A MEMORY THAT ONLY LOOKS LIKE ONE: A HYBRID VISION-LANGUAGE MODEL RELAYS THE PICTURE RATHER THAN STORING IT

Anonymous authors
Paper under double-blind review

ABSTRACT

A hybrid vision-language model carries two kinds of memory. Its attention layers keep a growing cache: every token the model reads, including every image patch, is stored as a key-value entry that later tokens can look up. Its recurrent layers keep a single state of fixed size, and every new token overwrites part of that state. Prior work proved that a fixed-size state must eventually lose information, but not *when* a given model forgets, *what* it keeps while forgetting, or *what breaks* downstream. We answer these three questions for Zamba2-VL-7B, whose 81 decoder layers all carry a Mamba2 mixer while 13 also carry attention. First, just after reading an image the state holds the gist, not the details: a linear probe reads which object was present at 92.5% on the best layer against 12.5% chance, but can barely read how many objects there were or where they sat. Second, how fast the state forgets can be computed from the weights alone: each head has a half-life of $\ln 2/(|A|\bar{\Delta})$ tokens, with medians of 6.6, 2.9 and 2.6 at three scales, nothing fitted. Third, and centrally, the memory a probe appears to find at distance is mostly not memory. With write amplitude held fixed, heads with a 2.3-token half-life still decode identity at 26.6% sixty-four tokens after the image, where their own parameters retain 4×10^{-9} of what was written. Evicting the image's key-value entries, without touching the recurrent state, drops the fast tercile from 51.6% to 7.8% at thirty-two tokens, below chance, while evicting an equal amount of pre-image text moves the readout the other way; at matched amplitude, eviction costs fast heads 2.4 times what it costs slow ones. Causally, exchanging all 81 recurrent states between runs that saw different images leaves 100 of 100 answers following the host's own image; exchanging seven of thirteen attention caches moves 94 to the donor's. The recurrent channel relays what attention holds; it does not store the picture. The redundancy premise behind visual token pruning survives down to a quarter of the entries, but the cost of pruning harder roughly doubles with distance from the image.

1 INTRODUCTION

A hybrid linear-attention model carries two kinds of memory at once. Most of its layers keep a recurrent state of fixed size; this is why the architecture promises constant-memory inference. A minority of layers keep ordinary attention caches, which grow with the context. Suppose such a model reads an image and answers a question about it thirty tokens later. Which of the two memories is the answer coming from?

Why does this question matter? The field is deploying these models on the assumption that the fixed state carries real work. Zamba2-VL-7B keeps $458{,}752$ numbers per layer across 81 layers, room enough to invite the intuition that a few hundred image tokens fit comfortably inside. This paper measures what is actually there. Our central finding is that the recurrent state is not a memory of the image but a relay of it: at distance, what a probe recovers is what the attention channel still holds and keeps re-supplying. The evidence is causal, and the direction was not the one we expected when we began.

Contributions.

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.

1. **A retention horizon computed from the weights, not fitted.** Mamba2 multiplies its state by $\exp(\Delta_t A)$ at every step, so a head's half-life is $\ln 2/(|A|\bar{\Delta})$; the median is 6.6, 2.9 and 2.6 tokens at three scales (Section 5). We also report where this prediction fails.
2. **The forgetting is content-selective.** At the instant of writing, object identity is highly decodable while count and position sit near chance (Section 4). The state keeps the gist and discards the particulars.
3. **The relay result.** With write gain held fixed, heads split by their own predicted half-life are ordered as decay predicts, yet decode identity far beyond what their own parameters permit. Evicting the attention channel's copy of the image collapses the fast heads' readout below chance; evicting an equal number of pre-image text positions moves the readout the other way, a double dissociation that rules out mere damage (Section 6).
4. **A dose-response predicted by the weights.** How much a head's readout depends on the attention channel follows that head's own decay rate: $2.4\times$ more for fast heads than slow at matched write amplitude, and $r = 0.80$ across ten strata on the uncontrolled split (Section 6.1).
5. **A causal localisation.** Exchanging all 81 recurrent states between runs that saw different images changes nothing. Exchanging seven of thirteen attention caches flips almost every answer (Section 7).
6. **A consequence for a deployed technique.** Visual key-value pruning is safe on this architecture down to a quarter of the entries. Below that, its cost grows with distance from the image, tripling between 0 and 512 tokens at 5% retention (Section 9).

Why this matters. Anyone deploying a hybrid must decide how much work the fixed state can be trusted with. This paper turns that judgement into a calculation: from a checkpoint's decay weight A and its gate statistics, the retention horizon can be computed before any experiment is run, and for this model it is single digits. The practical stake is visual token and key-value pruning (Chen et al., 2024; Yang et al., 2024; Shang et al., 2025; Zhang et al., 2023; Li et al., 2024; Xiao et al., 2024), which rests on the premise that most entries are redundant. The premise largely survives on this hybrid; the safety margin does not. The redundancy exists because the recurrent channel is relaying the picture, and it only relays for a few tokens, so a method tuned on short prompts looks safer than it is.

2 Related Work, and What Is New Here

Linear-attention and state-space sequence models replace the growing key-value cache of a transformer with a state of fixed size (Katharopoulos et al., 2020; Gu et al., 2022; Peng et al., 2023; Gu & Dao, 2023; Dao & Gu, 2024; Beck et al., 2024). A fixed state cannot grow with the context. For this reason most deployed systems keep a few attention layers alongside the recurrent ones, and the resulting hybrids are now common (Lieber et al., 2024; Glorioso et al., 2024; De et al., 2024; Ren et al., 2025; Waleffe et al., 2024).

We want to be clear about what is already known: the premise that a fixed-size state must eventually lose information from context is not ours. Jelassi et al. (2024) prove that transformers can copy exponentially long strings while generalized state space models are "fundamentally limited by their fixed-size latent state", and show that pretrained transformers "dramatically outperform state space models at copying and retrieving information from context". Other results reach the same conclusion from other directions. State-space models are limited to TC^0 and cannot reliably track state (Merrill et al., 2024). In-context retrieval is the specific skill that separates recurrent models from transformers (Wen et al., 2024). Most of the quality gap between efficient architectures and attention comes from recall (Arora et al., 2023; 2024). Our work builds on these results; it does not rediscover them.

What is new here is therefore not that a fixed state forgets, but when it forgets (a horizon computed rather than fitted), what survives (identity, not count or position), where the answer actually comes from (a causal localisation inside one model), and what that implies for a deployed technique. Table 7 in the appendix sets each item against the prior work it builds on.

