DEPTH-ORDERED FEATURE MEMORY: FROM SNAPSHOTS TO TRAJECTORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse dictionaries describe transformers layer by layer and crosscoders assign cross-layer identities, yet neither maintains state across depth. We introduce **Depth-Ordered Feature Memory (DOFM)**, a sparse recurrent observer trained to predict normalized residual updates across a frozen transformer. DOFM integrates three mechanisms: a shared decoder fixing directional coordinates across depth (***identity***); a per-token recurrent state tracking writes, retention, and attenuation (***history***); and cumulative master gates organizing coordinates into a retention hierarchy (***address***). Across diverse backbones, accumulated state reads word sense as accurately as the residual stream itself while revealing what probes cannot: which coordinate carries the sense, when it was written, and how long it persists. The retention prior hierarchically organizes coordinates, allowing a persistent tier to match full-dictionary accuracy and support addressable causal intervention; lexical competition resolves via write-time asymmetric recruitment, where competing senses receive reduced writes and persist without downstream erasure. Feature lifespan indexes persistence and generality over abstraction, turning cross-layer computation into an addressable state trajectory.

1 INTRODUCTION

Transformers build semantic representations across depth. While sparse autoencoders isolate layer-wise directions (Bricken et al., 2023; Huben et al., 2024; Gao et al., 2024) and crosscoders establish cross-layer identities (Lindsey et al., 2024), neither maintains internal state across depth. Retrospective comparisons cannot determine whether recurring activations reflect continuous retention or repeated re-derivation. Understanding intermediate computation requires analyzing the **feature lifecycle**: a consistent directional coordinate paired with a history of state updates across depth.

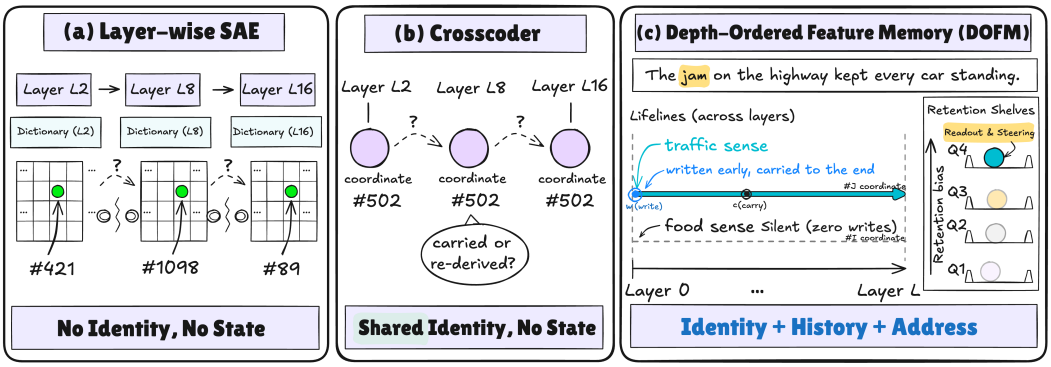


Figure 1: **From disconnected snapshots to an addressable feature lifecycle. (a) Layer-wise SAEs:** Independent dictionaries require post-hoc matching without state (*No Identity, No State*). **(b) Crosscoder:** Unifies coordinates but remains stateless, unable to separate retention from re-derivation (*Shared Identity, No State*). **(c) DOFM:** Unifies identity, history, and address. For *jam* (traffic), target coordinates are written (w) and carried (c) while food coordinates receive zero writes; retention shelves (Q1–Q4) isolate persistent tier Q4 for readout and steering.

Lifecycle analysis requires three connected properties: *identity* establishes consistent directional coordinates across depth, ensuring tracking without retrospective matching; *history* maintains an explicit recurrent state decomposing updates into writes, retention, and attenuation to separate persistence from re-derivation; and *address* organizes coordinates by retention timescale, making features

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.

queryable. To unify these requirements, we introduce **Depth-Ordered Feature Memory (DOFM)**, a sparse recurrent observer of residual updates in a frozen transformer. DOFM pairs a shared decoder across depth with a per-token recurrent state, while cumulative master gates adapted from ordered neurons (Shen et al., 2019) order coordinates by retention.

