ATTENTION DILUTION: TWO STATISTICAL FAILURE MODES OF VISUAL TOKEN INFLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

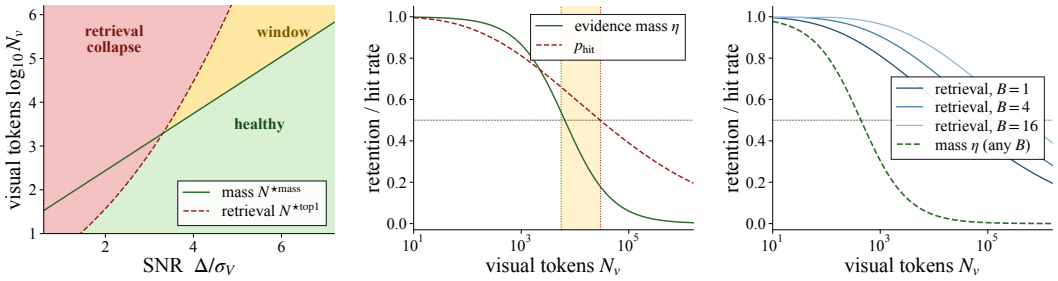
Multimodal language models encode each image as hundreds to thousands of visual tokens. As this count grows, instruction following and fine-grained retrieval degrade, while prune and merge heuristics succeed or fail in ways that appear model dependent. A diffuse attention map need not imply a wrong answer, and no method explains when token count alone causes which failure. We develop a statistical theory of these phenomena. Softmax attention on a two population visual context fails in two ways. Evidence mass on the tokens that carry the answer decays algebraically past a budget set by the margin, and a typical query retains systematically less mass than the average one. Top 1 retrieval of a visual needle collapses at a much larger, exponentially separated scale. When the mass budget lies below the retrieval budget, a diluted yet correct window opens in which attention is already diffuse while the needle remains the argmax. The same split classifies queries by which budget binds and maps when score pruning or covering merges can work for a single frozen attention operation. Monte Carlo studies, a trained attention layer, a CLIP proxy, frozen Qwen3-VL decoders, and a LLaVA-1.5 encoder check instantiate the two laws. Generation accuracy on a referring scene index probe crosses one half in the predicted window once the query is read at the first generated token and the margin is locked from logits alone.

1 INTRODUCTION

Multimodal large language models pay for perception with context. A single image becomes 576 patch tokens in LLaVA-1.5 (Liu et al., 2024a) and tens of thousands under dynamic resolution pipelines such as Qwen2-VL (Wang et al., 2024). A crowded engineering literature prunes or merges these tokens, with success that varies across tasks and models (Chen et al., 2024; Yang et al., 2025; Shang et al., 2025; Zhang et al., 2025c; Alvar et al., 2025; Jiang et al., 2025; Xing et al., 2025; Zhang et al., 2025a; Dong et al., 2026; Takezoe et al., 2026). Diagnostic studies add three observations that no existing method paper explains. Attention score pruning can lose to a random baseline and exhibits strong positional skews (Wen et al., 2025; Zhao et al., 2025). UniPruneBench finds random dropping competitive on many tasks, with optical character recognition (OCR) style benchmarks uniquely compression sensitive (Peng et al., 2025). Practitioners also observe that a visibly diffuse attention map need not imply a wrong answer. What is missing is not another selector but a statistical account of when token count alone causes which kind of failure. The three observations above are usually listed as separate engineering facts. They are one softmax, read at two different statistics, and they come apart as soon as the visual token count grows.

This paper supplies that account. We do not introduce a compressor. We model the logits seen by one attention query as a two population ensemble, with a handful of evidence tokens carrying margin Δ over N_v visual noise tokens of dispersion σ_V. Within this minimal model, visual token inflation induces two failure modes that are governed by different statistics and separated in token count, as Figure 1 shows. The first mode is mass dilution, which is algebraic and limits any query that needs many visual tokens to remain in the residual stream. The second mode is retrieval collapse, which is exponential and limits any query whose answer lives in one visual region. The two modes respond differently to spatial redundancy and to score pruning, and the gap between them is the regime in which a diffuse attention map and a correct answer can coexist. The same split also classifies queries by which budget binds and maps when score pruning versus covering merges can work for a single frozen attention operation.

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.



(a) Phase portrait. (b) Fixed SNR slice. (c) Block size effect.

Figure 1: Mass dilution and retrieval collapse are two statistics of one softmax. Green marks the healthy regime, yellow marks the diluted yet correct window, and red marks retrieval collapse. Block size B shifts retrieval only and leaves the mass law unchanged.

Mass dilution is algebraic. Softmax evidence mass η is a ratio of log normal sums. Conditioning on the signal tokens, the curve tracks a deterministic shape with a query dependent prefactor, as Theorem 1 shows. The annealed prefactor and the typical one differ by a Jensen gap. At high dispersion the typical query retains an order of magnitude less mass than the average theory suggests. Beyond a budget linear in e^{Δ}, retention decays as $1/N_v$. Instruction following, captioning, and any query that needs many visual tokens to remain in the residual stream follow this statistic. A grey attention map is then a true loss of evidence, not a visualization artifact. Retrieval collapse is exponential. Needle queries are an extreme value problem rather than a law of large numbers one. The top 1 hit probability collapses at a budget exponential in $(\Delta/\sigma_V)^2$, located to within half a unit of $\log N_v$ by a second order closed form, as Theorem 2 shows. When the mass budget lies below the retrieval budget, a diluted yet correct window $N_v^{\star\text{mass,typ}} < N_v < N_v^{\star\dagger\text{top1}}$ opens. Evidence mass has already fallen below one half, yet the needle remains the argmax of the visual logits, as Corollary 1 states. This is the regime in which a diffuse attention map and a correct answer can coexist, and in which random pruning can look competitive. Score selection below the retrieval wall is extremal sampling of noise. OCR style and referring queries live here.

The object of the retrieval law is the query that writes the next token. Prompt final attention is a convenient proxy and is often close in location, but it is not the statistic that Theorem 2 names. Generation accuracy is an empirical bridge through values and the vocabulary, not a restatement of the hit probability. We therefore estimate the ratio $c = \Delta/\sigma_V$ from the location and scale of needle versus noise logits on a held out prompt split, at moderate token counts that still contain distractors, and we lock the pair before evaluating unseen prompts. This protocol does not use hit events or generation accuracy as fitting targets. Fitting c to a hit curve would only recover the shape of the standard normal tail and would not be a point prediction. Spatial redundancy among patches is asymmetric. Block correlation rescales the retrieval threshold by an effective count $N_v^{\text{eff}} = N_v/B$ but leaves the mass law invariant. Redundancy protects the needle, not the instruction. Raising resolution therefore does two different things at once, and only one of them is protective. The same dichotomy is a query taxonomy. Caption like queries tend to be mass limited. Fine grained needle queries tend to be retrieval limited. Token merging is then a covering problem rather than a ranking problem, as Theorem 3 shows. Score pruning is near optimal only above the retrieval wall.

To address these gaps we develop a dual scale theory, an identifiable measurement protocol that locks the margin from logits rather than from the outcome one wishes to predict, and a frozen model check that reads attention at the first generated token. The theory names the two budgets, the protocol keeps the overlay from being tautological, and the check reads the query that actually writes the next token. An exact log sum exp merge with softmax weighted values reproduces the original output identically. Clustering merges with a $\log|C_k|$ correction approximate that equality. Recency bias in pruning benchmarks and a duration wall for multi turn and video contexts are corollaries of the same logits. Under regularly varying tails the wall scales as a power of c rather than as $\exp(c^2/2)$, so we report a Student t companion interval beside the Gumbel budget. The main contributions of this work are summarized as follows.

- A dual scale theory of visual token inflation, an algebraic dilution law with a typical versus annealed duality, and an exponential retrieval collapse with a second order budget together with a regularly varying companion.

- Asymmetric spatial protection, which postpones retrieval but not mass and yields a query taxonomy of which budget binds, with an explicit condition for the diluted yet correct window in which a grey map can coexist with a correct needle.
- A covering guarantee that maps when score pruning versus clustering merges can work for a single frozen softmax, including the exact merge equality, with recency bias and a Lambert W duration wall as corollaries.

2 Related Work

Visual token reduction. Training-free visual token reduction begins with token merging in vision transformers (Bolya et al., 2023) and now spans a wide range of approaches. Attention score pruning selects tokens by their weights (Chen et al., 2024; Zhang et al., 2025a;c; Takezoe et al., 2026), while similarity-based merging clusters redundant tokens (Shang et al., 2025; Yang et al., 2025). Staged dropping reduces tokens layer by layer (Xing et al., 2025), and diversity or coverage maximization selects subsets that preserve visual variety (Alvar et al., 2025; Dong et al., 2026). Other lines include graph-propagation selection (Jiang et al., 2025), graph foundation models (Zhu et al., 2026), extreme pre-fusion (Zhang et al., 2025b), first-principles subset construction (Wang et al., 2026), and multi-objective covering with a Hausdorff bound (Li et al., 2026). Diagnostics make any ranking of these schemes unstable. Score pruning exhibits positional and recency skews and can lose to a random baseline (Wen et al., 2025; Zhao et al., 2025). UniPruneBench reports that random dropping is competitive on many tasks, that no method is uniformly best, and that OCR style benchmarks are uniquely compression sensitive (Peng et al., 2025). Failures of text-guided pruning have been traced to cross-modal misalignment (Xu et al., 2026; Li et al., 2025; Gu et al., 2026). Methods which look like ranking, FastV-style score pruning being the prototype, and methods which look like covering, VisionZip-style merges and diversity subsets being the prototype, trade places as the task changes. These papers select tokens under a prescribed budget. The prior question is whether the token count itself has already induced a statistical failure and which kind of failure is occurring. The retrieval wall is a candidate account of OCR style fragility and of random dropping below threshold.

Statistical theories of attention. Statistical theories of attention analyze softmax as a capacity-limited selector. Existing analyses give vanishing weights of order $1/L$ (Mudarisov et al., 2025), show that circuits disperse as the input size grows (Veličković et al., 2025), and track entropy and rank collapse along depth (Zhai et al., 2023; Dong et al., 2021; Geshkovski et al., 2023; Giorlandino & Goldt, 2026). A parallel literature uses the phrase attention dilution for text long-context degradation (Zhang et al., 2024). Bohbot et al. (2026) give a token sample complexity for attention in a one-population Gaussian ensemble, while the present model is two-population. Mass dilution is a law of large numbers on the noise sum, whereas retrieval is an extreme-value contest between the signal and the noise maximum. Those scales separate when equation 7 holds, a split with no analogue in a homogeneous field. Chen et al. (2026) identify a critical logarithmic attention scaling that prevents rank collapse in a homogeneous token ensemble, and test-time training work finds a score dilution margin of $\Omega(\log T)$ (Bansal et al., 2026). Both speak to the mass law, which we strengthen to an explicit prefactor and complement with the extremal top-1 law. Our proofs use the Fisher-Tippett-Gnedenko theorem and Berman's condition for dependent Gaussian maxima, log-normal concentration, the random-energy-model freezing transition, and covering numbers, whose capacity control for attention was studied by Edelman et al. (2022). The two laws are independent properties that fail on different scales, so they are established separately, in the same spirit as safety versus liveness and almost-sure termination arguments for concurrent protocols (Qiu et al., 2026a;b).

Long-context retrieval and positional effects. Long-context retrieval and positional effects study how models lose access to information as the context grows. Long-context language models lose retrieval in a position-dependent way (Liu et al., 2024b; Hsieh et al., 2024), and zero-shot composed image retrieval poses an analogous visual needle problem (Tang et al., 2026), with multitask deep evidential fusion (Lou et al., 2026) a related reference-free visual read. Positional encodings remain a complementary mechanism for text length generalization (Xu et al., 2024; Barbero et al., 2025), and over-squashing explains why deep layers lose sensitivity to early tokens (Barbero et al., 2024). Attention sinks and visual attention re-sinking further complicate the picture (Xiao et al., 2024; Kan et al., 2026). The dual-scale statement is a consequence of the multimodal two-population structure. Mass dilution and retrieval collapse are different statistics, with opposite responses to spatial redundancy.

3 TWO STATISTICAL FAILURE MODES

We consider a single attention query, taken in applications to be the position that writes the first answer token, attending to N_t text tokens and N_v visual tokens. A subset $\mathcal{S}$ of size m carries the instruction evidence. The remaining visual positions form the noise set $\mathcal{V}$. With logits $\{s_i\}$ and weights $\alpha = \mathrm{softmax}(s)$, the evidence retention rate is

$$\eta(N_v) \;:=\; \sum_{i\in\mathcal{S}} \alpha_i \;=\; \frac{A}{A+B}, \qquad A := \sum_{i\in\mathcal{S}} e^{s_i}, \quad B := \sum_{j\in\mathcal{V}} e^{s_j}. \tag{1}$$

Needle retrieval is the probability that a distinguished visual logit s^* is the unique visual maximum, $p_{\mathrm{hit}}(N_v) = \Pr[s^* > \max_{j\neq j^*} s_j]$.

