat{p} = q\}$, and the within bin AUROC is $1/2$.*

Corollary 3 (Mixtures)**.** *Any mixture of generators that share η obeys equation 5.*

Proposition 3 (Data processing)**.** *If $Z' = \Psi(Z)$, then $\eta' \leq \eta$. Consequently, no post processing of Z lowers the floor of Theorem 1.*

Proposition 4 (Typical and annealed retention)**.** *If $\eta(I)$ is random and $\eta_{\mathrm{typ}} \leq \mathbb{E}\eta$, then the median image has floor at least $(1 - \mathbb{E}\eta)\widehat{\mathrm{MF}}_v$, which is the floor obtained by substituting the annealed retention.*

C PRELIMINARIES

C.1 GOOD TURING CONCENTRATION

Lemma 4 (McAllester and Schapire)**.** *For every distribution on a countable alphabet and every* $\delta \in (0, 1)$*,*

$$\left|p(U) - \widehat{\mathrm{MF}}_v\right| \;\le\; O\Big(\sqrt{\tfrac{\log(1/\delta)}{n}}\Big)$$

with probability at least $1 - \delta$ *over* n *independent and identically distributed draws (McAllester & Schapire, 2000).*

The hidden constant in Lemma 4 is absolute, and τ denotes the right hand side throughout the proofs. Every subsequent high probability statement is restricted to the event of that lemma. The bound holds for every factoid law and does not use Zipf structure, while a Zipf law is used only to obtain a useful range of $\widehat{\mathrm{MF}}_v$ in the experiments of Appendix F.1.

C.2 GENERATIVE COARSENING

Let $\mathcal{V}_b(g)$ be the adaptive partition of the statement simplex into at most b level sets of g, as in Kalai & Vempala (2024). Write $p^{\mathcal{V}_b(g)}$ for the coarsening of p that is uniform on each block and carries the p mass of that block. Generative miscalibration is the total variation

$$\mathrm{Mis}_G(g, p) := \left\|p^{\mathcal{V}_b(g)} - g\right\|_{\mathrm{TV}}.$$

If g is exactly a coarsening of p and b is large enough that $1/b \le \min_{s\in\mathrm{supp}(g)} g(s)$, then $\mathrm{Mis}_G = 0$. A finite b leaves an error of at most the bin width, and that error is absorbed in the $O(1/b)$ term of the average calibration variant.

Lemma 5 (Pinsker control of bin leakage)**.** *If* $\mathrm{Mis}_G = \mathrm{TV}(p^{\mathcal{V}_b(g)}, g)$*, then the mass that* g *places on* H *cannot fall short of* $p^{\mathcal{V}_b(g)}(H)$ *by more than* Mis_G*. For a coarsening, Assumption 1 gives* $p^{\mathcal{V}_b(g)}(H) \ge p(U) - re^{-s}$*. The missing true facts occupy at most an* re^{-s} *fraction of any unobserved bin.*

Lemma 5 is the comparison taken from the core lemma of Kalai and Vempala, in their Section 6, and used below. Total variation limits how far the generated mass on hallucinations can fall below the coarsened truth, and regularity limits how much true mass can sit in an unobserved bin.

Lemma 6 (Total variation of mixtures)**.** *For a mixture* $\eta P_1 + (1 - \eta)P_0$*,*

$$\mathrm{TV}\big(\eta P_1 + (1-\eta)P_0,\, \eta Q_1 + (1-\eta)Q_0\big) \le \eta\,\mathrm{TV}(P_1, Q_1) + (1-\eta)\,\mathrm{TV}(P_0, Q_0).$$

Applied to the visible and erasure components, Lemma 6 allows Theorem 1 to use a single miscalibration parameter Mis. A separate parameter on each slice is controlled by the same total variation, and the factor $1 - \eta$ already carries the weight of the erasure slice.

C.3 CROSS ENTROPY STATIONARITY

Let $g_s(x) \propto s \cdot g(x)$ on $\mathcal{A} = \{g > t\}$ and $g_s(x) \propto g(x)$ off $\mathcal{A}$, after renormalizing. Then

$$\frac{d}{ds}\mathbb{E}_{x\sim p}[-\log g_s(x)]\Big|_{s=1} = g(\mathcal{A}) - p(\mathcal{A}),$$

so the single threshold gap δ is the failure of stationarity for a one parameter rescaling of the mass on $\mathcal{A}$ (Kalai et al., 2026). The same derivative holds for pairs $x = (I, s)$ and for a generator $g(\cdot \mid I)$ that is renormalized on each image. Local optimization inside a class that is closed under this rescaling drives δ to zero. The resulting condition is weaker than expected calibration error, and it is the one used for the Visual *Is It Valid* algebra in Theorem 3.

D PROOFS

D.1 PROPOSITION 2

Grounded calibration implies average calibration by the law of total expectation. Indeed, $\mathbb{E}[y \mid \hat{p}] = \mathbb{E}\big[\mathbb{E}[y \mid \hat{p}, I] \;\big|\; \hat{p}\big] = \hat{p}$ whenever the inner conditional equals $\hat{p}$ almost surely. For the converse, take $y \mid I \sim \mathrm{Bern}(\rho(I))$ with ρ not constant, and set $\hat{p} \equiv \mathbb{E}\rho$. Then $\mathbb{E}[y \mid \hat{p}] = \mathbb{E}\rho = \hat{p}$, so average calibration holds, whereas $\mathbb{E}[y \mid \hat{p}, I] = \rho(I) \neq \hat{p}$ on a set of positive measure. A model that reports the dataset frequency of one fixed object on every image, and ignores Z, is an instance of this forecast.

If $\hat{p}(s \mid I) = g(s \mid I)$ and $\mathcal{V}_b(g)$ is the adaptive partition of Kalai & Vempala (2024), then each bin B satisfies

$$\big|g(s) - \mathbb{E}_p[\mathbf{1}_B]/|B|\big| \leq \mathrm{width}(B)$$

whenever g is a coarsening. In that case Mis_G is at most the bin width plus the forecast calibration error of the binned $\hat{p}$. Once b is large enough that $1/b \leq \min_{s \in \mathrm{supp}(g)} g(s)$, the two notions agree on the support of g.

Event algebra. Let $\mathcal{E}$ be the finite algebra generated by the events $\{o \in O(I)\}_{o \in \mathcal{O}}$. Coarsening is required only on the atoms of $\mathcal{E}$, and object hallucination is exactly the error on that algebra. Attribute and relation algebras are finer. Theorem 10 applies the same argument to those atoms, with evidence arity m.

D.2 LEMMA 1

Condition on the training sample $\mathbf{x}_{\mathrm{train}}$ and on $G = 0$. Assumption 2 then gives $Z \perp y$. Let $p^{(0)}$ and $g^{(0)}$ denote the conditional laws of the factoid given $G = 0$. Independence of G from (F, p) yields $p^{(0)} = p$, together with the same regularity constant r. For every $y \in U$,

$$\Pr\big[y \in F \mid \mathbf{x}_{\mathrm{train}}, G = 0\big] = \Pr\big[y \in F \mid \mathbf{x}_{\mathrm{train}}\big] \leq \frac{r}{|U|}\,\mathbb{E}\big[|F \cap U| \mid \mathbf{x}_{\mathrm{train}}\big].$$

Sparsity $|F| \leq e^{-s}|H|$ is unchanged on this slice, so Lemma 5 applies to the pair $(p^{(0)}, g^{(0)})$. A generatively calibrated $g^{(0)}$ therefore satisfies

$$g^{(0)}(H) \;\geq\; p(U) - \mathrm{Mis}_G^{(0)} - re^{-s}.$$

The left side equals $\mathbb{E}[\mathrm{hallu} \mid G = 0]$, and Lemma 4 replaces $p(U)$ by $\widehat{\mathrm{MF}}_v \pm \tau$.

It remains to control $\mathrm{Mis}_G^{(0)}$. By definition of the conditional laws, $\mathrm{TV}(p^{(0)}, g^{(0)}) \leq \mathrm{TV}(p, g)/\Pr(G = 0)$, and therefore

$$(1 - \eta)\mathrm{Mis}_G^{(0)} \;\leq\; \mathrm{Mis}_G$$

whenever Mis_G is a total variation distance, which covers definition (G). For average calibration the same comparison holds with expected calibration error in place of total variation, because that error is an L^1 distance between binwise means. Restricting to a subset of mass $1 - \eta$ inflates the distance by at most $1/(1 - \eta)$, and the prefactor $1 - \eta$ in Theorem 1 cancels the inflation. As $\eta \to 1$ the erasure mass vanishes, while the product $(1 - \eta)\mathrm{Mis}_G^{(0)}$ remains $O(\mathrm{Mis})$.

Training counts and the gate at test time. The Good Turing statistic is a property of the empirical occupancy of the training sample, and the gate G at test time is independent of that sample. Restricting attention to the erasure slice therefore leaves the set of training singletons unchanged. The rate $\widehat{\mathrm{MF}}_v$ is computed from training counts, and the gate enters the final bound only through the weight $1 - \eta$.

D.3 THEOREM 1

Decompose $\mathrm{hallu} = \eta\,\mathbb{E}[\mathrm{hallu} \mid G = 1] + (1 - \eta)\,\mathbb{E}[\mathrm{hallu} \mid G = 0]$. The first summand is nonnegative, and Lemma 1 bounds the second. Multiplying by $1 - \eta$ and applying $(1 - \eta)\mathrm{Mis}_G^{(0)} \leq \mathrm{Mis}$ produces equation 5.