Figure 1 grounds this lifecycle by contrasting it with per-layer dictionaries (Huben et al., 2024) and cross-layer architectures (Dunefsky et al., 2024; Lindsey et al., 2024) on the token *jam* in a traffic context. While prior frameworks leave unresolved whether recurring activations reflect continuous state or re-derivation, DOFM provides an explicit recurrent history: the traffic-leaning coordinate (#J) is written early and carried to the end, while the competing food coordinate (#I) receives no write at all—disambiguation is decided at write time, and what is written is carried to the output. Moreover, cumulative gates sort these lifecycles onto retention shelves, which isolate coordinates by computational persistence, separate resolved semantic state from transient noise, and query coordinates for downstream readout and causal intervention. By establishing shared cross-depth coordinates (***identity***), recurrent state updates (***history***) (Belrose et al., 2023), and retention-ordered tiers (***address***), we form an operational framework that renders feature trajectories *traceable*, *inspectable*, and *queryable*—**track a coordinate, inspect its history, and query its retention address**.

We evaluate DOFM on Gemma and Qwen backbones (Gemma Team, 2024; Qwen Team, 2025) across word-sense disambiguation (Miller, 1995) and behavioral steering, confirming three core findings. First, shared coordinates exhibit robust cross-context selectivity across depth without heuristic alignment. Second, the observer turns a readout into a history: it reads sense at the accuracy of a probe on the raw residual while resolving that reading into dated, coordinate-level writes, and it shows that sense is written in a few early layers and carried to the output rather than recomputed. Third, retention ordering provides an operative address: the persistent tier matches full-dictionary accuracy and captures three-quarters of steering coordinates—capabilities absent in unconstrained memory or post-hoc lifespan sorting. In summary, our contributions are threefold:

• **A sparse lifecycle representation:** We introduce DOFM, a sparse recurrent observer that formalizes feature lifecycles across transformer depth through consistent directional identity, explicit write-and-attenuation histories, and a retention-ordered address space.

• **Semantic lifecycle accounting:** We derive an exact decomposition of semantic readout margins into the surviving contributions of earlier writes to date and locate meaning inside the network: resolved sense is written in early layers, carried to the output, and read from the persistent band at the accuracy of the full dictionary and raw residual stream.

• **A retention-indexed control interface:** We demonstrate that retention-selected decoder subspaces effectively actuate behavioral steering under ambiguity across diverse backbones, establishing feature lifecycles as an actionable tool for mechanistic intervention.

2 Depth-Ordered Feature Memory

Depth-Ordered Feature Memory (DOFM) is a sparse recurrent observer of feature lifecycles across transformer depth. Figure 2 depicts DOFM alongside a frozen transformer backbone, functioning without weight alteration or gradient backpropagation.

2.1 Problem Setup and Preliminaries

Let $H_\ell \in \mathbb{R}^{d_{\text{model}}}$ be the residual stream entering block ℓ of a frozen transformer, with H_0 the embedding output (nostalgebraist, 2020; Belrose et al., 2023). An observer step $s = 1, \ldots, S$ reads a source depth ℓ_s and predicts the update to a target depth t_s; our full-depth configuration takes $t_s = \ell_s + 1$, with S equal to the block count. To prevent tracking scale drift across depth (Gao et al., 2024; Lieberum et al., 2024), inputs and targets are standardized with per-depth statistics (μ, r) estimated on 100,000 pilot training tokens, following Lindsey et al. (2024): $x_s = \frac{H_{\ell_s} - \mu_{\ell_s}}{r_{\ell_s}}$, $\quad y_s = \frac{(H_{t_s} - H_{\ell_s}) - \mu_{\Delta,s}}{r_{\Delta,s}}$, so that y_s is the normalized residual update computed by the block (Dunefsky et al., 2024). DOFM is trained without labels to predict y_s from $x_1, \ldots, x_s$, minimizing the mean squared error $\frac{1}{SBd_{\text{model}}} \sum_{s,n} \|\hat{y}_s^{(n)} - y_s^{(n)}\|_2^2$ over a batch of B tokens.