Our main instrument is a linear probe, the simplest possible reader: a classifier that predicts the trial's answer from a weighted sum of a representation's numbers, trained on some trials and tested on held-out ones (Alain & Bengio, 2016). A probe can succeed because the classifier is powerful, not because the model uses the information (Hewitt & Liang, 2019; Belinkov, 2022), so we never rely

on a probe alone. Every probe result is paired with an intervention on the model, in the tradition of causal mediation analysis (Vig et al., 2020) and activation patching (Meng et al., 2022; Wang et al., 2022), and we report what each intervention does to the model's own behaviour, not only to the readout (Zhang & Nanda, 2024).

3 Model and Instruments

The model. Zamba2-VL-7B (Zyphra, 2026) is the vision-language member of the Zamba family (Glorioso et al., 2024), which pairs a Mamba2 backbone (Dao & Gu, 2024) with a shared attention module, and takes its vision encoder from Qwen2.5-VL (Bai et al., 2025). It is Apache-2.0 licensed and has 81 decoder layers. Every one carries a Mamba2 mixer, and 13 additionally carry a shared attention block, at layer indices 6, 11, 17, 23, 29, 35, 41, 47, 53, 59, 65, 71 and 77. Each layer's recurrent state has shape $[112, 64, 64]$, which is $458{,}752$ numbers. One image occupies between 345 and 415 tokens depending on its aspect ratio.

Reading the state. The Mamba2 recurrence is

$$S_t = S_{t-1} \exp(\Delta_t A) + \Delta_t \, x_t B_t^\top, \qquad y_t = S_t C_t, \tag{1}$$

so writing uses (x, B), reading uses C, and retention of anything already present is governed by A and Δ alone: the multiplication by $\exp(\Delta_t A)$, a number strictly between zero and one, is the only operation that removes information, and reading through C_t changes nothing. Each fact was verified against the installed model source rather than assumed: $A = -\exp(A_{\log})$ is a per-head scalar, the gate is $\Delta = \mathrm{softplus}(\Delta_{\text{raw}} + b)$ clamped below at 0.001 with no upper clamp, and the decay factor is $\exp(\Delta_t A)$ elementwise.

The state is too wide to probe directly, at $458{,}752$ numbers per layer, so each layer's state is passed through a fixed Gaussian random projection to 512 features. A Johnson-Lindenstrauss projection preserves linear separability up to a controlled distortion, so a linear probe on the projected features is, to that tolerance, a linear probe on the state. Probes are ridge classifiers evaluated by block-grouped k-fold cross-validation, with ridge penalties chosen on inner training folds only. One rule matters enough to state on its own. Eviction means deleting the image's key-value entries from the attention caches. Every probe is fitted and evaluated *within a single condition*: a probe reported for the evicted condition is trained only on evicted trials and tested on held-out evicted trials, never transferred from the intact condition. A drop under eviction therefore means less linearly decodable information in that condition, not a shift between training and test distributions. p-values are exact binomial tails against the 12.5% chance rate and intervals are Wilson intervals.

Splicing the channels. The splice runs the model twice, on two different images, and then physically exchanges one memory channel between the runs while leaving the other untouched. The run that keeps its own question is the host; the run that donates its memory is the donor. After the exchange the host answers its question, and we record whose image the answer describes. Swapping the recurrent channel replaces all 81 states and convolution buffers; swapping the attention channel replaces the 13 key and value caches. The host cache is restored exactly afterwards, so one prefill supports many conditions.

Materials. Trials are built from COCO val2017 instance annotations (Lin et al., 2014), the corpus underlying most visual instruction tuning (Liu et al., 2023), in three question families: which object is present, how many, and where. Every question is eight-way forced choice, so chance is exactly 12.5%. Trials come in blocks of eight sharing one candidate list, each candidate correct exactly once per block, which is what makes the pre-image control meaningful: a classifier that has not seen the image cannot beat chance. For the splice, images are letterboxed onto a common square canvas, because caches of different length cannot be exchanged.

4 The State Records Gist, Not Particulars

We first probe the state at distance zero, the instant the image has just been read. The state is highly informative about which object is present and close to uninformative about how many or where. A

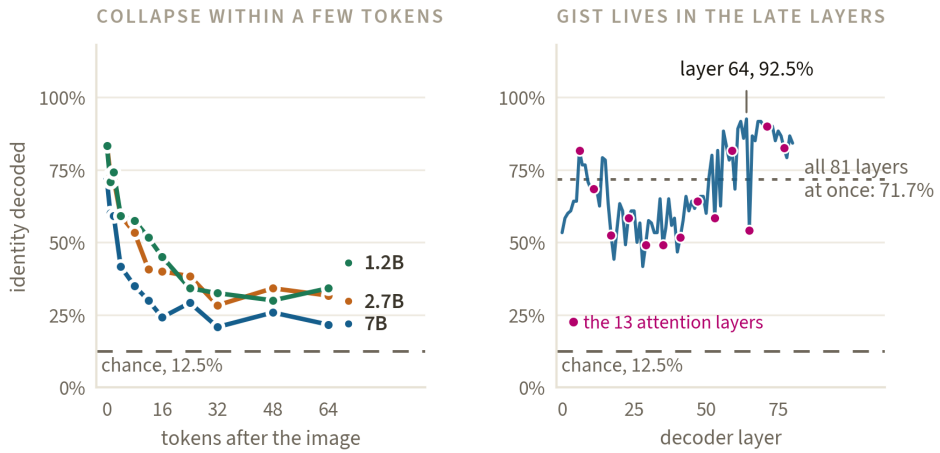


Figure 1: Left: identity decoded from the recurrent state against distance, at three scales, 120 trials per point; the pre-image control reads exactly 12.5%. **Right:** identity decoded from each layer's state at the instant of writing; red marks the 13 layers that also carry attention.

probe given all 81 layers at once reads identity at 83.1% ($p = 6 \times 10^{-92}$, $n = 160$), count at only 19.4% ($p = 0.009$), and position at only 17.5% ($p = 0.041$). The aggregate probe fits 41,472 features to 120 trials and is pulled down by overfitting, so we use it only to compare the three properties against each other; single layers do better. Layer 64 alone reads identity at 92.5%, and the last 21 layers average 85.2% against 59.0% for the middle third of the stack. The gist is strong, and it is concentrated in the late layers.

5 The Forgetting Horizon, Predicted From the Weights

Retention of whatever the state held, after k further tokens, is

$$R(k) = \exp\Big(A \sum_{t=1}^{k} \Delta_t\Big), \qquad \text{half-life} = \frac{\ln 2}{|A|\,\bar{\Delta}}. \tag{2}$$

A is a stored weight. $\bar{\Delta}$ is measured from the gate channels of the input projection over the tokens that follow the image, on the same prompts the decay experiment uses. Nothing here is fitted to a decay curve.