Average calibration. On a bin I_k of reported confidence, $\mathbb{E}[y \mid \hat{p} \in I_k] = \bar{q}_k \pm \mathrm{ECE}$. Under uniformity on U the missing mass bin has mean $\bar{q}_k = O(1/|U|)$, so the fraction of false statements in that bin is $1 - O(e^{-s})$. On the erasure slice the generator places mass $\widehat{\mathrm{MF}}_v \pm \tau$ on this bin. Subtracting the calibration error yields the average calibration form of the bound.

Event algebra. Replace $\mathcal{S}$ by the finite algebra generated by $\{o \in O(I)\}_{o \in \mathcal{O}}$, and require coarsening only on the atoms. The Kalai and Vempala lemma concerns a finite set of factoids and applies verbatim to those atoms. Object hallucination is exactly the error on this algebra.

Grounded calibration. Grounded calibration implies average calibration, so equation 5 holds. On the erasure slice, grounded calibration with respect to $\sigma(Z)$ forces $\hat{p} = \pi$, which is average calibration on that slice. Grounded calibration with respect to $\sigma(I)$ is a different statement, because it concerns an observer who sees the image itself.

Visual Is It Valid. For arbitrary facts, Theorem 3 of Kalai et al. (2026) gives $\varepsilon_0 \geq \frac{1}{2}\widehat{\mathrm{MF}}_v - O(n^{-1/2})$ after prompts and abstention are included. Theorem 9 multiplies this error by $1 - \eta$. Theorem 3 doubles the result and subtracts the sparsity ratio $|\mathcal{V}|/|\mathcal{E}| \leq e^{-s}$ together with the threshold gap δ. The resulting order matches equation 5. Generative coarsening and Visual *Is It Valid* therefore give the same floor, which is the robustness property stated in the introduction. The floor does not depend on a single calibration definition.

D.4 Theorem 3

Let D be the balanced mix of p and $\mathrm{Unif}(\mathcal{E}_c)$, and let $\hat{f} = +$ if and only if $g(s \mid I) > t$, where $t = 1/\min_c |\mathcal{E}_c|$. Then

$$\mathrm{err}_{\mathrm{iiv}} = \tfrac{1}{2} \Pr_{p}(\hat{f} = -) + \tfrac{1}{2} \Pr_{\mathcal{E}}(\hat{f} = +).$$

Let $\mathcal{A} = \{g > t\}$ and $\delta = |g(\mathcal{A}) - p(\mathcal{A})|$. The first term satisfies

$$\tfrac{1}{2} p(\mathcal{A}^c) = \tfrac{1}{2}\big(1 - p(\mathcal{A})\big) \leq \tfrac{1}{2}\big(1 - g(\mathcal{A}) + \delta\big).$$

For the second term, $|\mathcal{A} \cap \mathcal{E}|/|\mathcal{E}|$ counts the uniform errors above the threshold. Because $t = 1/\min|\mathcal{E}_c|$ and training is free of labeling error, the mass $g(\mathcal{E})$ is hallucinated and satisfies $g(\mathcal{E}) \leq \mathrm{hallu}$. The counting argument of Kalai et al. (2026), which does not use Z inside the algebra, then gives

$$\Pr_{\mathcal{E}}(\hat{f} = +) \leq \mathrm{hallu} + \frac{\max|\mathcal{V}_c|}{\min|\mathcal{E}_c|}.$$

Rearrangement is Theorem 3. The map from g to $\hat{f}$ may use Z. Any use of Z that reduces $\mathrm{err}_{\mathrm{iiv}}$ loosens the lower bound on hallu, as stated in the main text. A useful visual channel is allowed to help, and the theorem records how much help is required before the floor vanishes.

Prompts and abstention. If the generator may emit $\perp$, the Visual *Is It Valid* mix places $\perp$ in neither $\mathcal{V}_c$ nor $\mathcal{E}_c$. Abstention therefore contributes to neither error term. The cost appears instead as a loss of coverage, which is the tradeoff in Theorem 5.

D.5 Theorem 9

On $\{G = 1\}$ the label y is a function of Z, so the Bayes error is 0. On $\{G = 0\}$ the feature Z is independent of y, so the Bayes error equals the error ε_0 of the same test with no visual feature. The mixture of the two slices is $(1 - \eta)\varepsilon_0$.

For balanced POPE one has $\Pr(y = 1) = 1/2$ and therefore $\varepsilon_0 = 1/2$. For arbitrary facts with a uniform label on the unobserved pool, a random guess on U is wrong except on a fraction e^{-s} of the pool. The Visual *Is It Valid* mix places half of its mass on uniform negatives, which gives $\varepsilon_0 \geq \frac{1}{2} - O(e^{-s})$. The link from ε_0 to $\widehat{\mathrm{MF}}_v$ is Theorem 3 of Kalai et al. (2026), restricted to the erasure slice as in Lemma 1.

D.6 THEOREM 2

On $G = 1$, emit $s \sim \mathrm{Unif}(F \cap \mathrm{supp}(p))$, or emit a true caption. Then $g(H \mid G = 1) = 0$. On $G = 0$, let u be uniform on the unobserved pool $\hat{U}$ of size $|\hat{U}|$, place mass $\widehat{\mathrm{MF}}_v$ on u, and place mass $1 - \widehat{\mathrm{MF}}_v$ on the empirical distribution of observed facts, which is the two bin coarsening of p. Then

$$g(H \mid G = 0) = u(H \cap \hat{U}) = \frac{|H \cap \hat{U}|}{|\hat{U}|}, \qquad g(F \cap U \mid G = 0) = \frac{|F \cap U|}{|\hat{U}|} \le re^{-s}.$$

The uniform law on unobserved factoids is calibrated to p up to the Good Turing error τ and the sparsity term, so $\mathrm{Mis}_G = O(\tau + e^{-s})$. Taking the expectation over G yields $\mathrm{hallu}(g_\star) \le (1 - \eta)\widehat{\mathrm{MF}}_v + O(\tau + e^{-s})$.

The construction runs the Kalai and Vempala upper bound generator on the erasure slice and mixes it with a perfect visible branch. It shows that the order is attained by some calibrated generator. Together with Theorem 1, no calibrated generator can improve on that order. The construction is not proposed as a description of a deployed model.

D.7 COROLLARY 3

Hallucination $\mathrm{hallu}(g) = \mathbb{E}_{s\sim g}[\mathbf{1}\{y = 0\}]$ is linear in g. If $g = \sum_i \lambda_i g_i$ with $\lambda_i \ge 0$ and $\sum_i \lambda_i = 1$, then $\mathrm{hallu}(g) = \sum_i \lambda_i \mathrm{hallu}(g_i)$. Each summand is at least the right hand side of equation 5 at the shared retention η, and the mixture inherits the same lower bound. Under total variation, the miscalibration of a mixture is at most the mixture of the miscalibrations, so averaging generators does not improve the Mis term.

D.8 PROPOSITION 3

The chain $y - I - Z - Z'$ is Markov, so the data processing inequality gives $I(Z'; y) \le I(Z; y)$. On the erasure channel one has $I(Z; y) = \eta H(y)$ and $I(Z'; y) = \eta' H(y)$, and therefore $\eta' \le \eta$. Theorem 1 applied to the coarser observation Z' is the claim. Any decoder, merge, or contrastive feature that is a function of Z, possibly randomized, is an observation of this form.

D.9 PROPOSITION 4

Theorem 1 holds conditionally on the image I, with retention $\eta(I)$. The median of $1 - \eta(I)$ is $1 - \eta_{\mathrm{typ}}$. Whenever $\eta_{\mathrm{typ}} \le \eta_{\mathrm{ann}}$, one has $1 - \eta_{\mathrm{typ}} \ge 1 - \eta_{\mathrm{ann}}$, so the floor on the median image is at least the floor obtained by substituting the annealed mean. The map $\eta \mapsto 1 - \eta$ is linear, so Jensen's inequality contributes no further gap. The entire gap is the gap between the median and the mean of $\eta(I)$, and this gap is nonnegative when $\hat{A}$ is right skewed, as it is for lognormal signal mass.

D.10 LEMMA 2 AND THEOREM 4

Let $A = \sum_{i=1}^{m} e^{\Delta + \sigma_S \xi_i}$ and $B = \sum_{j=1}^{N_v} e^{\sigma_V \zeta_j}$. The lognormal moments are $\mathbb{E}e^{\sigma_V \zeta} = e^{\sigma_V^2/2}$ and $\mathbb{E}e^{2\sigma_V \zeta} = e^{2\sigma_V^2}$, so

$$\mathbb{E}[B/N_v] = e^{\sigma_V^2/2}, \qquad \mathrm{Var}(B/N_v) = N_v^{-1} e^{\sigma_V^2}(e^{\sigma_V^2} - 1).$$

Chebyshev's inequality gives $B = N_v e^{\sigma_V^2/2}(1 + \rho_n)$ with $|\rho_n| \le \sqrt{(e^{\sigma_V^2} - 1)/(N_v \delta)}\, e^{-\sigma_V^2/2}$, with probability at least $1 - \delta$. Set $\hat{A} = Ae^{-\sigma_V^2/2}$. Then

$$\eta = \frac{\hat{A}}{\hat{A} + N_v(1 + \rho_n)}, \qquad \eta - \frac{\hat{A}}{\hat{A} + N_v} = -\frac{\hat{A} N_v \rho_n}{(\hat{A} + N_v)(\hat{A} + N_v(1 + \rho_n))}.$$

For $|\rho_n| \le 1/2$ the absolute deviation is at most $2|\rho_n|\, N_v/(\hat{A} + N_v)$, and the ratio form is $1 + O_p(N_v^{-1/2})$. When $\sigma_S^2 \ll 2 \log m$, the average $\bar{W} = m^{-1} \sum_i e^{\sigma_S \xi_i}$ concentrates, and $\hat{A}$ converges in probability to $M := m \exp(\Delta + \frac{1}{2}(\sigma_S^2 - \sigma_V^2))$. This is the annealed law $\eta \approx M/(M + N_v)$.