2.2 Identity: One Direction Across Depth

Tracking a feature across depth requires a consistent directional coordinate system. Parameterizing this cross-depth mapping involves an efficiency trade-off: unconstrained cross-layer dictionaries

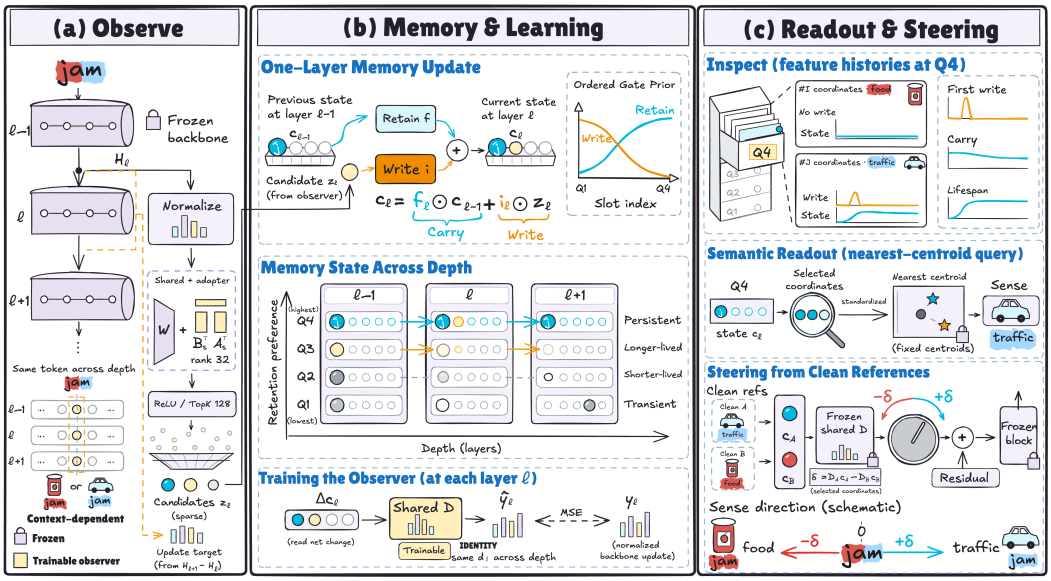


Figure 2: **DOFM architecture, memory dynamics, and operational interfaces. (a) Observe:** Standardized residual inputs x_ℓ from a frozen backbone pass through a shared encoder with source-specific rank-32 adapters ($B_s^\top A_s^\top$) to yield sparse candidates z_ℓ via TopK. Target y_ℓ is the normalized update between adjacent blocks ($H_{\ell+1} - H_\ell$). **(b) Memory & Learning:** State updates combine retained carry with gated writes ($c_\ell = f_\ell \odot c_{\ell-1} + i_\ell \odot z_\ell$), structured by cumulative master gates across slots Q1–Q4. Decoded net transitions $\Delta c_\ell = w_\ell - e_\ell$ optimize an MSE loss against y_ℓ via shared unit-norm decoder D. **(c) Readout & Steering:** Persistent tier Q4 enables three operations: inspecting lifelines, querying semantic states via nearest-centroid classification, and steering ambiguous representations using reference-derived vectors $\delta = D_A c_A - D_B c_B$.

(Lindsey et al., 2024) assign independent projection matrices to each depth, scaling parameters linearly with depth ($L\times$) and requiring billions of parameters for deep networks (e.g., ≈ 7.9B across 26 layers). Conversely, fully shared single dictionaries (Lawson et al., 2025) minimize parameters but lack layer-specific capacity to handle depth-varying activation distributions.

DOFM resolves this tension by decoupling directional identity from depth-specific activation dynamics. A shared base encoder combines with source-specific rank-r adapters ($A_s \in \mathbb{R}^{d_{\text{model}} \times r}, B_s \in \mathbb{R}^{r \times d_f}$, rank $r = 32$) in a LoRA-style residual parameterization (Hu et al., 2022), followed by per-step TopK selection (Gao et al., 2024):

$$a_s = W_e x_s + B_s^\top A_s^\top x_s + b_s, \qquad z_s = \text{TopK}_K\big(\text{ReLU}(a_s)\big) \tag{1}$$

with dictionary width $d_f = 131{,}072$ and candidate budget $K = 128$. Per-step TopK selection allows each depth to recruit distinct coordinates from the shared vocabulary. Compact rank-32 adapters restrict total 26-layer observer parameters to 0.72B, avoiding unconstrained parameter explosion. Directional identity originates from the decoder: a single matrix $D \in \mathbb{R}^{d_{\text{model}} \times d_f}$ with unit-norm columns ($\|D_{:,j}\|_2 = 1$) spans all depths, fixing direction $d_j = D_{:,j}$ by construction to enable lifecycle tracking without retrospective alignment.