Assumption 1 (Signal and noise decomposition)**.** *There exist a margin $\Delta > 0$ and scales $\sigma_S, \sigma_V > 0$ such that $s_i = \mu_V + \Delta + \sigma_S \xi_i$ for $i \in \mathcal{S}$ and $s_j = \mu_V + \sigma_V \zeta_j$ for $j \in \mathcal{V}$, where ξ_i, ζ_j are zero mean, unit variance, and mutually independent. For tail results we take them Gaussian. Softmax is shift invariant, so only Δ and the two scales enter.*

A query typically concerns $O(1)$ objects, and the remaining $1 - o(1)$ fraction of patches are semantically irrelevant. Their projections onto a query direction are well approximated as independent and identically distributed (i.i.d.) in the bulk, and spatial correlation is restored in Lemmas 2 and 3. Practical MLLMs operate at $N_v \in [10^2, 10^5]$, $m \in [1, 10^2]$, and σ_V of order 1 to 5. Asymptotics are in $N_v \to \infty$ with m, Δ, and the scales fixed. The two population split is the modeling choice that makes the two scales visible. A one population Gaussian field has a single extreme value law and no distinguished evidence mass, so it cannot produce a diluted yet correct window. Heterogeneity in the visual context is therefore not a nuisance parameter but the structure that separates instruction following from needle retrieval.

The signal mass A is a sum of m log normals and does not self average at small m, so retention is random across queries. Conditioning on the signals, the curve tracks a deterministic shape with a query dependent prefactor. The noise mass B concentrates by Lemma 1, and the ratio $\eta = A/(A + B)$ is governed by the relative size of A and B alone. The two statistics η and p_{hit} live on different scales, and the split between them is the main object of the theorems below.

Lemma 1 (Concentration of noise mass)**.** *Let $B = \sum_{j=1}^{N_v} e^{\mu_V + \sigma_V \zeta_j}$ with $\zeta_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Then $\mathbb{E}[B/N_v] = e^{\mu_V + \sigma_V^2/2}$ and, with probability at least $1 - \delta$, $\left|B/N_v - e^{\mu_V+\sigma_V^2/2}\right| \le e^{\mu_V}\sqrt{(e^{2\sigma_V^2} - e^{\sigma_V^2})/(N_v\delta)}$.*

Theorem 1 (Dilution law)**.** *Under Assumption 1 with Gaussian noise, fix $m, \Delta, \sigma_S, \sigma_V$. Define $\hat{A} := A\, e^{-(\mu_V+\sigma_V^2/2)}$. As $N_v \to \infty$,*

$$\frac{\eta(N_v)}{\hat{A}/(\hat{A} + N_v)} \;\xrightarrow{p}\; 1. \tag{2}$$

Let $\bar{W} := \frac{1}{m}\sum_{i=1}^{m} e^{\sigma_S \xi_i}$ and $\kappa(m, \sigma_S) := \sigma_S^2/2 - \mathbb{E}\log \bar{W} \ge 0$, so that $\log \hat{A} = \log m + \Delta - \sigma_V^2/2 + \log \bar{W}$. The annealed and typical half retention budgets are

$$M_{\mathrm{ann}} = m\, e^{\Delta + (\sigma_S^2 - \sigma_V^2)/2}, \qquad M_{\mathrm{typ}} = m\, e^{\Delta - \sigma_V^2/2 + \mathbb{E}\log \bar{W}}. \tag{3}$$

The median curve sits a factor $e^{-\kappa}$ below the annealed one. Beyond the budget, η decays as $\Theta(1/N_v)$. The finite N_v Chebyshev bound, the self averaging versus freezing regimes, and the REM correction are in Appendix B.

Higher σ_V accelerates dilution at fixed mean margin, because cluttered backgrounds are intrinsically harder to attend through. The quantity κ that the typical law uses is the numerical moment $\sigma_S^2/2 - \mathbb{E}\log \bar{W}$, and the REM formula $\frac{1}{2}(\sigma_S - \sqrt{2\log m})_+^2$ is recovered only as $m \to \infty$ inside the freezing regime. For the operating point $(m, \sigma_S) = (16, 3.5)$ that gap is 2.24, while the leading REM formula gives 0.66, so a typical query calculation that uses the infinite m correction would overstate retention by a large constant factor. An annealed overlay fitted to a mean attention map will therefore look healthier than the map of a typical prompt, and benchmark averages can hide a long lower tail of diluted queries. The gap between the typical and annealed prefactors is the statistical origin of the unpredictability reported in the diagnostic literature.

Theorem 2 (Retrieval phase transition)**.** *Let* $b_n = \sqrt{2\log n} - \frac{\log\log n + \log 4\pi}{2\sqrt{2\log n}}$ *and* $a_n = (2\log n)^{-1/2}$*. Under Assumption 1 with Gaussian noise,*

$$\Big| p_{\mathrm{hit}}(N_v) - \mathbb{E}_G\, \bar{\Phi}\big(b_{N_v} + a_{N_v} G - \Delta/\sigma_V\big) \Big| \to 0 \qquad (N_v \to \infty), \tag{4}$$

where G *is standard Gumbel,* Φ *is the standard normal cumulative distribution function, and* $\bar{\Phi} = 1 - \Phi$ *is its tail. Hence*

$$p_{\mathrm{hit}}(N_v) = \Phi\big(\Delta/\sigma_V - \sqrt{2\log N_v}\big) + o(1). \tag{5}$$

The leading budget $N_v^{\star\mathrm{top1}} = \exp(\Delta^2/2\sigma_V^2)$ *underestimates the half crossing. One fixed point correction of* $b_n = c$ *yields*

$$N_v^{\star\dagger\mathrm{top1}} = \sqrt{2\pi}\,\frac{\Delta}{\sigma_V}\,\exp\Big(\frac{\Delta^2}{2\sigma_V^2}\Big). \tag{6}$$

The fall from $1-\epsilon$ *to* ϵ *has relative width* $\Theta(\sigma_V/\Delta)$ *in* $\log N_v$*. Heteroscedastic and* m *signal lifts are in Appendix C.*

If the noise upper tail is regularly varying of index α, maxima belong to the Fréchet domain and the wall scales as $(\Delta/\sigma_V)^\alpha$ rather than $\exp(c^2/2)$. Heavier tails move the wall left of the Gumbel prediction, so the retrieval budget is sensitive to the noise distribution in a way that the mass budget is not. Student-t_ν is a computable intermediate, and we report $N_v^{\star\dagger\mathrm{top1}}$ together with a t_5 companion whenever the empirical Hill estimator of the tail index is finite. The Gumbel budget remains the main statement because CLIP noise and the trained layer are close to Gaussian in the bulk, and the companion is a robustness interval rather than a replacement for Theorem 2. The mass law depends on the mean of $e^{\sigma_V\zeta}$, which is finite for any noise distribution with a finite second moment, while the retrieval law depends on the tail of ζ. Tail misspecification therefore shifts the retrieval wall but leaves the mass budget untouched.

Corollary 1 (Two thresholds)**.** *Let* $c = \Delta/\sigma_V$*. The typical mass budget and the second order retrieval budget satisfy* $N_v^{\star\mathrm{mass,typ}} < N_v^{\star\dagger\mathrm{top1}}$ *if and only if*

$$\log m + \mathbb{E}\log \bar{W} \;<\; \tfrac{c^2}{2} + \log\big(c\sqrt{2\pi}\big) - c\,\sigma_V + \tfrac{\sigma_V^2}{2}. \tag{7}$$

When the inequality holds, a nonempty window $N_v^{\star\mathrm{mass,typ}} < N_v < N_v^{\star\dagger\mathrm{top1}}$ *exists. When it fails, retrieval collapse can precede mass half loss. Failure occurs at small* c *or large* m*. Figure 1 uses* $(m, \sigma_V) = (16, 1.5)$*, where the window opens for* $c \gtrsim 3.5$*.*

The window is therefore a property of the query, not a universal phase of token inflation. Caption like queries with many contributing patches sit closer to that failure region, while referring queries with a small m sit farther from it. The distinction between the two query classes is sharpened by the m signal retrieval lift. If the answer may use any of m signal tokens, the signal maximum contributes an extremal lift $\sigma_S\sqrt{2\log m}$, and the retrieval budget becomes $\exp((\Delta + \sigma_S\sqrt{2\log m})^2/2\sigma_V^2)$. Redundancy therefore raises the retrieval budget multiplicatively in log scale, while it raises the mass budget only linearly by Theorem 1. A caption query with many contributing patches can be mass limited while a referring query with the same total margin remains retrieval limited, because the two corrections act on different statistics and are not interchangeable. The taxonomy organizes the diagnostic facts into one statistical picture. A caption like request at moderate c and large m loses mass first, so a grey attention map is a true loss of evidence, while a referring request at the same c and small m keeps the needle as the visual argmax long after mass has dropped, so a grey map coexists with a correct answer. The same split also predicts when random dropping can look competitive, because below the retrieval wall the retained set is an exchangeable sample of noise and a random rule loses nothing.

Lemma 2 (Block correlation on the retrieval side)**.** *If the* N_v *noise logits partition into* N_v/B *independent blocks of* $B = \ell^2$ *perfectly correlated logits, the needle alone in its block, then Theorem 2 holds with* N_v *replaced by* $N_v^{\mathrm{eff}} = N_v/B$*. The raw retrieval budget inflates by exactly* B*.*

Lemma 3 (Block correlation on the mass side)**.** *Under the same block structure, blockwise multiplicities cancel in* B_{noise}/N_v*, and the annealed dilution law is unchanged. Block correlation only widens fluctuations and weakens the freezing correction.*

Corollary 2 (Resolution invariance)**.** *Doubling image resolution at fixed physical content quadruples* N_v *but also doubles the correlation length* ℓ *in patch units, leaving* $N_v^{\mathrm{eff}} \approx N_v/\ell^2$ *invariant.*

Raising pixel count is therefore not a statistical free lunch for instruction retention, and it does not by itself create new retrieval noise once correlation length scales with the grid. The asymmetry between the two failure modes is the reason resolution invariance holds on the retrieval side but not on the mass side. The two lemmas together imply that spatial redundancy protects the needle but not the instruction, which drives the query taxonomy and the method map of Section 4.

4 GUARANTEED MERGING

A merging scheme partitions $\mathcal{V} = \bigsqcup_{k=1}^{K} C_k$ and represents each cluster by a pair $(\hat{s}_k, \hat{v}_k)$, where $\hat{s}_k$ is a logit and $\hat{v}_k$ is a value vector. Let O and O' denote the attention outputs before and after merging, and suppose $\|v_i\| \le V$ for every visual token. The question is how much the output moves under a given partition and a given representative choice. The answer splits into two independent errors, one on the logit side and one on the value side, and the split is what makes the method map tractable. The statement applies to any model at inference time, because it concerns one frozen softmax rather than a training procedure.

Theorem 3 (Guaranteed merging)**.** *Let $\tilde{s}_k = \log \sum_{j \in C_k} e^{s_j}$ and $\tilde{v}_k = (\sum_{j \in C_k} e^{s_j} v_j)/e^{\tilde{s}_k}$ denote the exact merge of cluster C_k. Then $O' = O$ identically. If a scheme instead satisfies $\max_k |\hat{s}_k - \tilde{s}_k| \le \epsilon_s$ and $\max_k \|\tilde{v}_k - \hat{v}_k\| \le \delta'_v$, then*

$$\|O - O'\| \;\le\; V\,\epsilon_s + \delta'_v. \tag{8}$$

The value bound holds whenever $\max_k \max_{j \in C_k} \|v_j - \hat{v}_k\| \le \delta_v$. The statement is an output perturbation for one frozen operation, not a training time guarantee.