Across all $81 \times 112 = 9{,}072$ heads the predicted median half-life is 6.6 tokens, with an interquartile range of 2.8 to 17.8. 85.0% of heads fall below 32 tokens and only 0.08% exceed 512. Measured over eight different photographs the median moves only between 6.44 and 6.86 tokens, so the horizon is a property of the model rather than of the stimulus.

Does the measured decay match this prediction? On the 7B it appears to: a fine sweep of eleven distances (Table 6, appendix) tracks it closely, the above-chance signal halves at 4.0 tokens, inside the predicted interquartile range, and normalised decodability correlates with predicted mean retention at $r = 0.973$ with no free parameters. Throughout this collapse the model's own accuracy never moves: 94 to 98% at every distance, including where the probe reads near chance.

5.1 The within-model test, with write amplitude held fixed

A curve comparison alone cannot rule out coincidence. So we ran a sharper test. Every head has its own predicted half-life, so we can sort them: a *fast* head, with a half-life around two tokens, should empty almost immediately; a *slow* head should hold information longest. We probe the fast group and the slow group separately, using the same projection basis so the strata are comparable. If the state is genuinely retaining, decay makes a clear prediction: whatever signal survives at distance should sit in the slow heads.

Before giving the result, we must explain a confound, because it changed our own conclusions. The gate $\bar{\Delta}$ appears in both the decay and the write term, so it does double duty: it sets how fast a head

stratum	predicted $t_{1/2}$	d=0	d=16	d=32	d=64	signal kept at d=64
fast	2.3	81.8%	50.0%	35.9%	26.6%	0.20
slow	16.4	91.7%	61.5%	46.9%	34.4%	0.28

Table 1: Identity decodability by gain-matched stratum, $n = 192$ per cell, chance 12.5%. "Signal kept" is the above-chance excess at d=64 as a fraction of the excess at d=0. The slow stratum keeps more at every distance, which is the decay ordering; both strata keep far more than their own decay parameters permit.

stratum	distance	visual KV intact	visual KV evicted	change
fast	0	79.7%	42.2%	−37.5
fast	32	51.6%	**7.8**%	−43.7
fast	64	21.9%	4.7%	−17.2
middle	0	62.5%	43.8%	−18.8
slow	0	62.5%	54.7%	−7.8
all	0	59.4%	59.4%	0.0
all	32	6.2%	6.2%	0.0

Table 2: Identity decodability from the recurrent state with and without the attention channel's copy of the image, $n = 64$ per cell, chance 12.5%. The fast heads' apparent memory at 32 tokens is entirely dependent on the attention channel: removing it drops them below chance.

forgets and how strongly it writes, by about a factor of seven between the fastest and slowest thirds, and a head that writes with larger numbers leaves a signal that is easier to read whatever it contains. Split into naive terciles, the fast third decodes identity at 61.7% after sixty-four tokens while the slow third falls to 16.7%, the opposite of the decay ordering. That reversal has nothing to do with memory. It is the write strength.

The fix is to compare only heads that write equally hard. We keep the $3{,}628$ heads whose mean gate falls in the middle forty percent of its distribution, and split those by $|A|$ alone. This gives strata of 907 heads whose predicted median half-lives still differ sevenfold (2.3 against 16.4 tokens) at nearly equal write gain. Under that control the reversal vanishes (Table 1): the slow stratum decodes better at all seven distances measured, which is the decay ordering. But now look at the amounts. A stratum with a predicted half-life of 2.3 tokens retains $2^{-64/2.3} \approx 4 \times 10^{-9}$ of anything it wrote sixty-four tokens ago. Yet it reads 26.6% there, twice the chance rate ($p = 1 \times 10^{-7}$, $n = 192$), and the measured half-life of its readout is 19.3 tokens, eight times its prediction. In short, the strata are ordered exactly as decay predicts, and both hold far more signal than decay permits.

One caveat: the strata draw unequal numbers of heads per layer, and depth matters for decoding, so the comparison *between* strata carries a depth component. The comparison of each stratum with its own prediction does not, and each reads far above its own bound.

With write amplitude held fixed, the strata are ordered exactly as decay predicts, and both hold far more signal than decay permits. A head with a two-token half-life cannot be holding something written sixty-four tokens earlier: by then its state consists almost entirely of recent writes. If such heads still decode object identity at twice chance, the identity information in their state arrived recently. Something else in the model must be supplying it.

6 The Recurrent Channel Is a Relay, Not a Store

What could be supplying it? Only one component still holds the image at distance: the attention channel, whose key-value entries do not decay. Each trial was run twice, identically, except that in one run the image's key-value entries were evicted from all 13 attention layers immediately after the image was read, before a single filler token was processed; the recurrent state is untouched and identical in both runs at that instant. The two hypotheses now make opposite predictions: if the state remembers on its own, eviction changes nothing; if it is refreshed from attention, eviction collapses decodability at distance, worst for the fast heads, which depend most on being refreshed.

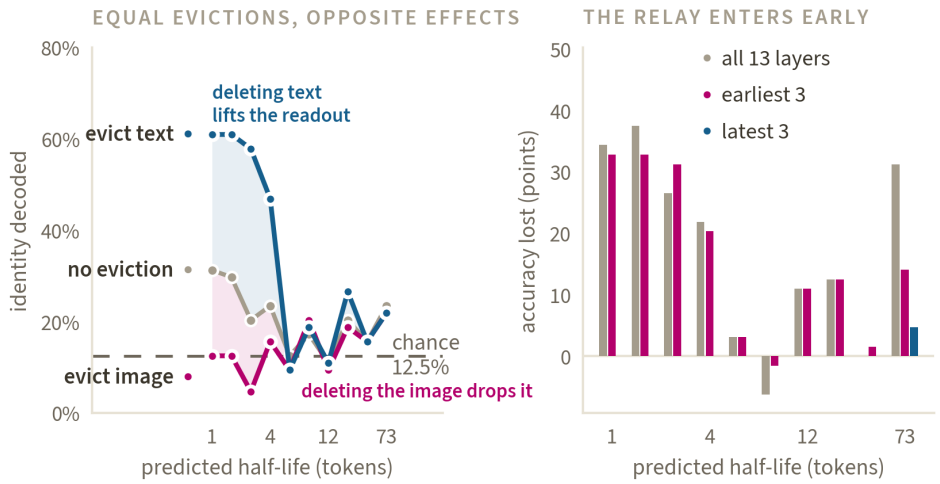


Figure 2: Left: against the no-eviction baseline, text eviction sits above and image eviction below at every distance; both delete the same number of positions, so damage alone cannot explain opposite signs. **Right:** accuracy lost when visual entries are zeroed at all 13 attention layers, the earliest three, or the latest three: the earliest three carry the effect.