2.3 History: Recurrent Updates and Lifecycles

Each token position carries a recurrent state $c_s \in \mathbb{R}^{d_f}$, initialized at $c_0 = 0$ and updated once per depth step with input and forget gates $i_s, f_s \in [0,1]^{d_f}$ (Section 2.5):

$$c_s = \underbrace{f_s \odot c_{s-1}}_{\text{carry}} + \underbrace{i_s \odot z_s}_{w_s\ \text{(write)}} = c_{s-1} + w_s - e_s, \qquad e_s = (1 - f_s) \odot c_{s-1} \qquad \widehat{y}_s = D\,(c_s - c_{s-1}) + b_s^{\text{tgt}} \tag{2}$$

The decoder reconstructs the *net change* of state, so c_s is free to hold whatever history best predicts the next update: it is the observer's account of the residual stream (Belrose et al., 2023; Ghandeharioun et al., 2024), and we hold it to the standards of an account, namely what it predicts, what it selects, and how it matches native backbone dynamics (validated in Section 3.4). Unrolling the

recurrence writes the state as a retention-discounted sum of past writes (Shen et al., 2019),

$$c_{s,j} = \sum_{t=1}^{s} w_{t,j}\, R_j(t,s), \qquad R_j(t,s) = \prod_{u=t+1}^{s} f_{u,j}, \quad R_j(s,s) = 1 \tag{3}$$

Where $R_j(t,s)$ is the fraction of a write made at depth t that survives to depth s. A coordinate's trajectory therefore has an explicit history: where it was written, how fast it attenuates, and what it still contributes. We summarize such histories with three lifespan metrics (census, effective, and mass-weighted residence), defined and compared against the unordered baseline in Appendix A.

2.4 From History to Semantic Evidence

The recurrence makes semantic readout exactly attributable to writes. Let J be the coordinates used by a frozen nearest-centroid word-sense readout (Raganato et al., 2017; Alain & Bengio, 2017) ($m = |J|$), $u_{s,j} = (c_{s,j} - \mu_j)/\sigma_j$ the training-standardized state, and $p_g, p_f \in \mathbb{R}^m$ the gold and foil centroids. The mean-square-distance margin is linear in the state,

$$M_s = \frac{\|u_s - p_f\|_2^2 - \|u_s - p_g\|_2^2}{m} = b + \sum_{j \in J} a_j\, c_{s,j}, \qquad a_j = \frac{2(p_{g,j} - p_{f,j})}{m\,\sigma_j} \tag{4}$$

With a calibration constant b that does not depend on the trajectory. Substituting $c_s - c_{s-1} = w_s - e_s$ (Eq. 2) and the unrolled state $c_{s,j} = \sum_{t \le s} w_{t,j} R_j(t,s)$ (Eq. 3) gives

$$M_s - M_{s-1} = \sum_{j \in J} a_j\,(w_{s,j} - e_{s,j}), \qquad M_s = b + \sum_{j \in J} \sum_{t \le s} \underbrace{a_j\, w_{t,j}\, R_j(t,s)}_{\gamma_{t \to s,j}} \tag{5}$$

Every change in sense confidence is the score added by new writes minus the score removed by attenuation, and the margin at any depth decomposes exactly into the surviving contributions $\gamma_{t \to s,j}$ of earlier writes.

2.5 Address: Cumulative Retention Prior and Querying

To organize coordinates by persistence, the dictionary ($d_f = 131{,}072$, following Lieberum et al., 2024) is partitioned into $M = 64$ contiguous slots of 2,048 features that share one input and one forget gate ($W_f, W_i \in \mathbb{R}^{M \times d_{\text{model}}}$); for evaluation and intervention, slots are grouped into four quartile bands Q1–Q4.[1] The ordered observer derives the slot gates from cumulative-softmax envelopes adapted from ordered neurons (Shen et al., 2019), with $\mathrm{cum}(v)_m = \sum_{m' \le m} v_{m'}$:

$$f_s^m = \mathrm{cum}\big(\mathrm{softmax}(W_f x_s + b_s^f)\big), \qquad i_s^m = 1 - \mathrm{cum}\big(\mathrm{softmax}(W_i x_s + b_s^i)\big), \tag{6}$$

$$f_s = \widetilde{f}_s \odot \omega_s + (f_s^m - \omega_s), \qquad i_s = \widetilde{i}_s \odot \omega_s + (i_s^m - \omega_s), \qquad \omega_s = f_s^m \odot i_s^m \tag{7}$$