Sub Gaussian tails. If ζ_j is sub Gaussian with variance proxy σ^2, the bound $\mathbb{E}e^{\lambda\zeta} \le e^{\lambda^2\sigma^2/2}$ still yields a finite logarithmic moment generating function in a neighborhood of σ_V, and exact Gaussian tails are not required. Bernstein's inequality then replaces Chebyshev's inequality, and $|\rho_n| = O_p(\sqrt{\log(1/\delta)/N_v})$. The annealed shape $N_v/(M+N_v)$ is unchanged, and only the rate of concentration changes. Logits with infinite variance would replace Chebyshev's inequality by a slower rate. The monotone dependence of $1-\eta$ on N_v remains as long as $\mathbb{E}e^{\sigma_V\zeta} < \infty$.

Retrieval and the mass budget. Under Gaussian tails, the top one retrieval event $\{\max_i s_i > \max_j \zeta_j\}$ has probability of order $\Phi_{\mathrm{N}}(\Delta/\sigma - \sqrt{2\log N_v})$, where Φ_{N} is the standard normal distribution function. This probability collapses at $N_v^{\mathrm{top1}} = \exp(\Delta^2/(2\sigma^2))$. Typically $\log N_v^{\mathrm{top1}} \gg \log M$, so there is a window in which a needle query, such as POPE existence for one named object, still succeeds, while free form coverage of many objects, as in CHAIR, has already risen with the mass budget floor. The two thresholds are different functionals of the same logit model, and both statements can hold together.

Theorem 4. Substitute $1-\eta = N_v/(M+N_v)+o_p(1)$ into Theorem 1. The map $N_v \mapsto N_v/(M+N_v)$ has derivative $M/(M+N_v)^2 > 0$, so the floor is strictly increasing in N_v. It equals $\widehat{\mathrm{MF}}_v/2$ at $N_v = M$, and it tends to $\widehat{\mathrm{MF}}_v$ as $N_v \to \infty$.

D.11 Theorem 5

Restrict Lemma 1 to the event $\{G = 0, s \neq \perp\}$, which has probability $(1-\eta)\alpha_0$. Conditional hallucination on that event is at least $\widehat{\mathrm{MF}}_v - \mathrm{Mis} - o(1)$. The identity $\alpha = \eta + (1-\eta)\alpha_0$ is the law of total probability for a policy that always answers when the image is visible. Solving for α_0 produces the factor $\max(0, \alpha-\eta)\widehat{\mathrm{MF}}_v$.

The three vertices in the main text are $\alpha_0 = 0$, $\alpha_0 = 1$, and the joint requirement $\mathrm{hallu} = 0$ with $\alpha = 1$. The third vertex forces miscalibration, since otherwise equation 9 would fail. Abstention rate $1-\alpha$ plays the role of a rate, hallucination plays the role of a distortion, and calibration restricts the feasible set. The Pareto line is the boundary of that set under erasure. ConfLVLM occupies the honest gating vertex by filtering (Li et al., 2025), and POPE occupies the full coverage vertex through the scoring rule of Theorem 6.

D.12 Theorem 10

Let $G_{\mathrm{eff}} = \prod_{j=1}^{m} G_j$ with G_j drawn i.i.d. from $\mathrm{Bern}(\eta)$. Then $G_{\mathrm{eff}} \sim \mathrm{Bern}(\eta^m)$, and G_{eff} is independent of y when the revealability assumption is applied to each piece of evidence. Theorem 1 with η replaced by η^m is the claim, and the numerical retention gap at $\eta = 1/2$ equals $1-\eta^m$.

The OR rule. The gate $G_{\mathrm{eff}} = 1 - \prod_j (1-G_j)$ has mean $1-(1-\eta)^m$, so the floor decreases as m grows. Redundant captions, in which the same object is named twice, are closer to an OR rule. AMBER relations, in which two distinct objects must both be in view, are closer to an AND rule (Wang et al., 2023).

Block correlation. If the pieces of evidence arrive in fully correlated blocks of size B, for instance adjacent patches of one object, then the AND of m objects uses m block representatives. The same power η^m applies, with η computed from $N_v^{\mathrm{eff}} = N_v/B$ in Lemma 2. Spatial redundancy therefore changes the signal mass M, while the algebraic role of m stays the same. Partially correlated gates lie between the two extremes, and the effective arity is the number of independent blocks.

Inflation of $\widehat{\mathrm{MF}}_v^{(m)}$. Relation tuples live in a product space of size $|\mathcal{O}|^m$. A Zipf law on that product is more peaked, so the monofact rate $\widehat{\mathrm{MF}}_v^{(m)}$ typically increases with m even at fixed η. Combinatorial dilution of η and inflation of $\widehat{\mathrm{MF}}_v$ therefore move the floor in the same direction.

D.13 THEOREM 6

Let $u(\text{Yes}) = (1+c)q - c$, $u(\text{No}) = 1 - (1+c)q$, and $u(\text{IDK}) = 0$. Comparing the three actions, Yes beats IDK when $q \geq c/(1+c)$, No beats IDK when $q \leq 1/(1+c)$, and Yes beats No when $q \geq 1/2$. For $c = 0$ one has $u(\text{Yes}) = q \geq 0$ and $u(\text{No}) = 1 - q \geq 0$, so IDK is never strictly better. Strict propriety of the Brier score $S(q, y) = -(q - y)^2$ and of the log score $S = \log q_y$ is given by Gneiting & Raftery (2007). A zero one score with abstention payoff 0 is proper for the hard decision only after q is already known, and it does not elicit q. On $\{G = 0\}$, Assumption 2 forces $q = \pi$, and $c = 0$ selects $\arg\max\{\pi, 1 - \pi\}$.

Incentive gap. Relative to honest abstention, the expected score gain on the erasure slice is $\mathbb{E}[\max(\pi, 1 - \pi)]$. For a Zipf prior with $\beta = 1.3$ and $K = 400$, as in the Zipf world, this gap equals 0.93. Almost every queried object has $\pi \neq 1/2$. The penalty c must therefore rise past a threshold that depends on the prior before coverage moves in the experiments of Section 6.

D.14 THEOREM 12

Conditionally on $\hat{p} = q$, average calibration says that $y \sim \text{Bern}(q)$. A detector $d(\hat{p})$ is a function of q alone, possibly randomized, and is therefore independent of y given q. Hence $\Pr(d = 1 \mid y = 0, q) = \Pr(d = 1 \mid y = 1, q) = d(q)$. Integrating over q can push the AUROC above $1/2$ only through the marginal laws of $\hat{p}$ on $\{y = 0\}$ and on $\{y = 1\}$.

If those marginals differ, for instance because hallucinations occupy a lower confidence bin, then the between bin AUROC can be large. This is compatible with undetectability inside a bin. A probe that only rereads a calibrated $\hat{p}$ therefore cannot separate true statements from false statements that the model has placed in the same bin (Kadavath et al., 2022). Detection needs information outside Definition 2, such as the attention retention η, a perturbation Kullback Leibler divergence (Leng et al., 2024), or semantic entropy over several samples.

Binned calibration. Replace the exact conditional $\mathbb{E}[y \mid \hat{p} = q] = q$ by its binned version with error ECE_b. The true and false positive rates then satisfy $|\text{TPR}(I_k) - \text{FPR}(I_k)| \leq \text{ECE}_b / \Pr(\hat{p} \in I_k)$ on each bin, so the within bin AUROC is $1/2 + O(\text{ECE}_b)$. The experiment in Section 6 uses $\text{ECE} = 0$ and records an AUROC of exactly $1/2$.

D.15 THEOREM 11

If $\hat{\eta}$ recovers G and the policy abstains on $\{G = 0\}$, then $\alpha = \eta$ and $\text{hallu} = 0$, which is the first vertex of Theorem 5. For a soft retention $\eta(I)$, abstaining on $\{\eta < \eta_\star\}$ gives

$$\text{hallu} \leq \mathbb{E}\big[(1 - \eta)\mathbf{1}_{\eta \geq \eta_\star}\big]\widehat{\text{MF}}_v + O(\text{Mis} + \tau)$$

by applying Theorem 2 conditionally on the image. Any decoder that answers on a set $\mathcal{A}$ measurable with respect to Z has $\eta_{\text{ans}} = \mathbb{E}[\eta \mid \mathcal{A}]$ and $\alpha_{\text{ans}} = \Pr(\mathcal{A})$. Theorem 5 on that subset is the stated lower bound.

Data processing. Let η_{Bayes} be the retention of a Bayes decoder that observes Z. A contrastive or reweighted decoder is a function of Z, possibly randomized, so its internal state Z' satisfies $I(Z'; y) \leq I(Z; y)$. Its erasure parameter therefore cannot exceed η_{Bayes}. Evidence aware decoding cannot drive the floor to zero unless the decoder abstains or receives new information about I.