The result matches the relay prediction, and the effect is large. At 32 tokens the fast tercile reads 51.6% intact and 7.8% evicted, below the 12.5% chance rate; in raw counts, thirty-three correct of sixty-four become five. The slow heads at distance zero, whose content was written while the image was still present, barely move (62.5% to 54.7%), showing the intervention does not simply break the model.

6.1 Controls: the eviction removes the image, not merely positions

A fair objection: the deletion shortens every cache by roughly 400 positions and changes the softmax normalisation for every later token, so perhaps the readout collapsed because we damaged the machinery. The control is to delete the same *number* of positions from the same caches, but text instead of the image; identical mechanical damage should push both readings the same way. We ran this on ten strata formed by deciles of predicted half-life, 64 trials per cell (Table 8, appendix).

That is not what happens: the two evictions move the readout in opposite directions. At 32 tokens, evicting the image costs the fastest decile 18.8 points, while evicting an equal number of the most recent pre-image text positions *raises* it by 29.7 points (-6.4 against $+12.5$ averaged over the ten strata). Deleting text plausibly helps because the text was competing with the image for the residual stream; the key point is the signs, because identical mechanical damage cannot push a reading up. The dissociation holds on the gain-matched strata at three times the trial count ($n = 192$ per cell): image eviction costs the fast stratum 22.4 points where matched text eviction *raises* it 13.5.

The same collection locates where the relayed information enters. Zeroing the visual entries at only the three earliest attention layers reproduces 15.8 of the 17.2 points lost when all thirteen are zeroed; the three latest cost 0.5. Visual information enters early and is consumed downstream. (Per-layer conditions zero entries rather than delete them; where both operations exist, at d=32, they agree within 2.34 points.)

Finally, the decay parameters predict the size of the dependence: a head that forgets quickly must depend more heavily on being refreshed, so evicting the refresh's source should hurt it more. At matched write gain the fast stratum loses 22.4 points beside the image against 9.4 for the slow, a $2.4\times$ dose-response at equal amplitude ($n = 192$ per cell); on the uncontrolled decile split the drop correlates with the decay rate $1/t_{1/2}$ at $r = 0.800$ (95% CI $[0.34, 0.95]$, $n = 10$) at 32 tokens. The decay calculation carries predictive content about the mechanism rather than the curve: it identifies which heads are relaying and which are retaining.

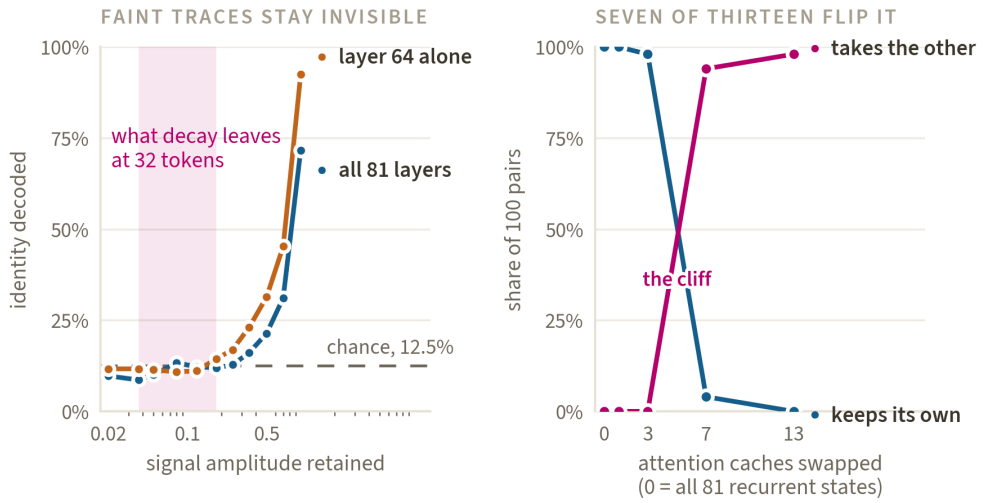


Figure 3: Left: the probe's detection floor. The shaded band is what decay leaves of a stored trace after 32 tokens; both probes are at chance throughout it, so a visible reading there cannot be retained content. **Right:** how many attention caches must be exchanged before the answer follows the other image; the leftmost point is the full recurrent swap, with no effect.

What a probe reads out of the recurrent state at distance is not what the state remembered but what the attention channel is still supplying. Our evidence does not distinguish re-encoding at every token from a single re-supply shortly after the image, so we claim the dependence, not its period. The recurrent channel relays the picture; it does not store it.

6.2 What the probe can see, and why that matters here

A second objection is that the probe may simply miss faint signals: freshly re-encoded content sits at higher amplitude than a decayed trace, so a probe insensitive to faint content could produce all of these observations even if the state never stopped storing the image. This can be measured, and we measured it. We take the distance-zero features and weaken each trial's own contribution while substituting unrelated variation from another trial at constant total variance, $x_i(\alpha) = \mu + \alpha(x_i - \mu) + \sqrt{1-\alpha^2}\,(x_j - \mu)$, which mimics decay followed by overwriting. The probe loses half its signal at $\alpha = 0.78$ and is at chance by $\alpha \approx 0.25$: it cannot see content below roughly a quarter of full amplitude. The weights say a retained trace at thirty-two tokens sits at 3.7% of full amplitude at the median head and 18% at the mean, both below that floor, so a genuinely retained trace *could not* produce a visible reading. Yet the intact probe reads the fast tercile at 51.6% there. Whatever it reads must be high in amplitude, which means recently written; and removing the attention channel's copy removes it.

The probe's insensitivity to faint signals is what makes the positive reading informative: a signal it can see at thirty-two tokens is necessarily fresh. The same measurement bounds the claim. Whether faint retained content persists below the floor is not answerable with this instrument, so we do not claim the state stops storing the image; we claim that what a probe reads at distance is not retention, and that it depends causally on the attention channel.

7 Causally, the Answer Follows the Attention Channel

The splice tests the same claim on the model's own behaviour, with no probe involved. Two runs see different images, one memory channel is exchanged, and the host run answers its question. A pair is admitted only if host and donor each answered correctly beforehand, so the ceiling is 100% by construction; of 111 attempted pairs, 11 were rejected on that criterion, and every recurrent swap was verified to have changed the state, 100 of 100.