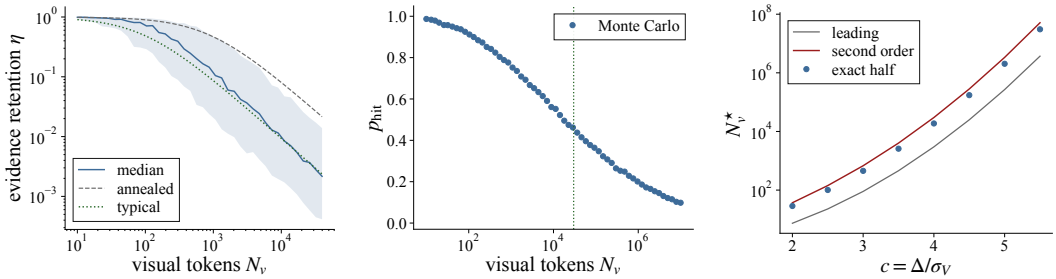
The equality case is the object that later experiments implement as an oracle. Adding $\log |C_k|$ to a centroid logit is a first order approximation to $\tilde{s}_k$, and the correction is exact when all logits in C_k are equal. Score pruning is near optimal only above the retrieval wall, as Proposition 3 in Appendix D shows. Below that wall the retained set is an exchangeable extremal sample of noise, so a random baseline can match or beat a score rule because the ranking of noise logits carries no information about the ranking of noise values. Clustering merges are SNR robust once a mass correction is applied, and principled subset selection corresponds to Δ weighting, covering, and covering number minimization, as Propositions 4 and 5 record.

The resulting map is a statement about one frozen softmax, not a claim that any named compressor is uniformly best. A method that is near optimal on an OCR style needle at small N_v can look no better than random once the same scene is past the retrieval wall, while a covering merge that looks wasteful in the healthy regime becomes the more stable default after mass has already dropped. The regime label therefore predicts which method is better for a given query at a given token count, not which method is better in an absolute sense. The experiments in Section 5 verify that prediction on reconstruction error, and frozen decoders replace the error ranking by a generation ranking that is allowed to depend on the model.

Corollary 3 (Covering number budget)**.** *If visual value vectors have covering number $\mathcal{N}(\delta_v/2)$, a merger with $K \le \mathcal{N}(\delta_v/2)$ clusters achieves $\|O - O'\| \le \delta_v$. For r objects whose values lie on ν-dimensional manifolds of radius R, $K^\star(\epsilon) = \Theta(r(2R/\epsilon)^\nu)$, independent of N_v. Combined with Theorem 1, any $K \le \hat{A}\frac{\eta_0}{1-\eta_0}$ restores retention to $\eta \ge \eta_0$.*

The covering budget $K^\star$ is independent of N_v, which is the formal reason that spatial redundancy does not inflate the number of clusters needed to preserve a target retention. Combined with Theorem 1, the same budget that restores mass also controls the output error, so the mass law and the merge guarantee share a single scale. The second order retrieval budget $N_v^{\star\dagger\mathrm{top1}}$ tracks the exact half crossing to within half a unit of $\log N_v$ over the operating range $c \in [2, 5.5]$, while the leading formula $N_v^{\star\mathrm{top1}} = \exp(\Delta^2/2\sigma_V^2)$ understates the token count at which a needle remains recoverable by a gap that grows with c.

The same logits read along position recover the recency skew of top K pruning and the reweighting that undoes it, as Propositions 6 and 7 show. A linear tilt $\Delta(i) = \Delta_0 - \lambda d(i, q)$ makes keep probability a decreasing function of distance to the query, which is the positional skew reported by pruning diagnostics. Reweighting by $\hat{\lambda} d(i, q)$ undoes that skew if and only if $\hat{\lambda}$ matches the true tilt. Read along time they give a Lambert W duration wall for a single instruction stretched over T



(a) Typical versus annealed mass. (b) Hit probability versus N_v. (c) Budgets versus c.

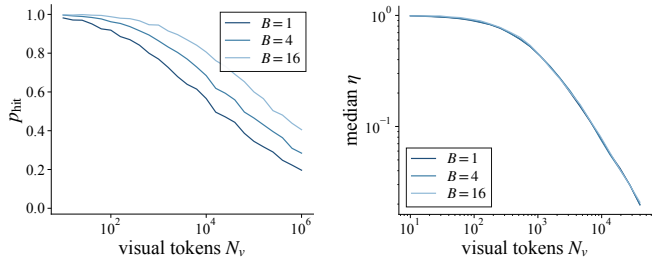
Figure 2: Monte Carlo checks of Theorems 1 and 2. Median retention follows the typical law rather than the annealed curve, and the second order budget locates retrieval collapse.

turns, as Proposition 8 records. A persistent instruction hits an algebraic wall linear in e^{Δ_0}, while a fresh one each turn does not accumulate. Video and multi turn settings are therefore the same two population softmax with a growing noise pool, not a separate phenomenon. The duration wall is the time analogue of the mass budget, and the recency skew is the position analogue of the retrieval wall.

5 Experiments

5.1 Mass and retrieval on controlled logits

Figure 2 compares simulated retention and hit probability with Theorems 1 and 2. The dilution law concerns a typical query rather than the average attention map. When the margin is large, the typical curve tracks the median retention, and the relative error is 2.3%. As the signal scale increases, the median moves away from the annealed curve by the Jensen gap of Theorem 1. At signal scale 3.5 the gap equals 2.24, and the typical law still tracks the median, so a report that shows only the mean map describes a typical prompt as healthier than it is. This is one reason that pruning by attention score can change rank from one prompt to the next. Far into the diluted regime, the slope on logarithmic axes equals -1, which matches the algebraic tail. Retrieval obeys a different limit. Although the hit curves follow the Gumbel law across six decades of N_v, the leading budget still misses the half crossing by 1.36 to 2.11 units of $\log N_v$. A single fixed point correction reduces that error to between 0.26 and 0.51 for c in the operating range. A keep budget copied from the leading expression would therefore drop needles that remain recoverable. Both comparisons still treat visual tokens as independent draws. The next question is what changes when neighboring patches share a logit.



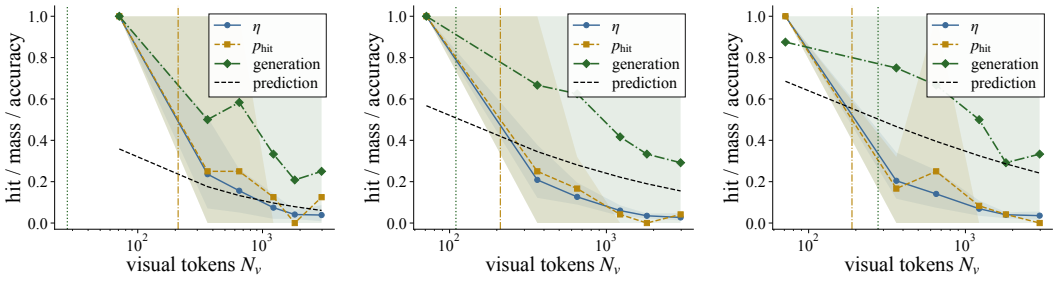
(a) Retrieval shifts with B. (b) Mass stays on one law.

Figure 3: Block correlation is asymmetric. Retrieval moves by the block size B, while mass stays on one annealed law.

Table 1: Regime occupancy among 24 CLIP queries.

N_v	healthy	diluted yet correct	retrieval collapse
144	0.67	0.17	0.17
256	0.63	0.17	0.21
576	0.42	0.33	0.25
1024	0.29	0.46	0.25
2304	0.08	0.58	0.33
4096	0.08	0.54	0.38
9216	0.00	0.54	0.46
16384	0.00	0.46	0.54

Figure 3 shows how the two laws separate once neighboring patches share a logit. In Figure 3(a), hit curves for block sizes 1, 4, and 16 fall onto one curve once N_v is replaced by $N_v^{\text{eff}} = N_v/B$. That collapse is Lemma 2. Perfectly tied logits inside a block contribute one extreme value draw, so the retrieval wall moves by the factor B and not by a small location shift. Figure 3(b) is the matching statement for mass. The three curves lie on the same annealed law, and a larger block only widens



(a) Qwen3-VL-2B. (b) Qwen3-VL-4B. (c) Qwen3-VL-8B.

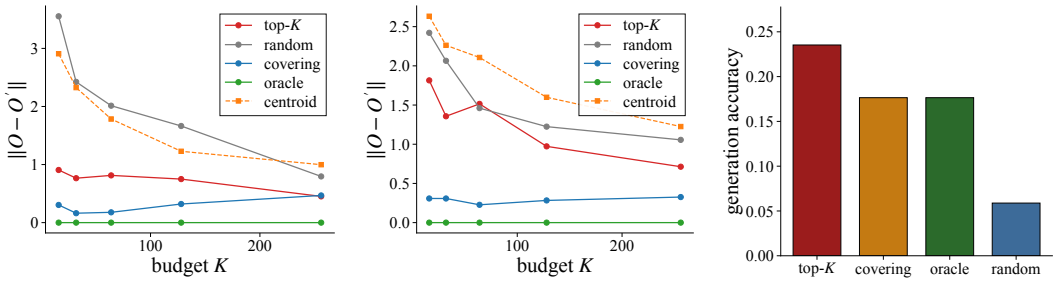
Figure 4: Locked margin overlay on frozen Qwen3-VL. (a), (b), and (c) add distractors. Generation crosses one half only after hit probability and evidence mass have fallen.

the fluctuations, which is Lemma 3. Raising resolution at fixed scene content is this mechanism on a real grid. Moving from 576 to 2304 patches leaves the hit probability at 0.88, and a further rise to 9216 patches leaves it there. Corollary 2 is therefore already visible at sizes used by current vision encoders. A shared token count is an incomplete coordinate for the retrieval wall. A fine grid over one scene and a collage of unrelated tiles can report the same N_v and still fall in different regimes, because only the grid has a long correlation length. The noise tail is a separate coordinate from that geometry. At $c = 3.2$, a Student t law of index 5 places the wall near 50 tokens, while the Gumbel wall is 1342 and the exact Gaussian crossing is 882. Under that change the mass budget stays where Theorem 1 places it, since dilution uses a moment of $e^{\sigma_V \zeta}$ and retrieval uses the domain of the maximum.

5.2 The window on real queries and frozen models

Once known, two budgets sort real prompts. Patch text scores from a CLIP ViT-B/32 encoder (Radford et al., 2021) on 160 COCO images (Lin et al., 2014) give pooled noise within distance 0.012 of a fitted Gaussian, although 10 of the 24 prompts have upper tails heavier than Gaussian. As N_v grows from 98 to 3136, hit rate falls from 0.96 to 0.58 and retention falls from 0.65 to 0.17. At the matched margin $c = 3.2$, the empirical tail crosses near 860 tokens, close to the Gaussian crossing of 882 and far from the Student t wall. These scores therefore support the Gumbel budget, with the heavy tail formula kept as a companion bound. Fits with five signal tokens give geometric mean budgets of 325 and 8.1×10^3, a ratio of about 25, and every prompt satisfies equation 7. Table 1 shows how the prompts then fill the three regimes. At 576 tokens one third are already diluted yet correct, and at 9216 tokens the healthy share is zero. Object prompts remain retrievable past the measured range, whereas scene prompts with c near 1.45 have collapsed after a few dozen tokens. The ratio of the two budgets grows with the signal to noise ratio, which is the taxonomy of Section 3 read from data. A fit on 120 images, evaluated on the other 40, has median relative error near 0.22, so the regime counts are not an artifact of testing the images used to estimate c.

Figure 4 tests the window on frozen Qwen3-VL decoders (Bai et al., 2025), with attention read at the first generated token. That position is the query in Theorem 2. The ratio c is locked from the location and scale of needle and noise logits on a held out third of the prompts, and the fit is formed from those two quantities alone. The protocol is recorded in Appendix F. On the 8B model the lock gives $c = 2.72$ and a retrieval budget of 277. On unseen prompts the hit probability crosses one half at 189 tokens, a gap of 0.38 in $\log N_v$. Generation falls from 0.88 to 0.50 as N_v grows from 71 to 1223, while mass at the generated token is already 0.20 by 363 tokens. Mass is therefore lost first, the needle remains the visual argmax, and the written answer fails later. Qwen3-VL-4B keeps this order, with a gap of 0.65 in $\log N_v$ and a generation half crossing near 950 tokens. On the 2B model the locked budget of 27 lies below the hit crossing at 211, so the Gaussian wall is conservative there, and generation stays at or above one half through 651 tokens. Prompt final estimates of c stay within 0.22 of the generation token values, so the residual is not a mismatch of readout position. LLaVA-1.5 (Liu et al., 2024a) holds N_v fixed at 576 and reduces the needle share by a collage. Generation then falls from 1.00 at one tile to 0.31 at four tiles, and mass follows the share. A trained head fitted at $(8.45, 2.13)$ matches the hit law within 0.024. Appendix G shows that the same logits also recover the recency skew and the duration wall. The failures arrive in the same order on every system checked here, while the token count of each failure moves with the model.



(a) Synthetic error above the wall. (b) Synthetic error below the wall. (c) Generation below the wall.