Where $\omega_s = f_s^m \odot i_s^m$ and $\widetilde{f}_s, \widetilde{i}_s \in [0,1]^M$ are free sigmoid gates. Because $\mathrm{cum}(\mathrm{softmax}(\cdot))$ is non-decreasing in slot index, f_s^m forms a rising envelope favoring retention and i_s^m forms a falling envelope regulating writes, with free gates modulating their overlap ω_s. This soft retention prior allows high slots to accept writes while resisting erasure, concentrating 77% of cumulative feature births into Q4 (Appendix B.2). Slot index thereby becomes an address. For a slot subset A and a training-derived discriminability score $r_{a,j}$ for sense a, $J(A) = \bigcup_a \mathrm{TopK}_{j:\ \mathrm{slot}(j) \in A}\, r_{a,j}$, so the score says *what* to select and the address says *where* to look; $J(A)$ is the feature set entering the margin of equation 4.

3 Experiments

Experimental Setup and Baselines. We evaluate DOFM on Gemma-2-2B (Gemma Team, 2024), Qwen3-1.7B (Qwen Team, 2025), Qwen3.5-2B (Qwen Team, 2026), and Gemma-4-E2B (Gemma Team, 2026); held-out and steering evaluations (Sec 3.2–3.3) add observers on Llama-3.2-1B-Instruct and GPT-2 small (App H.1). Baselines and controls: (1) ***layer-wise SAEs*** (Huben et al.,

[1] The cumulative construction below leaves the last master input gate at zero, so slot 63 is never written; the writable Q4 is slots 48–62 (129,024 writable coordinates in total).

2024; Lieberum et al., 2024) with post-hoc matching; (2) a ***static crosscoder*** (Lindsey et al., 2024), our shared no-memory arm; (3) ***DOFM ablations***, without the retention prior (free sigmoid gates, exchangeable slots) and with a scalar decay (EMA); (4) a ***post-hoc address***, the ablation's coordinates ranked by measured lifespan; and (5) ***CAA*** (Rimsky et al., 2024) as the dense steering upper bound (manifests in Appendix B). The prior costs 0.048 FVU and 2.5 points of cross-entropy recovery relative to the ablation (Appendix B.2); Section 3.3 shows what it buys, a known address for the coordinates that carry and steer sense.

3.1 Identity: Cross-Depth Consistency and Tracking

Feature lifecycle analysis requires coordinates to represent consistent directional concepts across disparate contexts and depths. We evaluate this property through two fundamental questions.

Semantic Validity: Consistent Concept or Random Projection? Do shared decoder coordinates capture consistent semantic concepts rather than random residual geometry? Evaluations on lexical ambiguity benchmarks (Miller et al., 2017; Loureiro et al., 2021) confirm semantic consistency (Table 1, Appendix D.1): DOFM coordinates re-activate in same-sense over cross-sense contexts by $8.54\times$ (Gemma-2) and $7.97\times$ (Qwen3.5), quadrupling baseline observer separation ($1.89\times$–$2.15\times$). While static Crosscoders lack recurrence and admit transient activations that degrade retrieval, DOFM couples directional identity with cumulative retention: high slots lock in persistent state while low slots prune transient noise, purifying representations across depth.

Table 1: **Selectivity & Chain Survival.** Within/across retrieval ratios on held-out sentences, and 26-layer chain survival rates across cosine thresholds (Appendix D.1, D.2).

Method / Paradigm	Selectivity ↑		Chain Survival	
	Gemma-2	Qwen3.5	@ 0.70	@ 0.10
Untrained (random)	$2.15\times$	$1.89\times$	–	–
Layer-wise (greedy)	–	–	**0.0%**	36.2%
Layer-wise (Hungarian)	–	–	**0.0%**	0.02%
Crosscoder (static)	$3.37\times$	$3.84\times$	100%	100%
DOFM (ours)	$\mathbf{8.54}\times$	$\mathbf{7.97}\times$	**100%**	**100%**
w/o prior	$3.10\times$	$3.88\times$	100%	100%