Three things follow from the table. Replacing the recurrent memory has no effect at all: exchanging all 81 states leaves 100 of 100 answers on the host's own image. The attention channel is causally

condition	layers swapped	follows host	follows donor
no swap	0	$100/100$ $[96, 100]$	$0/100$ $[0, 4]$
recurrent, all 81 states	0	$100/100$ $[96, 100]$	$0/100$ $[0, 4]$
attention, 1 layer	1	$100/100$ $[96, 100]$	$0/100$ $[0, 4]$
attention, 3 layers spread	3	$98/100$ $[93, 99]$	$0/100$ $[0, 4]$
attention, 7 layers	7	$4/100$ $[2, 10]$	$94/100$ $[88, 97]$
attention, all 13	13	$0/100$ $[0, 4]$	$98/100$ $[93, 99]$
attention, earliest 3	3	$94/100$ $[88, 97]$	$1/100$ $[0, 5]$
attention, latest 3	3	$100/100$ $[96, 100]$	$0/100$ $[0, 4]$
both channels	13	$0/100$ $[0, 4]$	$100/100$ $[96, 100]$

Table 3: The splice at 100 pairs with a 100% ceiling. Brackets are 95% Wilson intervals in percent. Replacing the entire recurrent memory of all 81 layers changes nothing; replacing seven of the thirteen attention caches flips the answer to the other image.

model	layers	attn	probe d=0	probe d=64	measured $t_{1/2}$	predicted $t_{1/2}$
Zamba2-VL-7B	81	13	71.7%	21.7%	4.0	6.64
Zamba2-VL-2.7B	54	9	84.2%	31.7%	9.6	2.91
Zamba2-VL-1.2B	38	6	83.3%	34.2%	14.3	2.58

Table 4: The same protocol at three scales, 120 identity trials per distance. The pre-image control reads exactly 12.5% at every scale. Half-lives are in tokens: measured is the distance at which above-chance decodability halves, predicted is the median over heads of $\ln 2/(|A|\bar{\Delta})$.

sufficient: seven of thirteen caches move 94 of 100 answers to the donor's. And the dependence is graded and distributed: one layer changes nothing, seven move almost everything, and the earliest three matter where the latest three do not, matching the layerwise eviction of Section 6.1.

Why does replacing the entire recurrent memory change nothing? The transplanted state is decayed away within a few tokens and rewritten from the attention channel, which the swap leaves untouched. The probe result and the splice result are the same fact seen from two sides.

8 Scale, and Where the Prediction Fails

We repeated the core protocol at three model scales. The instruments read every tensor shape from the loaded checkpoint's own configuration, which is a requirement rather than a convenience: the smaller checkpoints differ from the 7B in group count and state dimension.

Two conclusions follow, the second a negative result we report plainly. The parameter-derived horizon is short at every scale: median predicted half-lives of 6.64, 2.91 and 2.58 tokens, with 85.0 to 97.9% of heads under 32 tokens. But the predicted horizon does not predict the measured one: across scales the two move in opposite directions, so the $r = 0.973$ agreement on the 7B does not replicate and we do not claim the decay calculation explains the measured curve. Section 6 shows why it cannot: the probe at distance was never measuring retention in the first place.

9 Visual Key-Value Eviction Removes the Only Copy, Eventually

Pruning methods discard visual key-value entries on the premise that they are redundant. Our results say the redundancy comes from the relay, and the relay is short-lived, so the premise should degrade with distance. We test this directly on the model's own behaviour. We evict a fraction of the visual key-value entries from all 13 attention layers immediately after the image is read, let the filler and the question stream past the pruned cache, and measure the model's accuracy. Retention ρ is the fraction of visual positions kept, sampled uniformly, and the question is asked d tokens after the image.

With a quarter of the entries retained there is no distance effect at all: the cost against the unpruned baseline is 6.2, 0.0, 4.2 and 6.2 points, flat within noise, so at that level the redundancy premise holds. Below it, distance matters a great deal. At a twentieth retained the cost triples across the range, from 10.4 points beside the image to 33.3 at 512 tokens; with every visual entry removed it runs from 33.3

visual KV retained	d=0	d=32	d=128	d=512
100%	97.9%	95.8%	100.0%	97.9%
25%	91.7%	95.8%	95.8%	91.7%
5%	87.5%	77.1%	75.0%	64.6%
0%	64.6%	37.5%	31.2%	22.9%

Table 5: Behavioural accuracy under visual key-value eviction, against distance between the image and the question. $n = 48$ per cell, chance 12.5%, no failures. The unpruned row is flat, so distance alone costs nothing.

to 75.0 (Table 11, appendix). The unpruned row is flat, so long prompts are not hard in themselves; they become hard once the visual entries are gone, because the recurrent channel then has nothing left to relay from. A method validated at short range therefore understates its long-range cost threefold at 5% retention.

The zero-retention row measures what the recurrent channel contributes on its own: accuracy falls from 64.6% beside the image to 22.9% at 512 tokens, a behavioural measurement that agrees with the probe. One discrepancy is worth stating plainly. At zero retention the model still answers at 37.5% at thirty-two tokens while the probe reads the state near chance, so the model uses something the probe does not recover. The probe is a lower bound, and our evidence does not support the sentence "the state is empty".

10 Limitations

The probe is a lower bound, and our own behavioural data prove it. With every visual entry removed, the model answers at 37.5% at thirty-two tokens while the probe reads near chance, so every statement about what the state contains is a statement about what is linearly decodable under this instrument, and accuracies are comparable only within a collection.

A sensitivity asymmetry is bounded but not eliminated. The floor argument of Section 6.2 assumes decay plus overwriting acts on the discriminative component roughly as our attenuation does. A model with no attention channel would settle this without the assumption; we have not tested one. The route is also located but not traced past the early attention layers.

The dose-response is confirmed at matched amplitude only at distance zero. At thirty-two tokens the gain-matched fast stratum reads too close to chance even when intact for the contrast to be informative. Two strata at one distance in one model is a direction, not a curve.

One model family, and a second family we tried and failed on. All results come from three Zamba2-VL checkpoints, the principal limit on generality. On Granite 4.0 h-tiny (IBM Granite Team, 2025) the horizon calculation transfers cleanly (median 8.7 tokens) but the relay contrast does not: the probe reads that model's state at chance even at distance zero on a task it answers at 100% (Appendix B), so no relay claim extends beyond Zamba2-VL on this evidence. The sharper test remains a vision-language model with no attention at all, such as Cobra (Zhao et al., 2025) or VL-Mamba (Qiao et al., 2024), both Mamba1 models needing the read operator ported first.

One task format, one dataset, one filler distribution. Eight-way forced choice on COCO buys an exact chance rate and a clean pre-image control; it costs generality to free-form answering. $\bar{\Delta}$ is input dependent: the median half-life is stable across images (6.44 to 6.86 tokens), but a filler with very different gate statistics would give a different horizon, which the same calculation predicts.

11 Conclusion

The recurrent state of this hybrid vision-language model is not a memory of the image in any useful sense. It is a selective summary of the gist, held for about four tokens, a lifetime computable from the decay parameters before any measurement is taken. What the model answers from lives in the thin attention channel, and the practical consequence is direct: a hybrid's visual key-value entries are redundant only while the recurrent state is still relaying the picture, and that is not long.