Figure 5: Under a locked retrieval wall, score pruning stays near the exact merge above the wall and moves toward random dropping below it. Covering holds the error near 0.31. Generation on Qwen3-VL-4B still favors score pruning.

5.3 Which rule the retrieval budget selects

The budgets above also determine which compression rule can work. Theorem 3 makes the exact merge an identity, and it predicts that score pruning should be reliable only above the retrieval wall. Figure 5 compares reconstruction error on the two sides of that wall, for a target of ten patches. Above the wall, at a keep budget of 32, score pruning preserves the target and the output error is 0.77, compared with 2.42 for random dropping. Adding $\log |C_k|$ to a centroid logit restores retained mass from 0.46 to 0.87 and reduces the error to 0.16, while the exact merge stays at numerical zero. Below the wall the ordering changes. The error under score pruning rises to 1.36, and the corrected covering merge holds the error near 0.31. Proposition 3 accounts for the reversal. Past the wall the retained tokens are an exchangeable sample of noise, so ranking them by attention score no longer ranks the target. A random rule can then match a score rule because the score has stopped carrying information about the target.

Generation ranks the rules differently from reconstruction, because the next token depends on the value vectors and on the vocabulary, while Theorem 3 bounds only the attention output. On Qwen3-VL-4B, below the wall and at the same keep budget, Figure 5(c) shows top K accuracy 0.24, covering accuracy 0.18, and random accuracy 0.06. Score pruning still leads on this decoder and this probe after reconstruction error has already favored the covering merge. The oracle merge attains an output error of 4×10^{-5} here, and the same near equality holds on Qwen3-VL-2B, on Qwen3-VL-8B, and on LLaVA-1.5, with errors 10^{-4}, 4×10^{-5}, and 2.5×10^{-6}. The heuristic ranking is less stable, as the additional bars in Appendix G show. Covering leads on the 2B decoder, with accuracy 0.29 against 0.21 for top K. Random dropping leads on the 8B decoder, with accuracy 0.17 against 0.08 for covering, and it also leads on the pooled LLaVA collage, with accuracy 0.24 against 0.20 for covering and 0.12 for top K. Which rule is better is therefore a function of the model and of the budget that binds. A single keep ratio, quoted without the query class, has no scale on which those changes can be compared. A benchmark is still required for systems that train the merge, route across layers, or edit the vision encoder. For those systems the two laws indicate which axis to vary after c has been read from the logits.

6 Conclusion

This paper accounts for visual token inflation as two statistics of one softmax. Evidence mass decays algebraically past a typical budget that sits below the annealed mean by a Jensen gap, while top 1 retrieval of a visual needle collapses at an exponentially larger scale that a second order budget locates to within half a unit of $\log N_v$. Between those budgets a diluted yet correct window opens, so a grey map can coexist with a needle that remains the argmax. Block correlation postpones retrieval through $N_v^{\text{eff}} = N_v/B$ and leaves mass on its annealed curve, so a fine grid and a collage of the same length fall under different budgets, as does each query. The same wall selects the compressor. Score pruning stays close to the exact log sum exp merge above the wall, and that merge is an identity, while a covering merge keeps the output faithful once scores sample noise. Monte Carlo curves, a trained head, a CLIP proxy, and frozen Qwen3-VL and LLaVA-1.5 checks, with c locked from the first generated token, place the hit law and referring generation in this window.

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

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9392–9401. IEEE, 2025.

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.

Rachit Bansal, Aston Zhang, Rishabh Tiwari, Lovish Madaan, Venkata Sai Surya Subramanyam Duvvuri, Devvrit Khatri, David Brandfonbrener, David Alvarez-Melis, Prajjwal Bhargava, Mihir Kale, et al. Let's (not) just put things in context: Test-time training for long-context llms. In *International Conference on Learning Representations*, pp. 112130–112153, 2026.

Federico Barbero, Andrea Banino, Steven Kapturowski, Dharshan Kumaran, João G Araújo, Alex Vitvitskyi, Razvan Pascanu, and Petar Veličković. Transformers need glasses! information over-squashing in language tasks. *Advances in Neural Information Processing Systems*, 37:98111–98142, 2024.

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! what makes rotary positional encodings useful? In *International Conference on Learning Representations (ICLR)*, pp. 1–31, 2025.

Léa Bohbot, Cyril Letrouit, Gabriel Peyré, and François-Xavier Vialard. Token sample complexity of attention. In *International Conference on Machine Learning (ICML)*, pp. 1–49, 2026.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. In *International Conference on Learning Representations (ICLR)*, pp. 1–20, 2023.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35, 2024.

Shi Chen, Zhengjiang Lin, Yury Polyanskiy, and Philippe Rigollet. Critical attention scaling in long-context transformers. In *International Conference on Learning Representations*, pp. 86454–86478, 2026.

Sixun Dong, Juhua Hu, Mian Zhang, Ming Yin, Yanjie Fu, and Qi Qian. Mmtok: Multimodal coverage maximization for efficient inference of vlms. In *International Conference on Learning Representations*, pp. 48075–48099, 2026.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning (ICML)*, pp. 2793–2803. PMLR, 2021.

Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning (ICML)*, pp. 5793–5831. PMLR, 2022.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. *Advances in Neural Information Processing Systems*, 36:57026–57037, 2023.

Alessio Giorlandino and Sebastian Goldt. Two failure modes of deep transformers and how to avoid them: a unified theory of signal propagation at initialisation. In *International Conference on Learning Representations*, pp. 63101–63127, 2026.

Shishen Gu, Jiequan Cui, Wenbo Hu, Zenglin Shi, Zhenzhen Hu, and Richang Hong. Visual token compression enhances robustness of mllms. *arXiv preprint arXiv:2607.22716*, 2026.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? In *Conference on Language Modeling (COLM)*, pp. 1–27, 2024.

Yutao Jiang, Qiong Wu, Wenhao Lin, Wei Yu, and Yiyi Zhou. What kind of visual tokens do we need? training-free visual token pruning for multi-modal large language models from the perspective of graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4075–4083, 2025.

Zhehan Kan, Xin Li, Yanlin Liu, Xiaochen Yang, Xinghua Jiang, Yinsong Liu, Deqiang Jiang, Xing Sun, Qingmin Liao, and Wenming Yang. Rar: Reversing visual attention re-sinking for unlocking potential in multimodal large language models. In *International Conference on Learning Representations*, pp. 39643–39660, 2026.

Kaiyuan Li, Xiaoyue Chen, Chen Gao, Yong Li, and Xinlei Chen. Balanced token pruning: Accelerating vision language models beyond local optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1–25, 2025.

Yangfu Li, Hongjian Zhan, Tianyi Chen, Qi Liu, Yu-jie Xiong, and Yue Lu. Why 1+ 1¡ 1 in visual token pruning: Beyond naive integration via multi-objective balanced covering. In *Advances in Neural Information Processing Systems*, volume 38, pp. 71999–72039, 2026.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024a.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024b.

Yiwei Lou, Yuanpeng He, Rongchao Zhang, Yongzhi Cao, Hanpin Wang, and Yu Huang. Multitasks-based deep evidential validation network for blind image quality assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7512–7520, 2026.

Timur Mudarisov, Mikhail Burtsev, Tatiana Petrova, and Radu State. Limitations of normalization in attention mechanism. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1–24, 2025.

Tianfan Peng, Yuntao Du, Pengzhou Ji, Shijie Dong, Kailin Jiang, Mingchuan Ma, Yijun Tian, Jinhe Bi, Qian Li, Wei Du, et al. Can visual input be compressed? a visual token compression benchmark for large multimodal models. *arXiv preprint arXiv:2511.02650*, 2025.

Longfei Qiu, Jingqi Xiao, and Zhong Shao. Mechanized safety and liveness proofs for the mysticeti consensus protocol under the lido-dag framework. In *2026 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*, 10:1152–1177, 2026b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22857–22867, 2025.

Rinyoichi Takezoe, Yaqian Li, Zi-Hao Bo, Anzhou Hou, Mo Guang, and Kaiwen Long. Learnpruner: Rethinking attention-based token pruning in vision language models. In *International Conference on Learning Representations*, pp. 66381–66400, 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.

Petar Veličković, Christos Perivolaropoulos, Federico Barbero, and Razvan Pascanu. Softmax is not enough (for sharp size generalisation). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 61190–61211, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Tinghao Wang, Yichen Guo, Rui Huang, Zheng Lu, Qizhe Zhang, Chenxi Li, Yuan Zhang, Jiajun Cao, Zhirong Shen, Yaosong Du, et al. Tops: First-principles visual token pruning via constructing token optimal preservation sets for efficient mllm inference. *arXiv preprint arXiv:2606.27161*, 2026.

Zichen Wen, Yifeng Gao, Weijia Li, Conghui He, and Linfeng Zhang. Token pruning in multimodal large language models: Are we solving the right problem? In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 15537–15549, 2025.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, volume 2024, pp. 21875–21895, 2024.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, et al. Conical visual concentration for efficient large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14593–14603. IEEE, 2025.

Mingyu Xu, Xin Men, Bingning Wang, Qingyu Zhang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. Base of rope bounds context length. In *Advances in Neural Information Processing Systems*, volume 37, pp. 87386–87410, 2024.

Rui Xu, Yunke Wang, Yong Luo, and Bo Du. Rethinking visual token reduction in lvlms under cross-modal misalignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 27323–27331, 2026.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19792–19802, 2025.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *International Conference on Machine Learning (ICML)*, pp. 40770–40803. PMLR, 2023.

Qizhe Zhang, Aosong Cheng, Ming Lu, Renrui Zhang, Zhiyong Zhuo, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20857–20867. IEEE, 2025a.

Shaolei Zhang, Qingkai Fang, Yang Yang, and Yang Feng. Llava-mini: Efficient image and video large multimodal models with one vision token. In *International Conference on Learning Representations*, pp. 53285–53310, 2025b.

Xuechen Zhang, Xiangyu Chang, Mingchen Li, Amit Roy-Chowdhury, Jiasi Chen, and Samet Oymak. Selective attention: Enhancing transformer through principled context control. In *Advances in Neural Information Processing Systems*, volume 37, pp. 11061–11086, 2024.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. Sparsevlm: Visual token sparsification for efficient vision-language model inference. In *International Conference on Machine Learning (ICML)*, pp. 1–18, 2025c.

Kai Zhao, Wubang Yuan, Yuchen Lin, Liting Ruan, Xiaofeng Lu, Deng-Ping Fan, Ming-Ming Cheng, and Dan Zeng. Attention debiasing for token pruning in vision language models. *arXiv preprint arXiv:2508.17807*, 2025.

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 15

B Proofs of the main theorems 15

C Identification, tails, and extra theory 18

D Method map statements and proofs 20

E Position effects and duration 21

F Experimental protocols 21

G Additional experimental results 22

A NOTATION

Table 2 records the symbols used in the theorems and the experiments. Softmax is shift invariant, so only Δ and the two scales enter the laws, and the query in the theorems is the position that writes the first answer token. Prompt final scores are kept only as a diagnostic comparison. Asymptotics are taken as $N_v \to \infty$ with m, Δ, and the scales held fixed. The two population split is the modeling choice that makes the two scales visible, whereas a one population Gaussian field has a single extreme value law and no distinguished evidence mass, and therefore does not produce a diluted yet correct window.

Table 2: Symbols used in the paper.

Symbol	Meaning
N_v, N_t	visual and text token counts
$\mathcal{S}$, m	evidence set and its size
$\mathcal{V}$	visual noise set
s_i, α_i	logit and softmax weight of token i
Δ, $c = \Delta/\sigma_V$	evidence margin and locked ratio
σ_S, σ_V	signal and noise scales
μ_V	shared location, cancelled by softmax
A, B, $\eta = A/(A+B)$	signal mass, noise mass, retention
$\hat{A}$	normalized signal mass
$\bar{W}$, κ	signal log normal mean and Jensen gap
M_{ann}, M_{typ}	annealed and typical mass budgets
$N_v^{\star\mathrm{top1}}$, $N_v^{\star\dagger\mathrm{top1}}$	leading and second order retrieval budgets
$N_v^{\star\mathrm{mass,typ}}$	typical mass half retention budget
b_n, a_n	Gaussian extreme value centering and scaling
G	standard Gumbel
$\bar{\Phi}$	standard normal tail
B, $N_v^{\mathrm{eff}} = N_v/B$	block size and effective count
ℓ	spatial correlation length in patch units
C_k, K	merge cluster and keep budget
$\tilde{s}_k$, $\tilde{v}_k$	exact merge logit and value
$\hat{s}_k$, $\hat{v}_k$	scheme logit and value
O, O'	attention outputs before and after a merge
T, n_0, λ	turns, tokens per turn, positional decay
W	Lambert W function

B PROOFS OF THE MAIN THEOREMS

This section proves Lemma 1 and Theorems 1, 2, and 3, together with the spatial lemmas and Berman's condition for weakly dependent Gaussian fields. The finite sample Chebyshev bound is included, as is the comparison between self averaging and freezing that was deferred from Section 3, and both sit inside the proof of Theorem 1.