A noisy gate. If $\Pr(\hat{G} \neq G) = \epsilon$, then at most an ϵ fraction of erased samples is answered, which contributes at most ϵ to hallucination. The contribution is at most $\epsilon \cdot O(\widehat{\text{MF}}_v)$ when the mistaken answers still follow $g_\star$. To first order, $\text{hallu} = O(\epsilon\widehat{\text{MF}}_v)$. A gate that is independent of G, such as a constant threshold on an uninformative statistic, recovers a full coverage policy and the original floor.

D.16 THEOREM 8

Construct a forecast with $\hat{p}(o \mid Z) = y$ on $\{G = 1\}$ and $\hat{p}(o \mid Z) = \pi(o)$ on $\{G = 0\}$. Then $\mathbb{E}[y \mid \hat{p} = q] = q$ almost surely. The visible bins have $q \in \{0, 1\}$ with $y = q$, and the erasure bin

has $q = \pi(o)$ with $y \sim \mathrm{Bern}(\pi(o))$ independent of Z. Hence the expected calibration error is 0. Thresholding at $t = 1 - \varepsilon$ yields

$$\Pr(\hat{p} \geq t,\, y = 0) = (1 - \eta)\Pr\big(\pi(o) \geq t,\, y = 0\big) \leq (1 - \eta)\Pr(\pi \geq t).$$

Under a Zipf prior the right side tends to 0 as $\varepsilon \to 0$. The same model may still sample caption objects from π on $\{G = 0\}$. That occupancy measure places missing mass on unobserved false objects, so

$$\big\|p^{\mathcal{V}_b(g)} - g\big\|_{\mathrm{TV}} \;\geq\; \tfrac{1}{2}\widehat{\mathrm{MF}}_v - O(\tau),$$

and generative calibration fails. This is the separation between the two definitions.

For the occupancy claim, draw a mention o uniformly from a caption S with $|S| \geq 1$, and let g_{occ} be the law of o. Apply Theorem 1 to g_{occ}. The event $\{o \notin O(I)\}$ is mention level CHAIR. A caption with $|S| = 0$ is an abstention and contributes nothing to that rate. If mentions are independent with common error $p_h = \mathrm{CHAIR}_i$ and expected length m_S, then

$$\mathrm{CHAIR}_s = 1 - \Pr(S \subseteq O(I)) \geq 1 - (1 - p_h)^{m_S}.$$

A bounded change of measure between mention sampling and sentence sampling is absorbed in the $O(\cdot)$ term of the occupancy bound.

D.17 Theorem 7

Let $u(\mathrm{Yes}) = (1 + c_{\mathrm{fp}})q - c_{\mathrm{fp}}$ and $u(\mathrm{No}) = (1 + c_{\mathrm{fn}})(1 - q) - c_{\mathrm{fn}}$, after removing an additive constant that does not affect the comparison. Then

$$u(\mathrm{Yes}) - u(\mathrm{No}) = (c_{\mathrm{fp}} + c_{\mathrm{fn}})q - c_{\mathrm{fp}},$$

so Yes beats No when $q > c_{\mathrm{fp}}/(c_{\mathrm{fp}} + c_{\mathrm{fn}})$. The symmetric case $c_{\mathrm{fp}} = c_{\mathrm{fn}} = c$ recovers the Yes versus No threshold $1/2$ from Theorem 6, while the Yes versus IDK threshold remains $c/(1 + c)$. Safety fine tuning that inflates c_{fp} relative to c_{fn} sends the Yes threshold to 1. On the erasure slice one has $q = \pi(o) < 1$ for almost every object, so $\alpha_0 = 0$ and the decoder occupies the gating vertex. An exogenous symmetric penalty c, written into the decision rule, moves the same decoder along the Pareto line without a weight update.

D.18 Lemma 3

The following comparisons delimit what an attention snapshot can certify. On a single image every visual token belongs to that image, so a span statistic that asks whether the needle image is attended is identically one. If y is independent of texture, or if (I_0, I_1) is a CLIP blind pair with $\Phi(I_0) \approx \Phi(I_1)$ and $y(I_0) \neq y(I_1)$ (Tong et al., 2024), attention may still concentrate on a peak, and η_{attn} can be large while $I(Z; y) = 0$. A top q concentration statistic is at least q under uniform attention, so the geometry of the simplex keeps it bounded away from zero even when the statistic carries no information about y.

Concatenating K independent distractor images adds K copies of the noise mass in the denominator of Lemma 2, which supplies an identifiable instrument for dilution. Masking the patches that carry y drives the signal mass toward zero, and CLIP blind pairs remove the coupling between A and y. If $\mathbb{E}[1 - \hat{\eta}] = (1 - \eta) - b$ with $b > 0$, then the reported product floor is biased downward by $b\,\widehat{\mathrm{MF}}_v$, and a small empirical value cannot be read as a violation of Theorem 1.

D.19 Proposition 1

If the channel is hard erasure, the posterior equals y on $\{G = 1\}$ and equals π on $\{G = 0\}$, so it takes values in $\{0, 1, \pi\}$. Conversely, set $G = \mathbf{1}\{P(y \mid Z) \in \{0, 1\}\}$. On $\{G = 0\}$ one has $P(y \mid Z) = \pi$, and therefore $Z \perp y$. Either description yields Theorem 1. For a binary symmetric channel, $\mathrm{logit}\, P(y \mid Z) = \mathrm{logit}(\pi) + \kappa(2Z - 1)$ with $\kappa \neq 0$ whenever the crossover is neither 0 nor $1/2$, so the posterior does not collapse to $\{0, 1, \pi\}$. Fano's inequality then supplies the lower bound $P_e \geq h_2^{-1}((1 - C)_+)$. This is the capacity relaxation derived in Appendix E, whereas the product $(1 - \eta)\widehat{\mathrm{MF}}_v$ is the identity for hard erasure.

E FANO, ERASURE, AND THE BINARY SYMMETRIC CHANNEL

Let $y \in \{0, 1\}$ be balanced, so that $H(y) = 1$, and consider three channels of capacity $C \in [0, 1]$. An erasure channel has capacity $C = \eta$ and Bayes error $P_e = (1 - \eta)/2 = (1 - C)/2$. Fano's inequality gives $H(y \mid Z) \le h_2(P_e)$ and $H(y \mid Z) \ge 1 - C$, so $P_e \ge h_2^{-1}((1 - C)_+)$. The map h_2^{-1} sends $[0, 1]$ onto $[0, 1/2]$ and is the increasing inverse of binary entropy on $[0, 1/2]$. A binary symmetric channel has capacity $C = 1 - h_2(\gamma)$, so the crossover equals $\gamma = h_2^{-1}(1 - C)$ and the Bayes error equals γ. Fano's bound lies strictly below the erasure curve except at the endpoints $C \in \{0, 1\}$, which is the usual slack of Fano's inequality on a binary alphabet. The binary symmetric channel meets the Fano inverse by the choice of γ. The main text uses erasure because the bound is tight and matches an attention hit, and because the factor $1 - \eta$ is then an identity in place of an inverse entropy transform. Figure 5 plots the three curves, obtained by inverting h_2 with 60 steps of bisection.

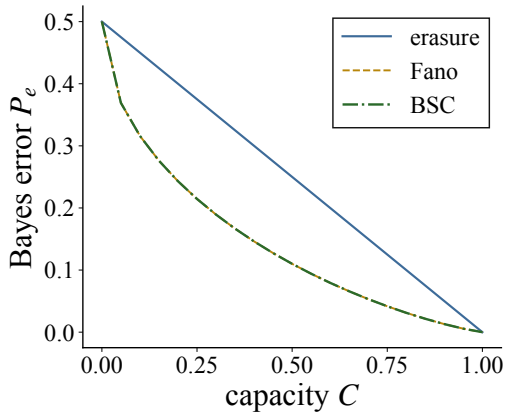


Figure 5: Bayes error versus capacity for erasure, Fano, and a binary symmetric channel. Erasure is the tight model in Theorem 1, and Fano is a general relaxation under a capacity constraint.

The same capacity axis can be read as a hallucination floor after the Bayes error is doubled. Figure 6 compares erasure with a binary symmetric channel of equal capacity under the Visual *Is It Valid* reduction. Under Assumption 3 alone, Theorem 3 together with binary Fano gives

$$\text{hallu} \;\ge\; 2\, h_2^{-1}\big((h_2(\rho) - C)_+\big) - \frac{|\mathcal{V}|}{|\mathcal{E}|} - \delta.$$

On the erasure channel one may identify C with $\eta H(y)$, and the display then has the same order as equation 5. The same qualitative conclusions hold for each of these channels. The floor is positive whenever $C < H(y)$, it vanishes under perfect vision, and it interpolates the open domain bound.

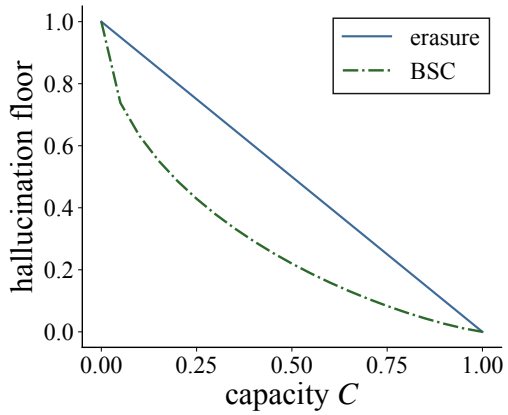


Figure 6: Visual *Is It Valid* hallucination floor under erasure and under a binary symmetric channel of equal capacity.