Reproducibility Statement

Every number in this paper comes from a scripted experiment against public checkpoints on public data. The model is Zamba2-VL-7B (Zyphra, 2026) together with its 2.7B and 1.2B siblings, all Apache-2.0; trials are built from COCO val2017 annotations (Lin et al., 2014). Section 3 gives the recurrence, the projection width, the probe and its cross-validation scheme, and the exact statistical tests; the decay prediction is stated in closed form in Section 5 with $\bar{\Delta}$ measured as described. The instruments read layer counts, head counts, group counts and state widths from each loaded checkpoint's own configuration rather than assuming the 7B geometry, which is what let the same code run at three scales and on a second family. Appendix A reports the full tables behind the summarised results, including the fine distance sweep, the per-decile control values and the seed repeats. Source for the instruments, the analysis and the figures accompanies the submission.

Ethics Statement

This work analyses publicly released model checkpoints on a public image dataset, involves no human subjects and collects no new data. COCO images depict people, and our trials ask only about object category, count and position; no attribute inference about depicted persons is performed. The interpretability findings are dual-use in the ordinary sense that understanding where a model holds information could inform both better and worse systems, but they identify no vulnerability and enable no capability that inspection of the open weights would not. We report a negative result rather than omitting it.

AI Use Statement

Large language models were used as coding and writing assistants: drafting and refactoring experiment and analysis code, and copy-editing prose. All experimental design, all interpretation of results, and all claims in this paper are the authors'. Every number reported was produced by the scripted pipeline and checked against its source output; no text describing results was accepted without that check.

References

G. Alain and Y. Bengio. Understanding Intermediate Layers Using Linear Classifier Probes. *arXiv:1610.01644*, 2016.

S. Arora, S. Eyuboglu, A. Timalsina, I. Johnson, M. Poli, J. Zou, A. Rudra, and C. Ré. Zoology: Measuring and Improving Recall in Efficient Language Models. *arXiv:2312.04927*, 2023.

S. Arora, S. Eyuboglu, M. Zhang, A. Timalsina, S. Alberti, D. Zinsley, J. Zou, A. Rudra, and C. Ré. Simple Linear Attention Language Models Balance the Recall-Throughput Tradeoff. *arXiv:2402.18668*, 2024.

S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, et al. Qwen2.5-VL Technical Report. *arXiv:2502.13923*, 2025.

M. Beck, K. Pöppel, M. Spanring, A. Auer, O. Prudnikova, M. Kopp, G. Klambauer, J. Brandstetter, and S. Hochreiter. xLSTM: Extended Long Short-Term Memory. *arXiv:2405.04517*, 2024.

Y. Belinkov. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 2022.

L. Chen, H. Zhao, T. Liu, S. Bai, J. Lin, C. Zhou, and B. Chang. An Image is Worth 1/2 Tokens After Layer 2: Plug-and-Play Inference Acceleration for Large Vision-Language Models. *ECCV*, 2024.

T. Dao and A. Gu. Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality. *ICML*, 2024.

S. De, S. L. Smith, A. Fernando, A. Botev, G. Cristian-Muraru, A. Gu, et al. Griffin: Mixing Gated Linear Recurrences with Local Attention for Efficient Language Models. *arXiv:2402.19427*, 2024.

P. Glorioso, Q. Anthony, Y. Tokpanov, J. Whittington, J. Pilault, A. Ibrahim, and B. Millidge. Zamba: A Compact 7B SSM Hybrid Model. *arXiv:2405.16712*, 2024.

A. Gu and T. Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv:2312.00752*, 2023.

A. Gu, K. Goel, and C. Ré. Efficiently Modeling Long Sequences with Structured State Spaces. *ICLR*, 2022.

J. Hewitt and P. Liang. Designing and Interpreting Probes with Control Tasks. *EMNLP*, 2019.

IBM Granite Team. Granite 4.0 h-tiny model card. `https://huggingface.co/ibm-granite/granite-4.0-h-tiny`, 2025.

S. Jelassi, D. Brandfonbrener, S. M. Kakade, and E. Malach. Repeat After Me: Transformers are Better than State Space Models at Copying. *arXiv:2402.01032*, 2024.

A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret. Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. *ICML*, 2020.

Y. Li, Y. Huang, B. Yang, B. Venkitesh, A. Locatelli, H. Ye, T. Cai, P. Lewis, and D. Chen. SnapKV: LLM Knows What You are Looking for Before Generation. *arXiv:2404.14469*, 2024.

O. Lieber, B. Lenz, H. Bata, G. Cohen, J. Osin, I. Dalmedigos, et al. Jamba: A Hybrid Transformer-Mamba Language Model. *arXiv:2403.19887*, 2024.

T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. Microsoft COCO: Common Objects in Context. *ECCV*, 2014. arXiv:1405.0312.

H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual Instruction Tuning. *NeurIPS*, 2023.

K. Meng, D. Bau, A. Andonian, and Y. Belinkov. Locating and Editing Factual Associations in GPT. *NeurIPS*, 2022.

W. Merrill, J. Petty, and A. Sabharwal. The Illusion of State in State-Space Models. *ICML*, 2024.

B. Peng, E. Alcaide, Q. Anthony, A. Albalak, S. Arcadinho, S. Biderman, et al. RWKV: Reinventing RNNs for the Transformer Era. *arXiv:2305.13048*, 2023.

Y. Qiao, Z. Yu, L. Guo, S. Chen, Z. Zhao, M. Sun, Q. Wu, and J. Liu. VL-Mamba: Exploring State Space Models for Multimodal Learning. *arXiv:2403.13600*, 2024.

L. Ren, Y. Liu, Y. Lu, Y. Shen, C. Liang, and W. Chen. Samba: Simple Hybrid State Space Models for Efficient Unlimited Context Language Modeling. *ICLR*, 2025.

Y. Shang, M. Cai, B. Xu, Y. J. Lee, and Y. Yan. LLaVA-PruMerge: Adaptive Token Reduction for Efficient Large Multimodal Models. *ICCV*, 2025.

J. Vig, S. Gehrmann, Y. Belinkov, S. Qian, D. Nevo, S. Sakenis, J. Huang, Y. Singer, and S. Shieber. Causal Mediation Analysis for Interpreting Neural NLP: The Case of Gender Bias. *arXiv:2004.12265*, 2020.

R. Waleffe, W. Byeon, D. Riach, B. Norick, V. Korthikanti, T. Dao, A. Gu, et al. An Empirical Study of Mamba-based Language Models. *arXiv:2406.07887*, 2024.