B.1 LEMMA 1

Proof. $\mathbb{E}e^{\sigma\zeta} = e^{\sigma^2/2}$ and $\mathbb{E}e^{2\sigma\zeta} = e^{2\sigma^2}$ give the mean and variance of $e^{\mu_V + \sigma_V \zeta_j}$. Independence divides the variance by N_v. Chebyshev's inequality concludes. □

B.2 SOFTMAX LIPSCHITZ

Lemma 4. *For $p = \mathrm{softmax}(x)$ and $q = \mathrm{softmax}(y)$, $\|p - q\|_1 \le \|x - y\|_\infty$ and the symmetrized KL divergence satisfies $\mathrm{KL}(p\|q) + \mathrm{KL}(q\|p) \le \|x - y\|_\infty^2$.*

Proof. $\mathrm{KL}(p\|q) = \langle p, x - y\rangle + \log Z_y - \log Z_x$. Symmetrizing, $\mathrm{symKL} = \langle p - q, x - y\rangle \le \|p - q\|_1 \|x - y\|_\infty$. Pinsker's inequality in both directions gives $\mathrm{symKL} \ge \|p - q\|_1^2$. Dividing yields $\|p - q\|_1 \le \|x - y\|_\infty$, and substituting back yields the second claim. □

B.3 THEOREM 1

Proof. Relative error. Condition on the signal logits, hence on A. By Lemma 1, $B/N_v \xrightarrow{p} e^{\mu_V+\sigma_V^2/2}$ in L^2, so

$$\eta = \frac{\hat{A}}{\hat{A} + B\, e^{-(\mu_V+\sigma_V^2/2)}}\,, \qquad B\, e^{-(\mu_V+\sigma_V^2/2)} = N_v\,(1+r_n)\,, \qquad r_n \xrightarrow{p} 0\,. \tag{9}$$

Then $\eta/(\hat{A}/(\hat{A}+N_v)) = (\hat{A}+N_v)/(\hat{A}+N_v(1+r_n)) \xrightarrow{p} 1$ by the continuous mapping theorem, which is the relative error claim. For the explicit bound, let $\mu_V = 0$ without loss, since it cancels in $\hat{A}$. Lemma 1 gives $|B/N_v - e^{\sigma_V^2/2}| \le \sqrt{(e^{2\sigma_V^2} - e^{\sigma_V^2})/(N_v\delta)}$. The relative fluctuation is $r_n = Be^{-\sigma_V^2/2}/N_v - 1$, hence $|r_n| \le e^{-\sigma_V^2/2}\sqrt{(e^{2\sigma_V^2} - e^{\sigma_V^2})/(N_v\delta)} = \sqrt{(e^{\sigma_V^2}-1)/(N_v\delta)}$. The map $g(b) = \hat{A}/(\hat{A}+b)$ has $|g'(b)| = \hat{A}/(\hat{A}+b)^2$. On $\{|r_n| \le 1/2\}$ one has $\hat{A} + N_v(1+r_n) \ge (\hat{A}+N_v)/2$, and $|\eta - \hat{A}/(\hat{A}+N_v)| = |g(N_v(1+r_n)) - g(N_v)| \le \frac{N_v\hat{A}}{(\hat{A}+N_v)^2}|r_n|$. Substituting the Chebyshev bound on $|r_n|$ yields the stated guarantee. There is no extra factor $e^{-\sigma_V^2/2}$ beyond the conversion from absolute to relative fluctuations, which already cancels.

Typical versus annealed prefactor. $\log\hat{A} = \log m + \Delta - \sigma_V^2/2 + \log\bar{W}$ is immediate. $\log\mathbb{E}\bar{W} = \sigma_S^2/2$ since $\mathbb{E}e^{\sigma_S\xi} = e^{\sigma_S^2/2}$. Jensen gives $\mathbb{E}\log\bar{W} \le \log\mathbb{E}\bar{W}$, so $\kappa \ge 0$. In the self averaging regime, $\mathrm{Var}(\bar{W}) = (e^{2\sigma_S^2} - e^{\sigma_S^2})/m$. If $\sigma_S^2 \ll 2\log m$ then $\bar{W} \xrightarrow{p} e^{\sigma_S^2/2}$ in relative error, $\log\bar{W} \to_p \sigma_S^2/2$, and $\kappa \to 0$. In the freezing regime, under REM scaling, $\log\sum_i e^{\sigma_S\xi_i}$ transitions from the annealed value to the extremal value $\sigma_S b_m$. Using $\log\bar{W} \ge \sigma_S\max_i\xi_i - \log m$ and $\mathbb{E}\max_i\xi_i = b_m + o(1)$ gives $\kappa \to \frac{1}{2}(\sigma_S - \sqrt{2\log m})_+^2$. The median statement follows because $\log\bar{W}$ concentrates around $\mathbb{E}\log\bar{W}$ at rate $m^{-1/2}$ in the self averaging regime and is dominated by the extreme value otherwise. The quantity used in the typical law is the numerical moment $\kappa = \sigma_S^2/2 - \mathbb{E}\log\bar{W}$, not the infinite m REM formula. At the operating point $(m, \sigma_S) = (16, 3.5)$ that moment equals 2.24, while the leading REM expression equals 0.66 and a second order Gumbel expression equals 3.05.

Budgets and tail. $\eta(N_v) = 1/2$ if and only if $\hat{A} = N_v(1 + o_p(1))$ from the relative error claim. The two budgets substitute the two values of the prefactor. The $\Theta(1/N_v)$ tail is the functional form $\hat{A}/(\hat{A}+N_v)$. □

Remark 1 (Freezing correction). *If $\sigma_V = \beta\sqrt{2\log N_v}$ with β fixed, then $\log B = \log N_v + \sigma_V^2/2 + o_p(\log N_v)$ for $\beta < 1$ and $\log B = \sigma_V b_{N_v} + O_p(1)$ for $\beta \ge 1$, where b_{N_v} is the Gaussian extreme value centering. The dilution law then uses the extremal mass on the frozen branch, $\bar{\eta} \asymp Me^{\mu_V}/(Me^{\mu_V} + e^{\sigma_V\sqrt{2\log N_v}})$. These are standard REM facts. The annealed branch is Lemma 1. The frozen branch is Fisher-Tippett-Gnedenko together with the fact that in the frozen phase the sum is carried by $O(1)$ extreme terms.*

B.4 THEOREM 2

Proof. Let $M_{n-1} = \max_j\zeta_j$ over noise, with $n = N_v$. By Fisher-Tippett-Gnedenko, $G_n := a_n^{-1}(M_{n-1} - b_n) \Rightarrow G$, a standard Gumbel. Then, with ζ^* the needle noise and $\sigma_S = \sigma_V$,

$$p_{\mathrm{hit}} = \Pr\big[\Delta + \sigma_V\zeta^* > \sigma_V(b_n + a_nG_n)\big] = \mathbb{E}\,\bar{\Phi}\big(b_n + a_nG_n - \Delta/\sigma_V\big)\,. \tag{10}$$

Slutsky's theorem and bounded convergence, using that $\bar{\Phi}$ is bounded and continuous, replace G_n by G at $o(1)$ cost and give equation 4. Since $a_n \to 0$ and $\bar{\Phi}$ is Lipschitz, $p_{\mathrm{hit}} = \bar{\Phi}(b_n - \Delta/\sigma_V) + o(1)$. With $b_n = \sqrt{2\log n} + o(1)$ this is equation 5.

Budgets. The b_n centered approximation solves $b_{n^\star} = c$ with $c = \Delta/\sigma_V$. This is not the exact finite n half crossing of p_{hit}, nor the Gumbel integral crossing of equation 4. The term a_nG contributes a Jensen shift of order a_n^2. Let $x = \sqrt{2\log n}$, so $b_n = x - \frac{\log(x\sqrt{2\pi})}{x} + O(x^{-3})$. One fixed point iteration of $x = c + \frac{\log(x\sqrt{2\pi})}{x}$ initialized at $x_0 = c$ gives $x_1 = c + \frac{\log(c\sqrt{2\pi})}{c}$. Hence $\log N_v^{\star\dagger\mathrm{top1}} = x_1^2/2 = c^2/2 + \log(c\sqrt{2\pi}) + o(1)$, which is equation 6.

Width. Differentiating $\bar{\Phi}(\sqrt{2\log n}-c)$ at $n = N_v^{\star\text{top1}}$ gives slope $-\phi(0)\,\sigma_V/\Delta = -(\sigma_V/\Delta)/\sqrt{2\pi}$, with ϕ the standard normal density, so the $(1-\epsilon) \to \epsilon$ window has width $\Theta(\sigma_V/\Delta)$ relative to $\log N_v^{\star\text{top1}}$. The Gumbel fluctuation $a_n G$ contributes an absolute width of order $a_{N_v^{\star\text{top1}}}\sigma_V = \sigma_V^2/\Delta$. □

B.5 Spatial lemmas

Proof of Lemma 2. The block maximum equals the maximum of the N_v/B i.i.d. representatives. Substitute $n \leftarrow N_v/B$ in Theorem 2. □

Proof of Lemma 3. With $B_{\text{noise}} = \sum_{k=1}^{N_v/B} B\, e^{\mu_V+\sigma_V\zeta_k}$ and block representatives ζ_k i.i.d., one has $B_{\text{noise}}/N_v = \frac{1}{N_v/B}\sum_k e^{\mu_V+\sigma_V\zeta_k} \xrightarrow{p} e^{\mu_V+\sigma_V^2/2}$, the same limit as the i.i.d. case. The annealed dilution law is therefore unchanged. The variance of B_{noise}/N_v is the i.i.d. variance times B, since only N_v/B independent terms remain. In the freezing regime, $\log B_{\text{noise}} \approx \log B + \sigma_V b_{N_v/B}$, so the extremal lift is computed at the effective count. □

B.6 Stationary Gaussian fields

Proposition 1 (Stationary Gaussian fields)**.** *If $\{\zeta_j\}$ is a stationary Gaussian sequence whose correlation $r(t)$ satisfies Berman's condition $r(t)\log t \to 0$, then $a_{N_v}^{-1}(M_{N_v} - b_{N_v}) \Rightarrow \text{Gumbel}$ with the same constants as the i.i.d. case. Weak spatial mixing leaves the leading order retrieval threshold unchanged and enters only as a finite N_v correction.*

Let $\{\zeta_j\}$ be stationary Gaussian, unit variance, $r(t) = \mathbb{E}[\zeta_1\zeta_{1+t}]$, and $u_n = b_n + a_n x$. By the normal comparison lemma,

$$\left|\Pr(M_n \le u_n) - \Phi(u_n)^n\right| \le K\sum_{t=1}^{n}|r(t)|\exp\Big(-\frac{u_n^2}{1+|r(t)|}\Big). \tag{11}$$

Split at t_0 with $|r(t)| \le \epsilon$ for $t > t_0$. The head contributes $O(t_0 e^{-u_n^2/(1+\rho)})$ with $\rho = \max_{t\le t_0}|r(t)| < 1$, which is $o(1)$ since $u_n^2 \approx 2\log n$. The tail contributes $O(n\epsilon e^{-u_n^2/(1+\epsilon)}) = O(\epsilon n^{1-2/(1+\epsilon)+o(1)}) = o(1)$ for ϵ small, and Berman's condition lets $\epsilon \to 0$ along $t_0 \to \infty$. Hence $\Pr(M_n \le u_n) - \Phi(u_n)^n \to 0$ and the Gumbel limit is unchanged. Block correlation of Lemma 2 is the opposite extreme, perfect dependence inside blocks of size B, and is not covered by Berman. That case rescales the retrieval threshold by $N_v^{\text{eff}} = N_v/B$ and leaves the mass law invariant, as proved for Lemma 2 above.

B.7 Theorem 3

Proof. Let $\tilde{s}_k = \log\sum_{j\in C_k} e^{s_j}$ and $\tilde{v}_k = \sum_{j\in C_k} e^{s_j}v_j/e^{\tilde{s}_k}$ be the exact merge, and let α over all tokens, β over the exact merge items, and β' the scheme distribution be the three attention distributions.

Exactness. By the distributive law, $\sum_{j\in C_k}\alpha_j v_j = \tilde{\alpha}_k\tilde{v}_k$, hence $O = O_\beta$ identically. No covering condition is needed. Value spread inside a cluster is generically nonzero, so the equality does not require $\delta_v = 0$.