F ADDITIONAL IDENTITIES

F.1 ZIPF MONOFACTS

Let $p_k = k^{-\beta}/H_{K,\beta}$, where $H_{K,\beta}$ denotes $\sum_{k=1}^{K} k^{-\beta}$. For n independent draws,

$$\mathbb{E}[\widehat{\mathrm{MF}}_v] = n \sum_{k=1}^{K} p_k (1-p_k)^{n-1}. \tag{11}$$

The expected missing mass is $\mathbb{E}[p(U)] = \sum_k p_k(1-p_k)^n$. For $\beta \in (1,2)$, if $K \to \infty$ and then $n \to \infty$ with $n = o(K^{\beta-1})$, a constant fraction of species remains unseen. This is the classical Zipf missing mass regime, so $\widehat{\mathrm{MF}}_v$ does not vanish. For $\beta < 1$ the law is flat enough that large n exhausts the support and $\widehat{\mathrm{MF}}_v \to 0$, which matches the cell $\beta = 0.6$, $n = 6400$ in the experiment, where both $\widehat{\mathrm{MF}}_v$ and hallucination equal 0. The theorems do not take equation 11 as a modeling assumption. They are valid for every law that satisfies Assumption 1, and the display only explains why sweeping β and n traces a useful range of $\widehat{\mathrm{MF}}_v$.

For $\beta \in (1,2)$ and infinite support, the expected number of unseen species after n draws is of order $n^{1-1/\beta}$. The experiments use finite vocabularies $K \in \{300, 400\}$ and do not rely on that asymptotic. They compare $\widehat{\mathrm{MF}}_v$ with $p(U)$ by Monte Carlo, with mean absolute error 0.0095 at seed 0. Figure 7 shows the absolute Good Turing error against n.

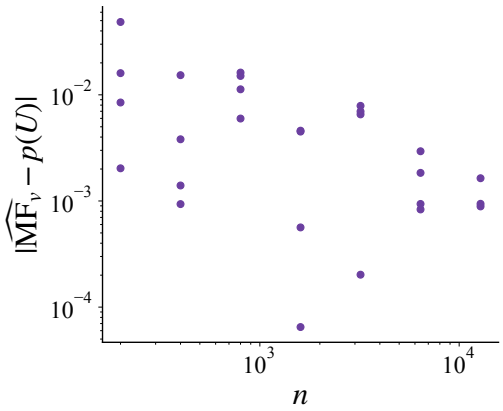


Figure 7: Absolute Good Turing error versus n, on a log log scale.

F.2 MIXTURE LOGITS AND LANGUAGE PRIOR DOMINANCE

Let $\ell(o) = \ell_{\mathrm{vis}}(o; Z) + \ell_{\mathrm{lang}}(o)$ at temperature T, and assume $\ell_{\mathrm{vis}} = 0$ on $\{G = 0\}$. Then $g(o \mid G = 0) = \pi_T(o)$. On $\{G = 1\}$, if the visual logit of the true object exceeds every other logit by Δ,

$$g(o^\star \mid G = 1) \;\geq\; 1 - (K-1)e^{-\Delta/T}.$$

The law of total probability gives

$$\mathrm{hallu} = (1-\eta)\,\mathbb{E}\big[1 - \pi(O(I))\big] + \eta \cdot O\big(Ke^{-\Delta/T}\big).$$

The second term is implementation error on the visible slice, while the statistical floor is the first term. As N_v grows, Lemma 2 drives η down and the first term dominates, so the language prior dominates the unread slice. Replacing $\mathbb{E}[1-\pi(O(I))]$ by $\widehat{\mathrm{MF}}_v$ under a long tailed prior, this mixture attains the order of Theorem 2. The construction is used only for the upper bound. The lower bound is proved without it. The same decomposition accounts for the empirical observation that vision language models rely heavily on the language prior (Leng et al., 2024; Fang et al., 2026). On the erasure slice of a calibrated generator that reliance is forced, and on the visible slice it remains an implementation error. Figure 8 compares emission of the prior mode on the unread slice with $g_\star$.

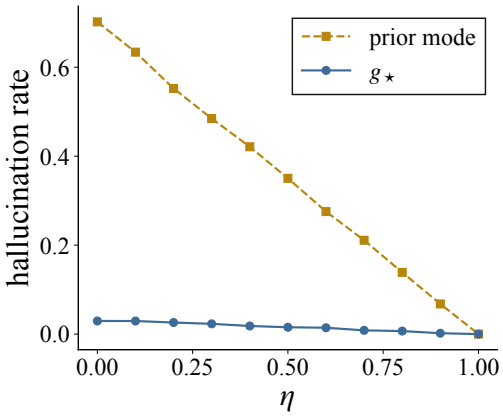


Figure 8: Emitting the language prior mode on the erasure slice lies well above $g_\star$.

F.3 Many facts per document

The one fact per document idealization can be dropped. Let each document x carry a finite set $S(x) \subset \mathcal{S}$ of visual statements, and let g be the law of an element drawn uniformly from $S(x)$, or the law of the whole set when hallucination means that at least one element is false. The latter choice only raises the rate. The factoid distribution p is the occupancy measure $p(s) = \mathbb{E}[\mathbf{1}\{s \in S(x)\}]/\mathbb{E}[|S(x)|]$, training counts are occupancy counts, and Good Turing still estimates the missing occupancy mass. Regularity is imposed on occupancy probabilities, and Lemma 1 applies to the occupancy law. The only change is notational, since $\widehat{\mathrm{MF}}_v$ is now computed on extracted statements. A union bound over $|S(x)| \le M_{\max}$ multiplies a set level rate by at most $M_{\max}$, and this factor is absorbed in the $O(\cdot)$ term when one reports hallucination per statement, which is the convention of CHAIR_i (Rohrbach et al., 2018). Sentence level CHAIR_s is a coarser functional and obeys the same lower bound up to that union factor. Figure 9 shows the occupancy monofact rate as the number of facts per document grows.

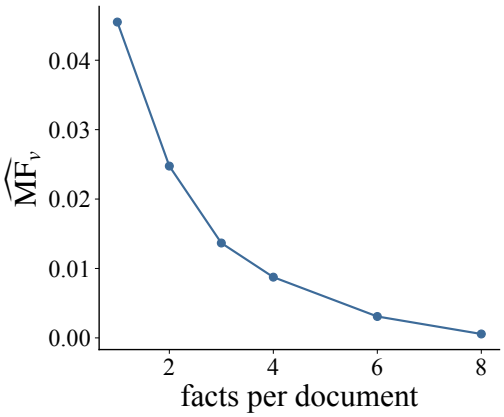


Figure 9: Occupancy monofact rate as the number of facts per document grows.

F.4 Innovation

Das & Srivastava (2026) define innovation as the mass that a generator places outside the training sample, and they prove that this mass characterizes Kalai and Vempala hallucination in text, up to lower order terms. A visual statement can lie outside the training sample and still be true of the current image. The innovation on the unread slice is

$$\mathrm{innov}_v := \Pr\big(s \notin \{\text{training factoids}\} \ \wedge \ G = 0\big).$$

Then $\mathrm{innov}_v \le (1-\eta)\,\mathrm{innov}_{\mathrm{text}}$, and the Das and Srivastava lower bound multiplied by $1-\eta$ recovers the order of Theorem 1. A novel object that is visible and true counts as innovation, and

it contributes to hallucination only on the unread slice. The factor $1 - \eta$ removes the mass of those visible novel objects. The split uses the visual channel, which a text argument does not have.

Equivalently, write $\mathrm{innov}_{\mathrm{text}} = \Pr(s \notin \mathrm{train})$ and decompose

$$\mathrm{innov}_{\mathrm{text}} = \Pr(s \notin \mathrm{train}, G = 1) + \Pr(s \notin \mathrm{train}, G = 0).$$

The first summand can be true of I. Only the second summand is a candidate hallucination when vision on the visible slice is free of error. Removing the first summand is again the factor $1 - \eta$.

G Scoring and reliability

G.1 Scoring rules

Table 3 lists scoring rules for a binary visual question that allows abstention. A strictly proper score elicits the report $\hat{p} = q$, and a later threshold, possibly with penalty $c > 0$, converts that report into a hard action. These are separate layers. The POPE convention $c = 0$ collapses them, because the hard action is scored, q is never elicited, and abstention is weakly dominated. An open rubric in the sense of Kalai et al. (2026) writes c into the prompt, so that the second layer follows the risk specified by the user, and no weight update is required.

Table 3: Scoring rules for a binary visual question with abstention. Only the Brier and log scores elicit q. Hard zero one loss with $c = 0$ is the POPE convention.

Rule	Payoffs (Yes, No, IDK)	Elicits q	Hard decision
$c = 0$ (POPE)	$\mathbf{1}_{y=1}, \mathbf{1}_{y=0}, 0$	no	always answer
weighted zero one	$1 - c\mathbf{1}_{y=0}, 1 - c\mathbf{1}_{y=1}, 0$	no	threshold $c/(1+c)$
Brier	$-(q-y)^2$ (report q)	yes	report q, then threshold
log	$\log q_y$	yes	report q, then threshold

CHAIR is the generative analogue of the same split. Emitting a false object is scored as an error, and saying nothing about an object is scored as a miss only when that object was required. In free form captioning the implicit penalty c is small, because a superfluous noun costs little, and the full coverage vertex is selected again.

G.2 Calibration of the matching generator

Factual confidence under $g_\star$ equals 1 on visible facts and on replayed training facts, and it equals 0 on uniform draws from the false unobserved pool. In eight equal bins this assignment has expected calibration error 0, as shown in Figure 10, and it is compatible with generative coarsening up to the Good Turing tolerance. The mass on the unobserved pool equals $\widehat{\mathrm{MF}}_v$, while the truth probability of a uniform draw from a pool of size much larger than $|F|$ is near 0. The overconfident baseline instead reports $\hat{p} = 0.95$ on every erased generation, which produces expected calibration error about $0.95(1 - \eta)$.