K. Wang, A. Variengien, A. Conmy, B. Shlegeris, and J. Steinhardt. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small. *arXiv:2211.00593*, 2022.

K. Wen, X. Dang, and K. Lyu. RNNs are not Transformers (Yet): The Key Bottleneck on In-context Retrieval. *arXiv:2402.18510*, 2024.

G. Xiao, Y. Tian, B. Chen, S. Han, and M. Lewis. Efficient Streaming Language Models with Attention Sinks. *ICLR*, 2024.

S. Yang, Y. Chen, Z. Tian, C. Wang, J. Li, B. Yu, and J. Jia. VisionZip: Longer is Better but Not Necessary in Vision Language Models. *arXiv:2412.04467*, 2024.

F. Zhang and N. Nanda. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. *ICLR*, 2024.

Z. Zhang, Y. Sheng, T. Zhou, T. Chen, L. Zheng, R. Cai, Z. Song, Y. Tian, C. Ré, C. Barrett, Z. Wang, and B. Chen. H_2O: Heavy-Hitter Oracle for Efficient Generative Inference of Large Language Models. *arXiv:2306.14048*, 2023.

H. Zhao, M. Zhang, W. Zhao, P. Ding, S. Huang, and D. Wang. Cobra: Extending Mamba to Multi-Modal Large Language Model for Efficient Inference. *AAAI*, 2025.

Zyphra. Zamba2-VL-7B model card. `https://huggingface.co/Zyphra/Zamba2-VL-7B`, 2026. Apache-2.0.

A Full Tables

This appendix carries the tables behind the summarised results in the body: the fine distance sweep, the matched-eviction control in full by decile, the seed repeat of the eviction contrast, and the pruning costs relative to the unpruned baseline at each distance. Cell counts and chance rates are stated in each caption.

B The Second-Family Attempt in Full

Granite 4.0 h-tiny (IBM Granite Team, 2025) is a Mamba2 hybrid from a different lineage: 40 decoder layers, of which 36 carry a Mamba2 mixer and 4 carry attention (at indices 5, 15, 25, 35), 48 heads, state dimension 128, one B/C group. Because its attention layers carry no mixer at all, the recurrent channel spans only 36 of the 40 layers, which is why the instruments ask each model which layers have mixers rather than assuming every layer does.

tokens	probe accuracy	p	behaviour	predicted retention (median)
0	71.7%	2×10^{-50}	94.2%	1.000
1	60.0%	2×10^{-34}	96.7%	0.901
2	59.2%	2×10^{-33}	97.5%	0.812
4	41.7%	1×10^{-15}	96.7%	0.659
8	35.0%	2×10^{-10}	97.5%	0.434
12	30.0%	3×10^{-7}	96.7%	0.286
16	24.2%	3×10^{-4}	97.5%	0.188
24	29.2%	1×10^{-6}	95.8%	0.082
32	20.8%	7×10^{-3}	95.0%	0.036
48	25.8%	6×10^{-5}	97.5%	0.007
64	21.7%	3×10^{-3}	98.3%	0.001

Table 6: The fine distance sweep of Section 5: decodability of object identity against distance from the image, 120 trials at every distance, with the retention predicted from the weights alongside. Chance is 12.5%; the pre-image control reads exactly 12.5%. Beyond 16 tokens the measured curve sits between the predicted median and mean, which is what the heavy tail of slow heads predicts.

Established by prior work	**New here**
A fixed-size state must lose information from context, with a capacity proof and behavioural evidence (Jelassi et al., 2024; Merrill et al., 2024; Wen et al., 2024).	*When*. A retention horizon in tokens, computed from the model's own A and Δ with nothing fitted, at three scales.
Recurrent models underperform at copying and retrieval as a class (Arora et al., 2023; 2024).	*What survives*. The failure is content-selective: object identity is highly decodable at the same instant that count and position are near chance.
Architectures are compared to one another from the outside.	*Causal localisation inside one model*, and the finding that the recurrent channel relays rather than stores, which makes probing it at distance a measurement of the attention channel.
Visual tokens and key-value entries carry substantial redundancy that can be pruned (Chen et al., 2024; Yang et al., 2024; Shang et al., 2025).	*What underwrites that premise*, and that its safety margin shrinks with distance from the image on a hybrid.

Table 7: Delineation of prior work and contribution.

The horizon calculation transfers without modification. Reading A and the gate channels of its input projection gives a median predicted half-life of 8.7 tokens, the same single-digit regime as the Zamba2-VL family, so Section 5 is not specific to one vendor.

The relay contrast did not transfer, for an instrumental reason we were unable to resolve. Granite is a text model, so the carried item is a code word rather than an image. On a code-word recall task the model answers correctly on 100%, 96.9% and 100% of trials at 0, 32 and 64 tokens, so the information is demonstrably in the model. Yet a probe on its recurrent state reads at chance at distance zero, where the Zamba2-VL probe reads 80 to 90%. Scaling to 256 trials did not change this (13.7% against a 12.5% chance rate, $p = 0.31$), and neither did probing single layers, where the best of 36 reached 19.9%.

We suspected the item was too small to write a detectable trace: a one-token code word against the 345-token image of the main experiments. We tested that by repeating the carrier sixteen times so the item occupied 192 tokens. It did not help (12.0%, 17.2% and 10.9% across the three strata). Repetition lengthens an item without diversifying what it writes, so the state ends up holding little more than the final repeat, which is not the manipulation we intended.

We do not know why the instrument fails on this model. Candidate explanations we cannot currently separate include the narrower recurrent channel, the different item modality, and a genuine architectural difference in what the state holds. A vision-language hybrid from a second family, where the item is a real image and the probe is the same one that works here, remains the right test and we have not run it.

	fastest decile ($t_{1/2} = 0.7$)		mean over ten deciles		
condition	d=0	d=32	d=0	d=32	effect
no eviction	56.2%	31.2%	44.2%	20.5%	baseline
evict image	18.8%	12.5%	25.5%	14.1%	−18.7, −6.4
evict matched text	78.1%	60.9%	58.9%	33.0%	+14.7, +12.5

Table 8: The matched control, $n = 64$ per cell, chance 12.5%. Both interventions delete the same number of positions from the same 13 caches, one taking the image and one an equal count of the most recent pre-image text. They move the readout in opposite directions, which a perturbation account cannot produce. The last column gives the mean effect in points at d=0 and d=32.