Distribution perturbation. By Lemma 4 and logit fidelity, $\|\beta - \beta'\|_1 \le \epsilon_s$, hence $\|O_\beta - \mathbb{E}_{\beta'}[\tilde{v}]\| = \|\mathbb{E}_\beta[\tilde{v}] - \mathbb{E}_{\beta'}[\tilde{v}]\| \le V\|\beta-\beta'\|_1 \le V\epsilon_s$, using Hölder with $\|\tilde{v}_k\| \le V$. Convex combinations preserve the bound.

Representative values. The two expectations use the same weights β' but different values $\tilde{v}$ versus $\hat{v}$, so the difference is not identically zero. $\|\mathbb{E}_{\beta'}[\tilde{v}] - \mathbb{E}_{\beta'}[\hat{v}]\| \le \max_k\|\tilde{v}_k - \hat{v}_k\| =: \delta_v'$. If also $\max_k\max_{j\in C_k}\|v_j - \hat{v}_k\| \le \delta_v$, then $\delta_v' \le \delta_v$ because $\tilde{v}_k$ is a convex combination of $\{v_j\}_{j\in C_k}$. Tokens outside the visual set appear in O and O' with identical logits and values. Their contribution is charged only through the softmax redistribution already bounded by $V\epsilon_s$. The triangle inequality concludes. □

C IDENTIFICATION, TAILS, AND EXTRA THEORY

The extensions below were deferred from Section 3. They record the m signal retrieval lift, the three half crossings of Theorem 2, the case of unequal signal and noise scales, bounds on the Jensen gap, the window condition, and companions for regularly varying noise.

C.1 m SIGNAL RETRIEVAL LIFT

If the answer may use any of m signal tokens, the signal maximum contributes an extremal lift $\sigma_S\sqrt{2\log m}$, and the threshold becomes

$$N_v^{\star\mathrm{top1}} = \exp\Big(\frac{(\Delta + \sigma_S\sqrt{2\log m})^2}{2\sigma_V^2}\Big). \tag{12}$$

Redundancy therefore raises the retrieval budget multiplicatively in log scale, while it raises the mass budget only linearly by Theorem 1. A finite m lift replaces Δ by $\Delta + \sigma_S\sqrt{2\log m}$ in the retrieval law and replaces $\hat{A}$ by a sum of m log normals in the mass law. The two corrections act on different statistics. A caption query with many contributing patches can therefore be mass limited, while a referring query with the same total margin remains retrieval limited.

C.2 THREE HALF CROSSINGS

Three distinct crossings are easy to confuse, and Table 3 separates them over $c \in [2, 5.5]$. The exact finite N_v crossing solves $p_{\mathrm{hit}}(N_v) = 1/2$ for the true Gaussian probability, either by Monte Carlo or by quadrature. The Gumbel integral crossing solves $\mathbb{E}_G\bar{\Phi}(b_n + a_nG - c) = 1/2$ and is the $n \to \infty$ limit of that exact crossing. The b_n centered approximation instead sets $b_n = c$ and ignores the Jensen shift of a_nG, and the leading closed form $N_v^{\star\mathrm{top1}} = \exp(\Delta^2/2\sigma_V^2)$ is this third object under $b_n \sim \sqrt{2\log n}$. equation 6 is one fixed point iteration of $b_n = c$, with a residue of order a_n^2. The leading budget $e^{c^2/2}$ therefore understates the token count at which a needle remains recoverable.

Table 3: Half crossing error in units of $\log N_v$ for the leading budget $e^{c^2/2}$ and the second order budget $\sqrt{2\pi}\,c\,e^{c^2/2}$, against exact quadrature.

$c = \Delta/\sigma$	exact $n_{1/2}$	leading err.	second order err.
2.0	29	1.36	0.26
2.5	1.0×10^2	1.49	0.35
3.0	4.5×10^2	1.61	0.40
3.5	2.6×10^3	1.73	0.44
4.0	1.9×10^4	1.84	0.47
4.5	1.7×10^5	1.94	0.49
5.0	2.0×10^6	2.03	0.50
5.5	3.1×10^7	2.11	0.51

C.3 UNEQUAL SIGNAL AND NOISE SCALES

When $\sigma_S \neq \sigma_V$, the needle is $\Delta + \sigma_S\zeta^*$ and it competes with a noise maximum $\sigma_V(b_n + a_nG_n)$. The hit probability is then

$$p_{\mathrm{hit}} = \mathbb{E}\,\bar{\Phi}\Big(\frac{\sigma_V}{\sigma_S}(b_n + a_nG_n) - \frac{\Delta}{\sigma_S}\Big) + o(1). \tag{13}$$

The leading budget still solves $\sigma_V\sqrt{2\log n} = \Delta$, so the leading order is unchanged, while the second order term inherits the factor σ_V/σ_S in the Gumbel width.

C.4 JENSEN GAP BOUNDS

The gap κ admits explicit bounds in both regimes of the signal average. Let $\bar{W} = m^{-1}\sum_{i=1}^m e^{\sigma_S\xi_i}$, so that $\mathrm{Var}(\bar{W}) = (e^{2\sigma_S^2} - e^{\sigma_S^2})/m$. Chebyshev plus the map $x \mapsto \log x$ on $\{x \geq e^{\sigma_S^2/2}/2\}$ yields,

in the self averaging regime $\sigma_S^2 \ll 2\log m$,

$$\kappa(m, \sigma_S) = O\big((e^{2\sigma_S^2} - e^{\sigma_S^2})/(me^{\sigma_S^2})\big)\,. \tag{14}$$

In the freezing regime a matching lower bound follows from $\log \bar{W} \geq \sigma_S \max_i \xi_i - \log m$ and $\mathbb{E}\max_i \xi_i = b_m + o(1)$, giving $\kappa \to \frac{1}{2}(\sigma_S - \sqrt{2\log m})_+^2$. The operating point comparison is in Table 4. Using the infinite m REM formula at $(m, \sigma_S) = (16, 3.5)$ would overstate typical retention by a factor of about $e^{2.24-0.66} \approx 4.8$.

Table 4: Finite m Jensen gap against the leading REM formula and a second order Gumbel expression, at $m = 16$.

σ_S	κ (Monte Carlo)	leading REM	Gumbel finite m	$e^{-\kappa}$
2.0	0.40	0.00	1.33	0.67
3.5	2.24	0.66	3.05	0.11

C.5 When the mass budget precedes the retrieval budget

Taking logarithms in $N_v^{\star\mathrm{mass,typ}} < N_v^{\star\dagger\mathrm{top1}}$ and substituting $N_v^{\star\mathrm{mass,typ}} = m\exp(\Delta - \sigma_V^2/2 + \mathbb{E}\log\bar{W})$ and $N_v^{\star\dagger\mathrm{top1}} = \sqrt{2\pi}\,c\,e^{c^2/2}$ with $c = \Delta/\sigma_V$ yields equation 7. If $\mathbb{E}\log\bar{W} = 0$ and $\sigma_S = \sigma_V$, the inequality reduces to $\log m < c^2/2 + \log(c\sqrt{2\pi}) - c\sigma_V + \sigma_V^2/2$. For the landscape parameters $(m, \sigma_V) = (16, 1.5)$ this first holds near $c \approx 3.5$. For $m = 100$ and $\Delta = \sigma_V = 1$, the leading retrieval scale is $e^{1/2}$ while the mass scale is $100e$, so retrieval fails first. On the CLIP proxy in Appendix G, all 24 queries satisfy $N_v^{\star\mathrm{mass}} < N_v^{\star\mathrm{top1}}$ at the $m = 5$ lift used there. The window is therefore a property of the query rather than a universal phase of token inflation.

C.6 Regularly varying companions

Proposition 2 (Regularly varying noise)**.** *Suppose the noise upper tail satisfies* $\Pr(\zeta > x) \sim L(x)\,x^{-\alpha}$ *as* $x \to \infty$*, with* $\alpha > 0$ *and* L *slowly varying. Then the noise maximum belongs to the Fréchet domain, and the retrieval wall for a fixed needle scales as* $n_{\mathrm{RV}}^\star \asymp (\Delta/\sigma_V)^\alpha / L(\Delta/\sigma_V)$ *rather than* $\exp(c^2/2)$*. Student* t_ν *has tail index* ν *and is the computable intermediate used in the overlays.*

Heavier tails move the wall to the left of the Gumbel prediction, and Figure 6 holds $c = 3.2$ fixed while varying only the tail. At that margin the Gumbel wall is 1342, the exact Gaussian crossing is 882, a t_5 wall sits near 50, and the CLIP empirical tail crosses near 860. The Hill estimator of the tail index on CLIP noise has median 4.69. The Gumbel budget remains the main statement because CLIP noise and the trained layer are close to Gaussian in the bulk, so the heavy tail formula is a robustness interval beside that budget. Fitting c to a hit curve would fold a wrong tail into the margin and would no longer test the extreme value law. Locking c from logits instead leaves that error visible as a horizontal shift of the predicted wall, which is the design of the overlays in Figure 4.

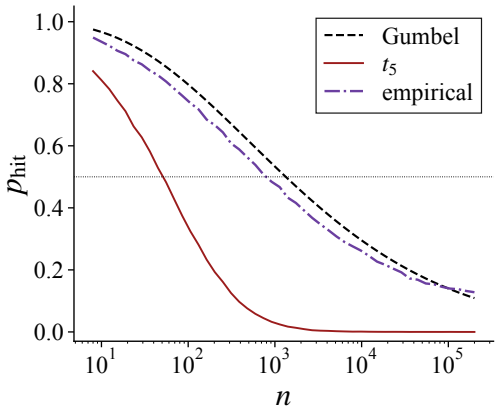


Figure 6: Same $c = 3.2$ under Gumbel, Student t_5, and the CLIP empirical tail. Heavier tails collapse sooner.

D METHOD MAP STATEMENTS AND PROOFS

The three propositions below describe the failure region of each method family through its selection rule, for one query and one frozen attention operation. They do not describe layer by layer benchmark behavior of a full system. Covering numbers for clustered visual values are recorded after the proofs.

Proposition 3 (Attention score pruning)**.** *Keeping the K largest logits, with centering $\tau_K := \sigma_V b_{N_v/K}$, the threshold fluctuates on the scale $\sigma_V a_{N_v/K}$, which the displayed boundary does not propagate. If $\Delta \geq \tau_K + \sigma_S \bar{\Phi}^{-1}(\delta/m)$, then all m signal tokens survive with probability at least $1-\delta$. If $\Delta \leq \tau_K - \sigma_S \bar{\Phi}^{-1}(\delta)$, the expected number of surviving signals is at most δ, and the retained visual set is an exchangeable extremal sample of noise logits. Exchangeability of logit ranks does not by itself control the covering radius of the associated values. If in addition the noise values are i.i.d. from a distribution whose $(\delta_v/2)$ covering number exceeds K, a size K sample fails to cover with high probability, and Theorem 3 is then unavailable as a covering guarantee.*

Proposition 4 (Clustering merges)**.** *Merging in the style of k-means satisfies the covering half of Theorem 3 whenever $K \geq \mathcal{N}(\delta_v/2)$, independently of the SNR. Its binding error term is logit fidelity. Centroid keys systematically underestimate cluster mass by about $\log|C_k|$. Adding that correction is exact when all logits in C_k are equal and is otherwise the first order approximation to $\tilde{s}_k$.*

Proposition 5 (Principled subset selection)**.** *Task relevance, information coverage, and semantic diversity correspond to Δ weighting from Theorem 1, the covering condition of Theorem 3, and covering number minimization from Corollary 3. The failure region is the large σ_V regime, where relevance estimates are corrupted by extremal noise, which is the same mechanism as Proposition 3.*

Proof of Proposition 3. The top K threshold is the $(N_v - K)$-th order statistic of the pooled logits, which concentrates at $\tau_K = \sigma_V b_{N_v/K}$ with fluctuations of order $\sigma_V a_{N_v/K}$. Each signal logit is $\mu_V + \Delta + \sigma_S \xi_i$. Then $\Pr[s_i < \tau_K] \leq \bar{\Phi}((\Delta - \tau_K)/\sigma_S) \leq \delta/m$ under the stated above threshold condition, and a union bound over the m signals gives survival with probability at least $1-\delta$. Under the below threshold condition, each signal survives with probability at most δ, so the expected survivor count is at most δ. Conditional on the survival pattern the retained logits are an exchangeable extremal sample of noise. The bound controls ranks. A covering statement for the values needs a separate assumption. If the noise values are i.i.d. from a law whose $(\delta_v/2)$ covering number exceeds K, a standard occupancy argument gives covering failure with high probability, and the covering half of Theorem 3 is then unavailable. Without an assumption that separates keys from values, covering failure does not follow from rank exchangeability. □

D.1 COVERING NUMBERS FOR CLUSTERED VISUAL VALUES

Suppose r objects have values in Euclidean balls of radius R in dimension ν. The ϵ covering number of the union is at most $r(1+2R/\epsilon)^\nu$. Corollary 3 then gives $K^\star(\epsilon) = \Theta(r(2R/\epsilon)^\nu)$, independent of

N_v. If each object occupies B perfectly correlated patches, the covering number does not grow with B. Redundancy inflates N_v but not $K^\star$. Combined with Theorem 1, any keep budget $K \le \hat{A}\frac{\eta_0}{1-\eta_0}$ restores retention to $\eta \ge \eta_0$ after an exact merge.