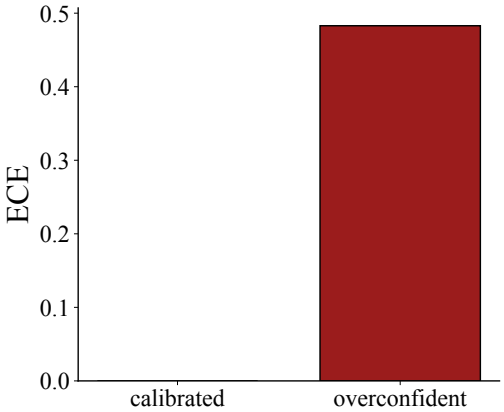


Figure 10: Eight bin expected calibration error of $g_\star$ and of the overconfident baseline at $\eta = 0.5$.

Figure 10 separates generative calibration from forecast calibration in one plot. Generative calibration forces mass onto unseen statements, while forecast calibration forces the reported probability of those statements to be small. Hallucination is the act of emitting the statements. Assigning them a small probability is a separate forecast. A model that assigns $\hat{p} = 0$ to a false statement and then emits it with positive mass is generatively uncalibrated whenever that mass exceeds what the coarsening of p would place there. The generator $g_\star$ emits the missing mass bin at rate $\widehat{\mathrm{MF}}_v$ and reports $\hat{p} = 0$ inside it, so both requirements hold.

H Worked example

The identities are easy to check on a vocabulary of size three. Let $\mathcal{O} = \{\mathrm{cat}, \mathrm{dog}, \mathrm{toaster}\}$ with $\pi = (1/2, 1/3, 1/6)$, and consider a world in which each image contains exactly one object. A training sample of size $n = 6$ yields counts $(3, 2, 1)$, so $\widehat{\mathrm{MF}}_v = 1/6$ because toaster is the unique singleton. The unobserved pool used at generation time is a large set of false strings, and sparsity holds with e^{-s} near 0.

At test time the image is a toaster, which appeared once in training. If $G = 1$, then $g_\star$ names toaster and $\mathrm{hallu} = 0$. If $G = 0$, then $g_\star$ places mass $1/6$ on the false pool and mass $5/6$ on the empirical frequencies $(3/6, 2/6, 1/6)$. A draw from the false pool is a hallucination. Replaying cat or dog on a toaster image is also false for that image, but those strings are true training facts, so they count as errors on observed mass. The Kalai and Vempala core lemma charges only the missing mass bin, and therefore $\mathbb{E}[\mathrm{hallu} \mid G = 0] \geq \widehat{\mathrm{MF}}_v - O(e^{-s}) = \frac{1}{6}$. Mixing with $G \sim \mathrm{Bern}(\eta)$ recovers $\mathrm{hallu} \geq (1 - \eta)/6$.

The same numbers show why multiplying a text bound by an indicator that an image was supplied does not yield a theorem. If the model observed I, toaster would be determined and the missing mass bin would not have to be emitted. The gate G carries the content of the reduction from the text theorem to the visual one. An overconfident policy that always names cat on $\{G = 0\}$ has hallucination $1 - \eta$ on toaster images, and its calibration error equals the reported confidence times $1 - \eta$. That policy lies well above the floor, and it is the analogue of a Yes bias on a high frequency object.

I Experimental protocol

The Zipf world draws object probabilities $p_k \propto k^{-\beta}$ on a vocabulary of size $K \in \{300, 400\}$. Training draws n fact identifiers independently and computes $\widehat{\mathrm{MF}}_v$ together with the missing mass $p(U)$. Unless a cell names another seed, the seed is 0. The false pool has size $2 \cdot 10^4$, so $|F|/|H| \leq K/(2 \cdot 10^4) \leq 0.02$ and $s \gtrsim 3.9$ in Assumption 1. Regularity is inherited from the Zipf law. On the cells we report, no unobserved atom exceeds a small multiple of the mean unobserved probability.

The matching generator $g_\star$ draws $G \sim \mathrm{Bern}(\eta)$, emits a true fact on $\{G = 1\}$, and on $\{G = 0\}$ answers with probability α_0. Of those answers, a fraction $\widehat{\mathrm{MF}}_v$ are uniform draws from the false pool and the rest are empirical replays. An overconfident baseline always hallucinates on $\{G = 0\}$ and reports confidence 0.95. Softmax retention uses Monte Carlo draws of Gaussian logits at a stated triple (m, Δ, σ). Hallucination is the fraction of test items with $y = 0$, abstention is scored as $y = 1$, and coverage is one minus the abstention rate. Expected calibration error uses eight bins of equal width. Within bin AUROC uses a constant $\hat{p}$ inside a mixed bin of size $2 \cdot 10^4$.

The no vision grid sweeps $\beta \in \{0.6, 0.9, 1.2, 1.6, 2.0\}$ and $n \in \{400, 800, 1600, 3200, 6400\}$ at $\eta = 0$, with 8000 test items. The product sweep fixes $\widehat{\mathrm{MF}}_v = 0.046$ and varies $\eta \in \{0, 0.1, \ldots, 1\}$, with 12000 test items. Softmax dilution uses $(m, \Delta, \sigma) = (4, 3, 1)$ and $N_v \in \{1, 2, 4, \ldots, 1024\}$. A coverage triangle fixes $\eta = 0.4$ and varies α_0. The AND sweep uses $m \in \{1, 2, 3, 4\}$ at $\eta = 0.5$. The scoring sweep uses $c \in \{0, 0.25, 0.5, 1, 2, 4, 9\}$ on 20000 binary items at $\eta = 0.35$. Gating compares $\alpha_0 = 0$ with $\alpha_0 = 1$, caption noise flips labels at rates in $[0, 0.5]$, and POPE style negatives use popular, random, and tail draws. A five seed band repeats the product sweep. Visual *Is It Valid* is checked on eight cells (β, n, η). Occupancy varies the number of facts per document from 1 to 8. Typical and annealed retention use $(m, \Delta, \sigma_S, \sigma_V) = (8, 3, 2.5, 1.5)$.

At seed 0, ordinary least squares of hallucination on $\widehat{\mathrm{MF}}_v$ through the origin has slope 0.979, and the mean absolute gap on the no vision grid is 0.0018. The product sweep has relative slope error $4.2 \cdot 10^{-3}$. The annealed mass is $M = 80.34$, and no calibrated point enters the forbidden region. Within bin AUROC is exactly $1/2$. The calibration error of $g_\star$ is 0, while the overconfident baseline has calibration error 0.57 at $\eta = 0.4$. The mean absolute Good Turing error is 0.0095, the five seed relative slope is 0.975 ± 0.018, and the gap between the typical and annealed floors is $7.3 \cdot 10^{-3}$. Cells with caption noise lie above the clean floor. These Zipf identities are recorded in Appendix J.

Frozen decoder measurements use Qwen3-VL Instruct checkpoints at 2B, 4B, and 8B, with weights held fixed (Bai et al., 2025). Official POPE questions on Common Objects in Context (COCO) val2014 contribute 8910 items after filtering. The first token logits for Yes, No, and Unsure implement the decision rule of Theorem 6. Last layer attention mass on the named object is recorded as a snapshot of η, subject to the bias in Lemma 3. One sentence captions on 200 COCO validation images are scored with a CHAIR style object lexicon against COCO instance labels. Input side lengths 224 and 448 vary the pixel count. Concatenating four independent distractor images, with a prompt that asks for a description of the first image only, is the identifiable dilution instrument. An attention threshold abstains when the snapshot falls below $\eta_\star$, and an exogenous error penalty c is applied to the same frozen logits without a weight update. The resulting measurements are recorded in Appendix K.

J ZIPF WORLD IDENTITIES

Table 4 records every cell of the no vision grid. Ordinary least squares of hallu on $\widehat{\mathrm{MF}}_v$ through the origin uses all twenty five points. The cell $\beta = 0.6$, $n = 6400$ is the exhausted support regime of Appendix F.1, and it contributes the origin of that regression.

Table 4: No vision Zipf world at seed 0. All twenty five cells enter the ordinary least squares slope in Section 6.

β	n	$\widehat{\mathrm{MF}}_v$	$\mathrm{hallu}(g_\star)$	$\lvert p(U) - \widehat{\mathrm{MF}}_v \rvert$
0.6	400	0.285	0.277	0.071
0.6	800	0.095	0.095	0.001
0.6	1600	0.019	0.018	0.003
0.6	3200	0.001	0.001	0.001
0.6	6400	0.000	0.000	0.000
0.9	400	0.205	0.196	0.010
0.9	800	0.096	0.099	0.027
0.9	1600	0.036	0.036	0.010
0.9	3200	0.008	0.007	0.005
0.9	6400	0.000	0.000	0.002
1.2	400	0.163	0.162	0.019
1.2	800	0.086	0.088	0.006
1.2	1600	0.046	0.042	0.005
1.2	3200	0.017	0.016	0.004
1.2	6400	0.007	0.004	0.003
1.6	400	0.075	0.076	0.003
1.6	800	0.051	0.054	0.003
1.6	1600	0.029	0.027	0.002
1.6	3200	0.026	0.024	0.008
1.6	6400	0.012	0.013	0.001
2.0	400	0.038	0.035	0.010
2.0	800	0.025	0.026	0.003
2.0	1600	0.013	0.013	0.003
2.0	3200	0.012	0.014	0.004
2.0	6400	0.007	0.009	0.001

Table 5 records the scoring sweep in the same world. Coverage remains 1 until $c = 2$, because the Yes threshold $c/(1 + c)$ has not yet entered the bulk of the prior. At $c = 4$ and at $c = 9$ the policy abstains on a set of positive mass, and both hallucination and coverage fall together.