		$d = 0$			$d = 32$	
decile	pred. $t_{1/2}$	intact	evict image	evict text	intact	evict image
1	0.70	56.2%	18.8%	78.1%	31.2%	12.5%
2	1.84	56.2%	17.2%	79.7%	29.7%	12.5%
3	2.75	53.1%	18.8%	73.4%	20.3%	4.7%
4	3.79	46.9%	21.9%	71.9%	23.4%	15.6%
5	5.38	31.2%	26.6%	42.2%	12.5%	9.4%
6	8.15	35.9%	40.6%	51.6%	17.2%	20.3%
7	11.8	37.5%	25.0%	54.7%	10.9%	9.4%
8	17.8	45.3%	29.7%	57.8%	20.3%	18.8%
9	32.0	28.1%	26.6%	32.8%	15.6%	15.6%
10	73.3	51.6%	29.7%	46.9%	23.4%	21.9%

Table 9: The matched-eviction control of Section 6.1 in full, by decile of predicted half-life, $n = 64$ per cell, chance 12.5%. "Evict image" and "evict text" delete the same number of positions from the same 13 attention caches. Single cells carry about 6 points of binomial noise at $n = 64$, and two of the twenty image-eviction cells fall slightly the wrong way; the trend rather than any one cell is the result. The drop under image eviction correlates with the decay rate $1/t_{1/2}$ at $r = 0.647$ at d=0 and $r = 0.800$ at d=32.

stratum	condition	seed 0	seed 1	seed 2
fast	d=0, intact	62.5%	64.1%	64.1%
fast	d=0, evicted	18.8%	20.3%	18.8%
fast	d=32, intact	32.8%	29.7%	31.2%
fast	d=32, evicted	7.8%	10.9%	10.9%
middle	d=32, intact	14.1%	12.5%	10.9%
middle	d=32, evicted	14.1%	12.5%	9.4%
slow	d=32, intact	12.5%	15.6%	17.2%
slow	d=32, evicted	14.1%	15.6%	17.2%

Table 10: The eviction contrast repeated across three independent Gaussian projection bases, $n = 64$ per cell, chance 12.5%. The fast tercile's dependence on the attention channel is stable across bases (intact 31.2 ± 1.6% against evicted 9.9 ± 1.8% at d=32, mean and standard deviation over seeds), while the slow tercile shows no eviction effect at any seed. The decay-curve collections elsewhere in the paper use a single fixed basis.

visual KV retained	d=0	d=32	d=128	d=512
25%	6.2	0.0	4.2	6.2
5%	10.4	18.8	25.0	33.3
0%	33.3	58.3	68.8	75.0

Table 11: Accuracy points lost to pruning, relative to the unpruned baseline at the same distance (Section 9).

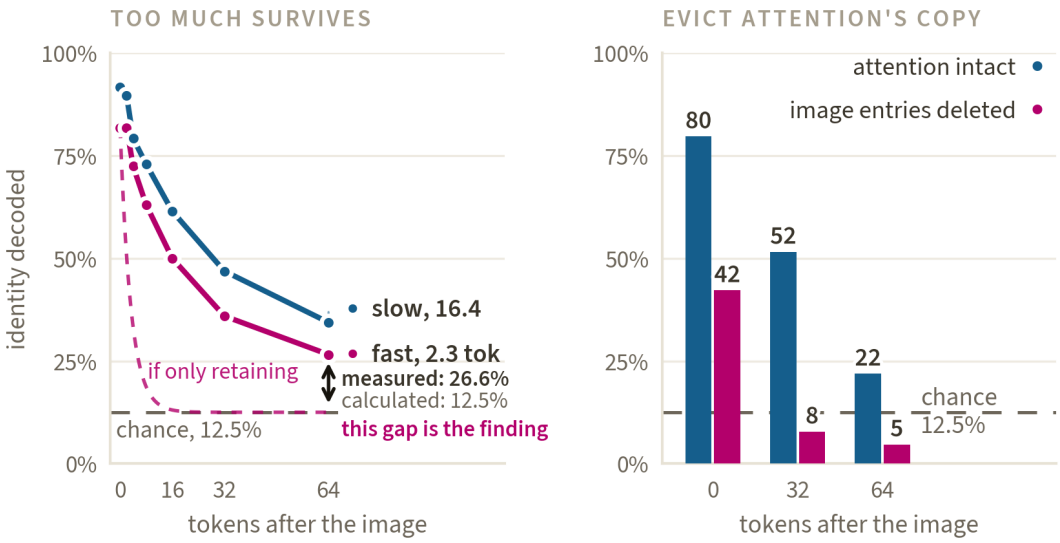


Figure 4: Left: how to read this panel. The two solid curves are measurements: identity decoded from the slow and fast strata at matched write gain, $n = 192$ per point. The dashed curve is not a measurement but a calculation: where the fast stratum would read if its signal came only from what its own 2.3-token half-life retains. The calculation reaches chance within ten tokens; the measurement ends at 26.6% at d=64. The slow-above-fast ordering is what decay predicts; the gap between the two fast curves is what decay cannot explain. **Right:** the same recurrent state probed with and without the attention channel's copy of the image, fast tercile, $n = 64$ per bar. The eviction never touches the recurrent state, which is identical in both runs at that instant.

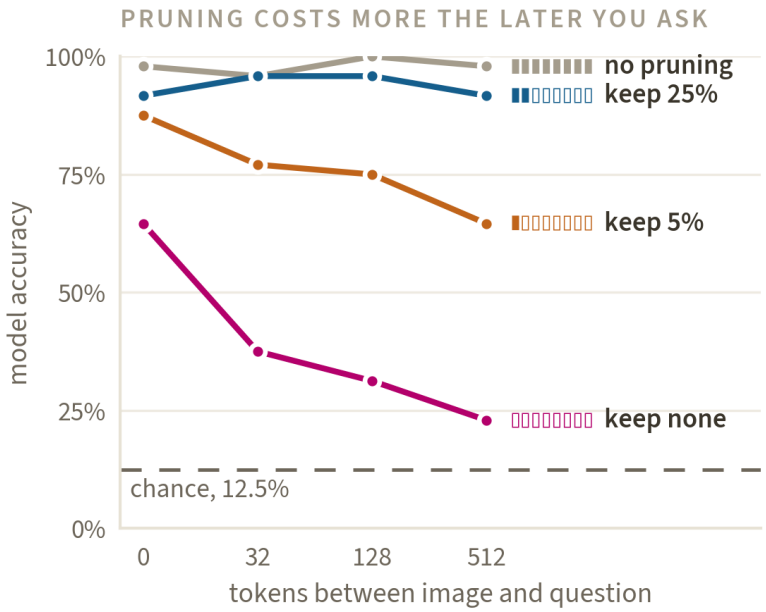


Figure 5: Model accuracy under visual key-value eviction, against how far the question sits from the image, $n = 48$ per point. The unpruned line is flat, so distance alone costs nothing. Keeping a quarter of the entries is also flat. Below that the lines fan out: the cost of pruning grows with distance only once the recurrent channel has stopped relaying the picture.