E Position effects and duration

These three propositions are the same logits read along position and along time. Recency skew of top K pruning, the reweighting that undoes it, and the Lambert W duration wall for a single instruction stretched over T turns are recorded here.

E.1 Recency and reweighting

Proposition 6 (Recency bias). *Under $\Delta(i) = \Delta_0 - \lambda d(i,q)$, top K pruning keeps a token at distance d with probability $p_{\text{keep}}(d) = \bar{\Phi}(b_{N_v/K} - \Delta_0/\sigma_V + \lambda d/\sigma_V) + o(1)$.*

Proof. Token i is kept iff s_i exceeds the extremal threshold $t_K \approx \sigma_V b_{N_v/K}$. Substituting $s_i = \Delta_0 - \lambda d + \sigma_V \zeta_i$ gives the formula. □

Proposition 7 (Reweighting). *Adjusted logits $\tilde{s}_i = s_i + \hat{\lambda} d(i,q)$ rank tokens by true content signal iff $\hat{\lambda} = \lambda^*$. Residual bias is proportional to $(\hat{\lambda} - \lambda^*)d$.*

Proof. $\tilde{s}_i = \Delta_0 + (\hat{\lambda} - \lambda^*)d_i + \sigma_V \zeta_i$. The position dependent term vanishes if and only if $\hat{\lambda} = \lambda^*$. □

E.2 Cumulative dilution

Proposition 8 (Cumulative dilution). *With T turns adding n_0 visual tokens each and margin decay $\Delta_T = \Delta_0 - \lambda(T-1)$ for a single instruction issued at turn 1, $\eta_T = \frac{me^{\Delta_0 - \lambda(T-1)}}{me^{\Delta_0 - \lambda(T-1)} + Tn_0} + o_p(1)$. If $\lambda = 0$, then $\eta_T = \Theta(1/T)$ with half retention wall $T^\star = me^{\Delta_0}/n_0$. If $\lambda > 0$ the numerator decays exponentially in T, and the wall has the closed form $T^\star = \lambda^{-1} W(\lambda m n_0^{-1} e^{\Delta_0 + \lambda})$, since $\eta_T = 1/2$ if and only if $\lambda T e^{\lambda T} = (\lambda m/n_0) e^{\Delta_0 + \lambda}$. As $\lambda \to 0^+$, $W(z) \sim z$ recovers the algebraic wall. If each turn issues a fresh instruction, $\eta_T = \frac{me^{\Delta_0}}{me^{\Delta_0} + n_0}$ independently of T. For needle retrieval in video, Theorem 2 with $N_v = Tn_0$ gives $T^\star_{\text{top1}} = n_0^{-1} \exp(\Delta^2/2\sigma_V^2)$.*

The decaying wall $T^\star(\lambda) = \lambda^{-1} W(\lambda m n_0^{-1} e^{\Delta_0 + \lambda})$ satisfies $T^\star(\lambda) \to me^{\Delta_0}/n_0$ as $\lambda \to 0^+$, because $W(z) \sim z$ as $z \to 0$. Differentiating under the implicit relation $\lambda T e^{\lambda T} = (\lambda m/n_0) e^{\Delta_0 + \lambda}$ shows $dT^\star/d\lambda < 0$ for $\lambda > 0$. Faster positional decay strictly shortens the usable horizon. In Monte Carlo, a decaying instruction with $\lambda = 0.02$ hits the Lambert W crossing near $T = 8$, a persistent instruction hits the algebraic wall near $T = 9$, and a fresh instruction each turn does not accumulate.

F Experimental protocols

Monte Carlo checks and synthetic attention checks are run on CPU, and Figure 7 uses a frozen CLIP ViT-B/32 encoder on COCO. Decoder reads use the public Qwen3-VL-2B-Instruct, Qwen3-VL-4B-Instruct, Qwen3-VL-8B-Instruct, and LLaVA-1.5-7B weights, again with no gradient updates. Where a theoretical curve is an exact integral, it is evaluated by Gauss Hermite quadrature with 200 nodes, and reported bands are standard errors over independent trials on log spaced grids in N_v. Dilution uses $m = 16$, the pairs $(\Delta, \sigma_V) \in \{(4,2),(6,2)\}$, and the high dispersion point $\sigma_S = 3.5$, with $N_v \in [16, 2^{17}]$ and 64 trials. Retrieval uses $\sigma_V = 1$, signal to noise ratios in $\{2,3,4\}$, $N_v \in [16, 10^6]$, and 4000 trials, and the budget comparison covers $c \in [2, 5.5]$. Block checks use $\Delta = 4$, $\sigma_V = 1$, sizes $B \in \{1,4,16\}$, and 3000 trials. The stationary field is a 64×64 torus with correlation lengths $\ell \in \{0,2,4,8\}$ and 400 trials, while resolution uses a 24×24 base grid, content scale 1.5, and resolutions $r \in \{1,2,4\}$. Merge checks use a 64×64 scene, target radius 12, key dimension 64, value dimension 32, eight text tokens at margin 3, the pairs $(4.5, 1.0)$ and $(2.0, 1.5)$, $K \in [2^3, 2^{11}]$, 16 trials, and 12 iterations of k-means. Recency uses $n = 8192$,

$K = 256$, $\Delta_0 = 4$, $\sigma_V = 1$, decays in $\{0, 5\times10^{-4}, 2\times10^{-3}\}$, 300 trials, and 40 position bins. Duration adds $m = 8$ tokens on each turn from $n_0 = 128$, with $\Delta_0 = 5$, $\sigma_V = 1$, decay 0.02, and 200 trials.

For the retrieval Monte Carlo, the noise maximum is drawn from the inverse cumulative distribution function $M_{N_v} \overset{d}{=} \Phi^{-1}(U^{1/N_v})$ with U uniform on $(0, 1)$, which is exact under the i.i.d. model. The hit probability in Theorem 2 is evaluated by the same Gauss Hermite rule, and doubling the node count from 200 changes values by less than 10^{-6}. The trained layer is a single head with input dimension 16, key dimension 8, and needle gain 4. It uses analytic gradients, checked against central finite differences to a maximum discrepancy of 10^{-10} before training. Training then runs Adam with moments $(0.9, 0.999)$, learning rate 0.03, batch size 64, and 3000 steps, and each grid point is evaluated on 600 scenes. After training, the fitted pair is $(\hat{\Delta}, \hat{\sigma}) = (8.45, 2.13)$, and the absolute hit error against Theorem 2 is at most 0.024.

CLIP patch features are taken after the final layer norm and the multimodal projection. Each image contributes 49 patches of dimension 512, normalized in ℓ_2 and scored as $\tau \cdot \cos(t, v_j)$ with the model temperature $\tau = 100$. All 160 images are center cropped to 224^2 with the standard CLIP normalization. The needle image for a query maximizes $\langle t, g_i \rangle$ over global image embeddings, and the signal patches are the $m = 5$ largest entries of the patch text scores of that image. This choice of m is a fixed operating point. Noise images are those whose global similarity lies below the median. The overlays in the main text use the full bank of 160 images. A split into 120 and 40 images reestimates (Δ, σ_V) on the training images and evaluates inflation on held out noise images, and uncertainty across prompts is a 2000 draw bootstrap of the 24 queries. Retrieval budgets larger than 3136 in Table 7 are extrapolations of the fitted pair rather than measured crossings.

Qwen3-VL-2B-Instruct, Qwen3-VL-4B-Instruct, and Qwen3-VL-8B-Instruct are run frozen in half precision. Each image is resized so that its longest side is 224, and the prompt places a needle image before distractor images and then asks which image contains the referent. Both the needle and the distractors come from the CLIP bank. Last layer attention is averaged over heads and read from the first generated token onto vision tokens, after a teacher forced prefix. A hit means that the argmax vision token lies in the needle image span, and mass is the share of vision attention on that span, so the probe is retrieval at the image level. The needle is shuffled among nearby distractors, and generation uses greedy decoding of at most twelve new tokens. Dropped visual positions are zeroed in the first decoder layer, so that rotary position embeddings and the vision grid stay aligned. The oracle path writes a softmax weighted mean into one representative and zeros the rest. That path implements the value half of Theorem 3 under a length constraint, and it is an exact merge only for the attended visual block. The margin c is the median ratio of logit location to scale on the estimation split, at distractor counts that still contain distractors, and it is not fitted to hit events or to generation accuracy. Prompt final estimates of c stay within 0.1 of the generation token values. LLaVA-1.5-7B takes a single image, so a collage of $K + 1$ tiles holds 576 tokens fixed and varies the needle share under the same locked margin protocol. When methods are compared, generation accuracy is pooled over collage sizes.

G Additional experimental results

G.1 Dilution error

Table 5 reports the relative error of the median dilution curve against the annealed law and the typical law in the regime $\eta < 1/2$. At $(\Delta, \sigma_S) = (6, 2)$ the typical law tracks the median with relative error 2.3%. At $\sigma_S = 3.5$ the annealed law is off by 90%, while the typical law remains within 24%. Table 4 compares the Monte Carlo gap with the leading REM formula and the finite m Gumbel expression at the same signal scales.

Table 5: Relative errors of the median dilution curve against the annealed and typical laws in the diluted regime $\eta < 1/2$.

Δ	σ	κ	$e^{-\kappa}$	max rel. err. (ann.)	med. rel. err. (typ.)
4	2.0	0.40	0.67	0.42	0.12
4	3.5	2.24	0.11	0.90	0.24
6	2.0	0.40	0.67	0.32	0.023

G.2 Retrieval budgets

Table 3 is the numerical content of Figure 2(c). The leading budget $e^{c^2/2}$ misses the half crossing by 1.36 to 2.11 units of $\log N_v$, and one fixed point correction cuts that error to between 0.26 and 0.51 for $c \in [2, 5.5]$. At signal to noise ratio 2, collapse is already under way by a few hundred tokens. At signal to noise ratio 5, the hit probability stays near 1 out to 10^5 tokens and then falls with the predicted width $\Theta(\sigma_V/\Delta)$. Figure 6 isolates the same comparison at $c = 3.2$ by changing only the noise tail.

G.3 Spatial correlation and resolution

Figure 3 in Section 5 is the block test of Lemmas 2 and 3. As shown in Figure 3(a), retrieval moves by exactly B once the token count is replaced by $N_v^{\text{eff}} = N_v/B$. As shown in Figure 3(b), the mass curves stay on one annealed law, while the fluctuation bands widen as $\sqrt{B}$ because a block structure leaves only N_v/B independent noise terms. A Gaussian field on a 64×64 torus recovers an effective count linear in N_v/ℓ^2. Quadrupling the resolution at fixed physical content, from 576 to 2304 patches, leaves the hit probability at 0.884 and then 0.879, and a further step to 9216 patches stays at 0.880. The main text rounds these three readings to 0.88. They are the controlled field check of Corollary 2, and two images with the same N_v can sit on opposite sides of the retrieval wall when one is a fine grid over a small scene and the other is a collage of unrelated tiles.

G.4 Compression and the method map

Figure 5(a) and (b) compare reconstruction error on the two sides of the wall, and Table 6 lists retained mass on the full keep grid. On the synthetic scene with a ten patch target and keep budget 32, score pruning above the wall has output error 0.77 against 2.42 for random dropping, with retained mass 0.99 versus 0.48. A centroid merge without a mass correction keeps 0.46 of the mass. Adding $\log |C_k|$ restores 0.87 and cuts the error to 0.16, while the exact merge sits at numerical zero. Below the wall, the top K error rises to 1.36 and the corrected covering merge stays near 0.31, as predicted by Proposition 3. Every exact merge trial lies on the line $\|O - O'\| = 0$ of Theorem 3. Generation bars that are not in Figure 5 appear in Figure 8.