Table 5: Scoring sweep in the Zipf world at $\eta = 0.35$ and seed 0.

c	$c/(1+c)$	coverage	hallu	IDK rate
0	0	1.000	0.076	0.000
0.25	0.200	1.000	0.076	0.000
0.50	0.333	1.000	0.076	0.000
1	0.500	1.000	0.076	0.000
2	0.667	1.000	0.076	0.000
4	0.800	0.805	0.017	0.195
9	0.900	0.729	0.008	0.271

The signed gap in Figure 11 stays inside ± 0.01 on a slice of (β, η). Figure 12 repeats the product sweep over five independent worlds and keeps the same slope, which shows that the identity is stable under both the Zipf parameters and the random seed.

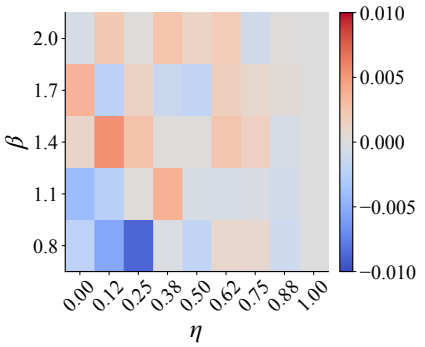


Figure 11: Signed gap $\text{hallu} - (1-\eta)\widehat{\text{MF}}_v$ on a (β, η) slice. The color scale covers ± 0.01.

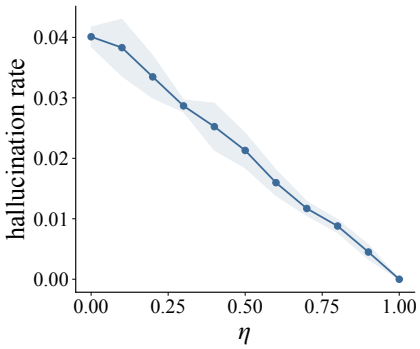


Figure 12: Mean and one standard deviation of $g_\star$ over five independent worlds.

Figure 13 follows the sweep of Table 5 and moves coverage down the Pareto line once $c/(1+c)$ enters the bulk of the prior. Figure 14 is the zero hallucination vertex of Theorem 11, obtained by abstaining whenever the image is unread. Full coverage $g_\star$ remains on the product floor, while the gated policy is identically free of hallucination.

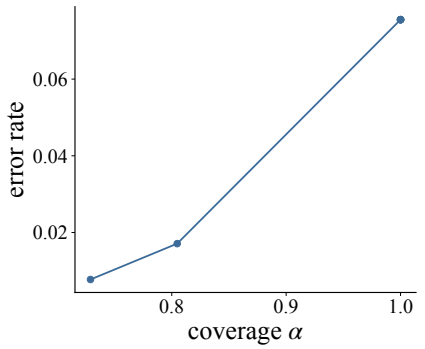


Figure 13: Raising c moves coverage down the Pareto line once the threshold enters the bulk of the prior.

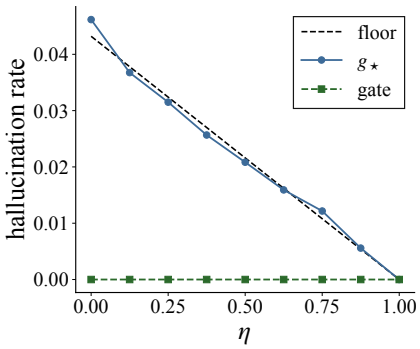


Figure 14: Full coverage $g_\star$ versus evidence gating. Gating has identically zero hallucination.

One Pareto line for each retention is shown in Figure 15. The dashed curves are $\max(0, \alpha - \eta)\widehat{\text{MF}}_v$. Calibrated generators lie on or above these curves, and the intercept moves with η exactly as Theorem 5 predicts.

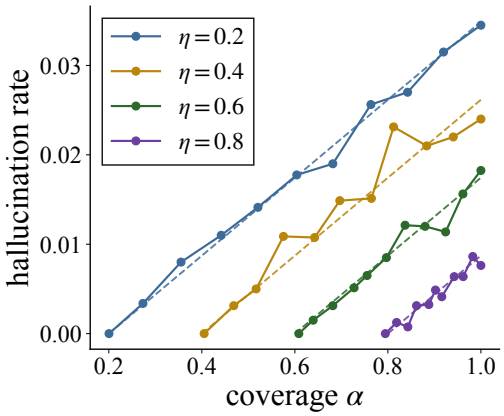


Figure 15: One Pareto line for each η. Dashed curves are $\max(0, \alpha - \eta)\widehat{\mathrm{MF}}_v$.

AND arity in Figure 16 raises the floor as $1-\eta^m$, and the OR rule reverses that increase, as predicted by Theorem 10. The isolated AND sweep in Figure 17 tracks $1 - \eta^m$ at the fixed retention $\eta = 0.5$. The two panels use the same monofact scale, so the change in the floor comes from the combination rule.

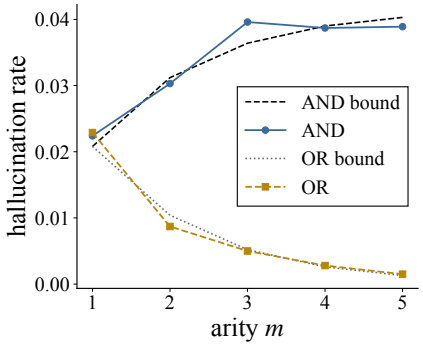


Figure 16: AND arity raises the floor as $1 - \eta^m$, and OR reverses the increase, as in Theorem 10.

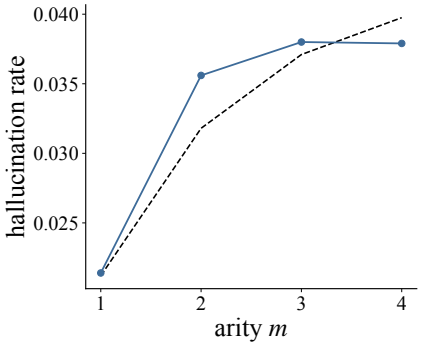


Figure 17: AND erasure at $\eta = 0.5$. The floor tracks $1 - \eta^m$.

A gate that flips G with probability ϵ traces the line $(1 - \eta)\epsilon\,\widehat{\mathrm{MF}}_v$ in Figure 18, which is the first order expansion in the proof of Theorem 11. Caption noise in Figure 19 raises hallucination above the clean floor and does not push any cell below it, which is the direction required of a lower bound under labeling error.

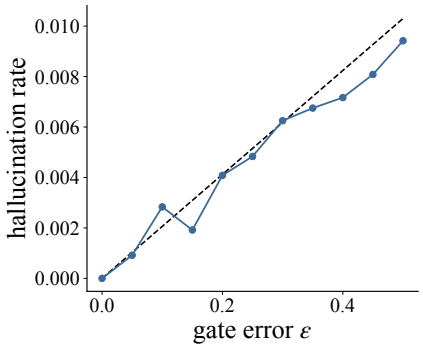


Figure 18: A gate that flips G with probability ϵ traces $(1 - \eta)\epsilon\,\widehat{\mathrm{MF}}_v$.

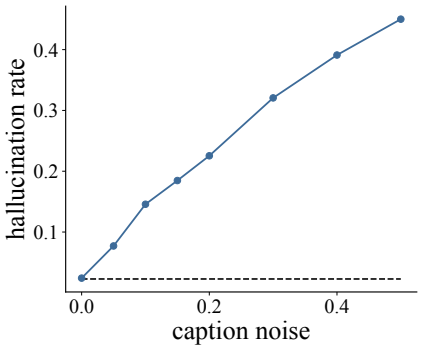


Figure 19: Caption noise raises hallucination above the clean floor.

POPE style negatives in Figure 20 rank popular objects above tail objects, because popular objects have larger prior mass π and therefore a higher floor on the erasure slice. Figure 21 checks Visual

Is It Valid on eight Zipf cells and compares hallucination with $2\,\mathrm{err}_{\mathrm{iiv}}$, as stated in Theorem 3. The cells track the predicted line once the sparsity correction is included.

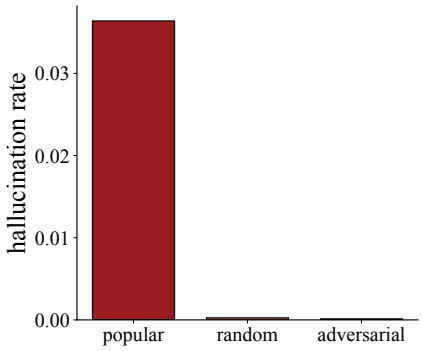


Figure 20: POPE style negatives. Popular objects have larger π and a higher erasure floor.

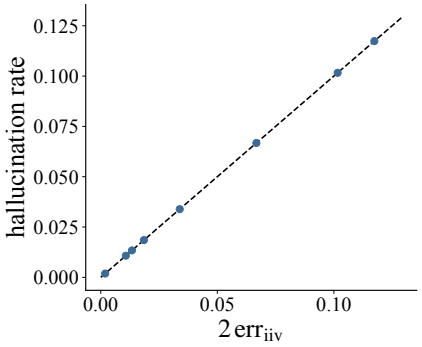


Figure 21: Hallucination versus $2\,\mathrm{err}_{\mathrm{iiv}}$ on eight Zipf cells, as in Theorem 3.