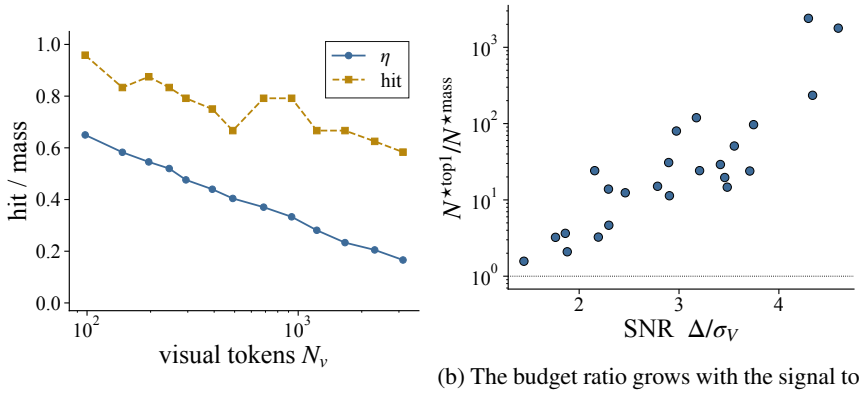
Table 6: Retained mass η' at budgets $K \in \{16, 32, 64, 128, 256\}$. Uncompressed η is 0.82 above the wall and 0.35 below it.

	above the wall					below the wall				
method	16	32	64	128	256	16	32	64	128	256
top K	0.98	0.99	0.96	0.91	0.88	0.67	0.51	0.63	0.53	0.40
random	0.24	0.48	0.57	0.62	0.76	0.06	0.08	0.26	0.23	0.22
k-means	0.32	0.46	0.52	0.62	0.66	0.03	0.04	0.10	0.14	0.15
k-means+	0.80	0.87	0.82	0.76	0.72	0.37	0.32	0.39	0.34	0.25
oracle	0.83	0.86	0.83	0.79	0.80	0.39	0.31	0.40	0.37	0.29

G.5 CLIP proxy and query taxonomy

Figure 7(a) is the inflation curve on the CLIP proxy. Hit rate falls from 0.96 to 0.58 and retention falls from 0.65 to 0.17 as N_v grows from 98 to 3136. The hit rate decays slowly because the needle is a region of m patches with residual spatial correlation, so an overlay that treats the needle

as one token underpredicts p_{hit}. Evidence mass tracks the annealed law up to a roughly constant factor. Figure 7(b) is the query taxonomy. Estimating (Δ, σ_V) per query with an $m = 5$ lift yields geometric mean budgets 325 and 8.1×10^3, a ratio of 25. High signal object queries such as cat and boat have fitted retrieval budgets in the millions, beyond the measured range $N_v \leq 3136$. Low signal scene queries such as clock on a wall and train on tracks have already collapsed after a few dozen tokens. All 24 queries satisfy equation 7 at $m = 5$.



(a) Hit rate and mass under token inflation.

(b) The budget ratio grows with the signal to noise ratio.

Figure 7: CLIP ViT-B/32 proxy on COCO. (a) Hit rate decays slowly because the needle is a region. (b) The budget ratio grows with the signal to noise ratio.

Table 7 records those fitted budgets query by query.

Table 7: Per query SNR and dual budgets on CLIP ViT-B/32 over COCO, sorted by SNR. Retrieval uses the $m = 5$ signal region lift. Values of $N^{\star\text{top1}}$ above 3136 are fitted extrapolations. All 24 rows satisfy $N^{\star\text{mass}} < N^{\star\text{top1}}$ at this m.

query	SNR	$N^{\star\text{mass}}$	$N^{\star\text{top1}}$	query	SNR	$N^{\star\text{mass}}$	$N^{\star\text{top1}}$
cat	4.59	4365	7.8×10^6	traffic light	2.90	292	3312
boat on the water	4.34	4257	1.0×10^6	vase with flowers	2.90	589	1.8×10^4
food on plate	4.30	1526	3.7×10^6	bird on a branch	2.78	320	4831
bed in a room	3.75	1404	1.4×10^5	car	2.46	101	1253
bowl of fruit	3.71	1527	3.7×10^4	motorcycle parked	2.30	82	384
suitcase on the ground	3.55	1260	6.4×10^4	airplane in the sky	2.29	103	1429
bench in a park	3.48	1289	1.9×10^4	skateboard on the street	2.19	64	208
dog	3.46	570	1.1×10^4	person holding an umbrella	2.16	221	5348
horse in a field	3.41	870	2.5×10^4	group of people	1.88	31	65
person riding a bicycle	3.21	571	1.4×10^4	sheep in a meadow	1.86	35	128
pizza on a table	3.17	734	8.8×10^4	train on tracks	1.76	27	88
truck on the road	2.97	384	3.1×10^4	clock on a wall	1.45	13	20

Table 1 in the main text records how the same queries fill the three regimes as N_v grows. At 576 tokens one third of the bank is diluted yet correct, with a prompt bootstrap interval $[0.25, 0.42]$. A split into 120 and 40 images keeps the median relative error of the training overlay near 0.22 on held out noise images. The pooled Kolmogorov Smirnov distance to a fitted Gaussian is 0.012, and the per query distances have median 0.016 and maximum 0.032. The median slope of the upper tail quantile plot is 1.00, and 10 of the 24 queries have upper tails heavier than Gaussian.

G.6 Decoder attention and generation

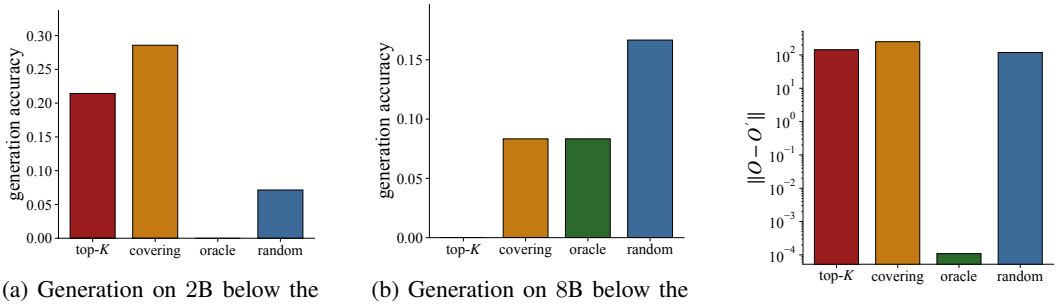
Table 8 records the locked margin and the observed half crossings for the frozen decoders in Figures 4 and 5. On Qwen3-VL-8B the logit estimate is $c = 2.72$ and $N_v^{\star\dagger\text{top1}} = 277$. On unseen prompts the generation token hit half crossing is 189, so $|\log 277 - \log 189| = 0.38$. Generation accuracy falls from 0.88 through 0.75 and 0.67 to 0.50 as N_v grows from 71 to 1223, while mass at the generated token is already 0.20 by 363 tokens. That lag is the diluted yet correct window on

the largest frozen decoder. On Qwen3-VL-4B the same protocol gives $c = 2.41$ and $N_v^{\star\dagger\text{top1}} = 110$ against a hit crossing of 211, so $|\log 110 - \log 211| = 0.65$, with a generation half crossing near 950. On the 2B model one has $c = 1.88$ and $N_v^{\star\dagger\text{top1}} = 27$ against a hit crossing of 211, and generation stays at or above one half through 651 tokens before it declines. The 8B overlay is the tightest hit prediction of the three.

Table 8: Locked logit margins and observed half crossings. LLaVA uses a fixed 576 tokens, so only the locked c is reported.

Model	c	$N_v^{\star\dagger\text{top1}}$	t_5	hit half	gen. half
Qwen3-VL-2B	1.88	27	7	211	363
Qwen3-VL-4B	2.41	110	11	211	950
Qwen3-VL-8B	2.72	277	14	189	1223
LLaVA-1.5-7B	2.34	90	collage	collage	collage

Oracle reconstruction error is 10^{-4} on Qwen3-VL-2B, 4×10^{-5} on Qwen3-VL-4B and on Qwen3-VL-8B, and 2.5×10^{-6} on LLaVA-1.5. Below the locked wall, generation accuracy is 0.29 for covering, 0.21 for top K, and 0.07 for random dropping on the 2B model. On the 4B model the same comparison is 0.24, 0.18, and 0.06. On the 8B model random dropping reaches 0.17 and covering reaches 0.08. Figure 8(a) and (b) show the 2B and 8B generation bars, and Figure 8(c) shows reconstruction error on the 2B model. Covering leads on 2B generation, while random dropping leads on 8B generation. The equality case of Theorem 3 is observed on the frozen Qwen decoders, and the winning heuristic on generation follows the model and the budget that binds.



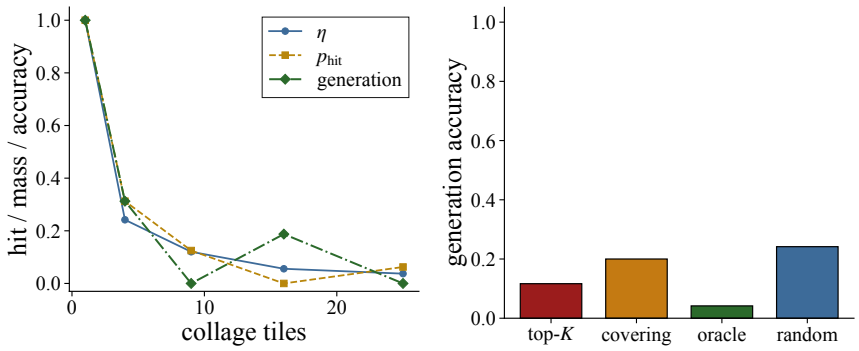
(a) Generation on 2B below the wall.

(b) Generation on 8B below the wall.

(c) Reconstruction error on 2B.

Figure 8: Frozen model bars. (a) Covering leads on 2B generation. (b) Random dropping leads on 8B generation. (c) Oracle reconstruction error on 2B is numerical zero.

LLaVA-1.5 takes one image, so a collage of $K+1$ tiles holds 576 tokens fixed and reduces the needle share. Generation accuracy falls from 1.00 at one tile to 0.31 at four tiles, and then reads 0.00, 0.19, and 0.00 at nine, sixteen, and twenty five tiles. Evidence mass tracks the needle fraction. Figure 9(a) is that collage overlay, and Figure 9(b) pools generation over collage sizes. Random dropping and covering both exceed score pruning once the needle share is small, which is the method map below the wall on this encoder. The pooled ranking is 0.24 for random dropping, 0.20 for covering, and 0.12 for top K.

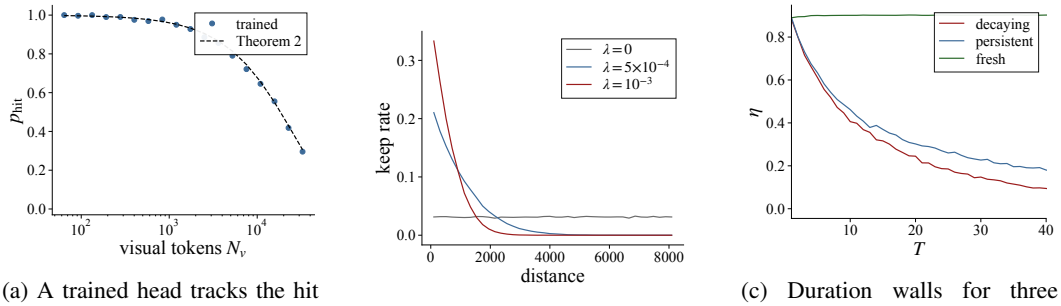


(a) Collage overlay at a fixed 576 tokens. (b) Generation pooled over collage sizes.

Figure 9: LLaVA-1.5 at fixed token count. (a) The needle share falls as tiles are added. (b) Random dropping and covering exceed score pruning once that share is small.

G.7 Recency, duration, and a trained layer

Figure 10(a) checks a trained layer against the retrieval law. The logits learned by stochastic gradient descent follow Assumption 1, and the induced hit rate tracks Theorem 2 at $(\hat{\Delta}, \hat{\sigma}) = (8.45, 2.13)$. Figure 10(b) shows that the top K keep rate falls with distance as Proposition 6 predicts, and reweighting restores a flat keep rate if and only if $\hat{\lambda} = \lambda^*$. Figure 10(c) is the duration wall. A decaying instruction hits the Lambert W crossing, a persistent instruction hits the algebraic wall me^{Δ_0}/n_0, and a fresh instruction on each turn does not accumulate.



(a) A trained head tracks the hit law.

(b) Keep rate falls with distance.

(c) Duration walls for three schedules.

Figure 10: Training and position checks. (a) A trained head tracks Theorem 2. (b) Keep rate falls with distance. (c) Duration walls for decaying, persistent, and fresh instructions.