Raising retention from η to a larger value η' lowers the floor to $(1-\eta')\widehat{\mathrm{MF}}_v$, as shown in Figure 22. The floor reaches 0 only at $\eta' = 1$. Intermediate gains in vision therefore scale the product, and they do not remove it.

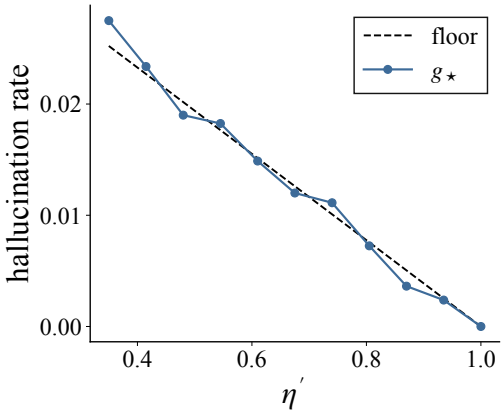


Figure 22: Raising retention from η to η' lowers the floor to $(1-\eta')\widehat{\mathrm{MF}}_v$, and the floor reaches 0 only at $\eta' = 1$.

J.1 Dilution

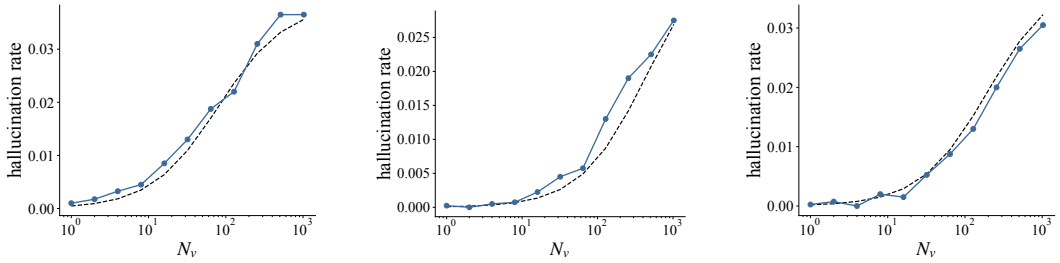
Table 6 records every other value of N_v in the softmax dilution sweep. The softmax column is a Monte Carlo mean of 300 Gaussian logit draws, and the annealed column is $M/(M+N_v)$ with $M = 80.34$. Hallucination of $g_\star$ follows the softmax column of the same table. Half mass occurs at $N_v = M \approx 80$, between the rows $N_v = 64$ and $N_v = 256$.

Table 6: Softmax dilution at $(m, \Delta, \sigma) = (4, 3, 1)$ and seed 0.

N_v	η (softmax)	η (annealed)	hallu$(g_\star)$
1	0.985	0.988	0.0005
4	0.937	0.953	0.0028
16	0.803	0.834	0.0080
64	0.519	0.557	0.0186
256	0.223	0.239	0.0304
1024	0.072	0.073	0.0380

At these parameters the retrieval probability $\Phi_{\mathrm{N}}(\Delta/\sigma - \sqrt{2\log N_v})$ remains above $1/2$ past $N_v = 256$. A needle query can therefore still succeed after the mass budget floor in Table 6 has more than

doubled from $N_v = 16$ to $N_v = 256$. This is the window between POPE and CHAIR stated after Theorem 4. Figure 23 repeats the sweep at three logit configurations. Figure 23a uses $(m, \Delta, \sigma) = (4, 3, 1)$, Figure 23b uses $(8, 4, 1.5)$, and Figure 23c uses $(16, 2.5, 1)$. In each panel the markers are $g_\star$ and the dashed curve is the annealed law.



(a) $(m, \Delta, \sigma) = (4, 3, 1)$. (b) $(8, 4, 1.5)$. (c) $(16, 2.5, 1)$.

Figure 23: Dilution at three logit configurations. Markers are $g_\star$, and dashed curves are the annealed law.

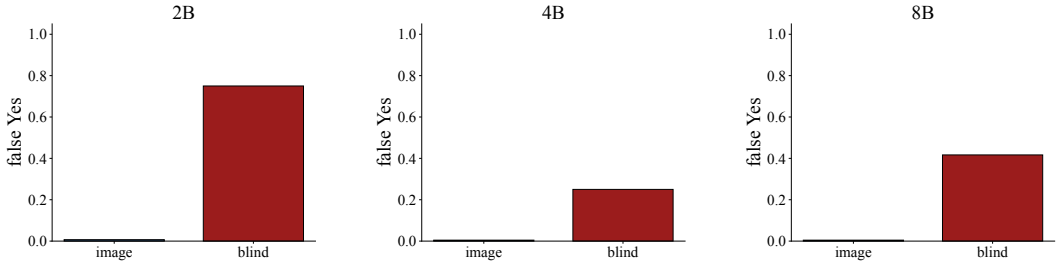
K FROZEN DECODER MEASUREMENTS

Table 7 records an exogenous penalty applied to frozen official POPE logits from Qwen3-VL-8B Instruct. Once $c > 1$, coverage and false Yes fall together. This is the Pareto movement of Corollary 2 on a real decoder, obtained by changing the decision threshold and leaving the weights fixed.

Table 7: Exogenous error penalty on official POPE with frozen Qwen3-VL-8B Instruct.

c	coverage	accuracy	false Yes
1	1.000	0.854	0.057
2	0.962	0.868	0.047
4	0.929	0.878	0.038
9	0.886	0.891	0.029
19	0.844	0.905	0.022
49	0.804	0.914	0.016

Figure 24 compares existence questions that include an image with the matched blind text slice, at three Instruct sizes. Figure 24a, Figure 24b, and Figure 24c are the 2B, 4B, and 8B checkpoints. Image conditioned false Yes stays near zero at each size, while the blind slice, which is the $\eta = 0$ case of Theorem 1, remains large.



(a) 2B. (b) 4B. (c) 8B.

Figure 24: Frozen Instruct existence probes. Image conditioned false Yes stays near zero, and the matched blind slice stays large.

The same attention threshold, now applied to official POPE, is shown in Figure 25. As the threshold rises, coverage falls and false Yes falls with it. The path is again a movement along the scoring line, here implemented by rejecting items whose attention snapshot is too small.

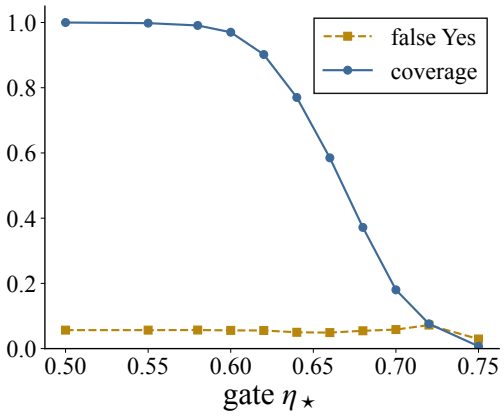


Figure 25: Attention threshold on official POPE. Coverage and false Yes fall together.

Table 8 records CHAIR style captions from the same frozen 8B checkpoint. Mention level rates remain several times the POPE false Yes of this model, as predicted by Theorem 8. No condition falls below the product floor estimated from the attention snapshot. The unrestricted prompt over four images is omitted from the comparison, because objects that are present on a distractor image would be scored as hallucinations and would contaminate the meter. The row that asks for the first image only avoids that contamination.

Table 8: CHAIR style one sentence captions on frozen Qwen3-VL-8B Instruct.

Condition	n	η snapshot	CHAIR_i	proxy floor
side 224	200	0.59	0.276	0.026
side 448	200	0.71	0.279	0.019
four images, first only	60	0.42	0.307	0.037
pixel noise $\sigma = 0.35$	40	0.55	0.238	0.029

Figure 26 plots mention level CHAIR and the attention snapshot against input side length. Changing the side length from 224 to 448 raises the snapshot and leaves CHAIR nearly unchanged, so the extra pixels do not act as new independent noise tokens. Figure 27 applies an attention threshold to single image captions. Coverage drops, while CHAIR barely moves, which is the proxy bias described in Lemma 3. Identifiable dilution therefore uses extra images, as in the four image row of Table 8. A threshold on a single image does not play the same role.

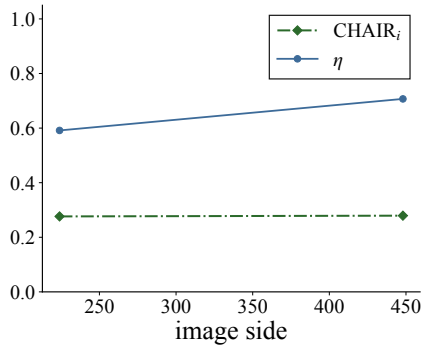


Figure 26: Mention level CHAIR and the attention snapshot versus input side length.

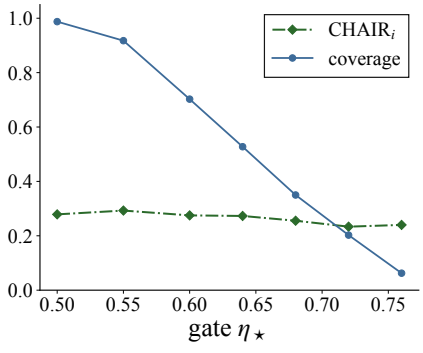


Figure 27: An attention threshold on single image captions lowers coverage, while CHAIR barely moves.