Tracking Continuity: Why Shared Decoders Over Post-Hoc Matching? Can post-hoc cosine matching over independent per-layer SAEs reconstruct continuous feature lifecycles? We link 26 per-layer SAEs across adjacent depths using greedy and one-to-one Hungarian cosine matching (Appendix D.2). At the standard 0.70 cosine threshold for shared latents (Paulo & Belrose, 2025), no coordinate forms a full 26-layer chain under either matching rule. Lowering the threshold to 0.10 recovers 36.2% of chains under greedy matching, but only through many-to-one collapse; Hungarian matching retains a single chain. Shared-decoder observers like DOFM sidestep this problem by construction: coordinate j is the column $D_{:,j}$ at every depth, so there is no matching step to fail.

3.2 History: Resolving Semantic Meaning Across Depth

Recurrent state dynamics unroll feature lifecycles into additions, carries, and attenuations. We evaluate how this historical accumulation constructs semantic representations through three questions.

State vs. Write: Depth Accumulation or Isolated Derivation? Does intermediate computation build semantic representations through cross-depth accumulation, or does each layer derive meaning in isolation? Across 2,797 instances from four unified WSD benchmarks (Raganato et al., 2017) (Senseval-2/3, SemEval-2013/2015; Edmonds & Cotton, 2001; Snyder & Palmer, 2004; Navigli et al., 2013; Moro & Navigli, 2015), reading accumulated

Table 2: **State vs. Write (Public WSD AUC).** * denotes Crosscoder (State ≡ Write). St: State, Wr: Write. Coordinates selected on state; write selection yields identical ordering (Appendix E). Reference rows evaluate unweighted accumulation and dense residual probes; observer state matches dense residual accuracy while resolving representations into dated, coordinate-level writes (Appendix I)

Backbone	Cross.	w/o Prior			DOFM (Ours)		
	St (Wr)	St	Wr	Δ	St	Wr	Δ
Gemma-2 (26L)	.5376*	.6273	.5321	$+\mathbf{.0952}$	**.6335**	.5376	$+\mathbf{.0960}$
Qwen3 (28L)	.5236*	.6197	.5361	$+\mathbf{.0836}$	**.6197**	.5319	$+\mathbf{.0878}$
Qwen3.5 (24L)	.5189*	.6089	.5089	$+\mathbf{.1000}$	**.6135**	.5164	$+\mathbf{.0970}$
Gemma-4 (35L)	.5268*	.6303	.5184	$+\mathbf{.1120}$	**.6367**	.5266	$+\mathbf{.1101}$
Reference readouts on Gemma-2-2B (same 2,797 test targets):							
Gate-free running sum of writes ($c_s = \sum_{t \le s} w_t$)					.5855	(-4.8 pts)	
Dense probe on raw residual stream H_ℓ (2,304-d)					**.6350**	(≈ DOFM)	

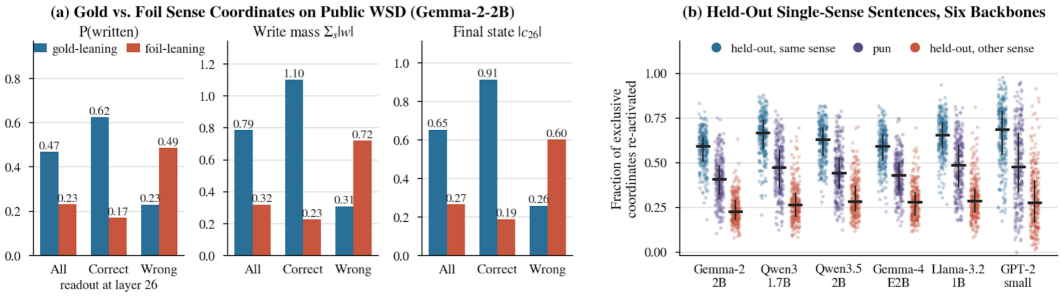


Figure 3: **Disambiguation dynamics: Asymmetric write recruitment without mid-trajectory erasure. (a)** Write probability, write mass $\sum_s |w_{s,j}|$, and final state $|c_{26,j}|$ per candidate coordinate across 2,752 public WSD test instances on Gemma-2-2B, stratified by layer-26 state readout correctness. **(b)** Fraction of sense-exclusive coordinate sets re-activated across held-out single-sense contexts, homographic puns, and cross-sense contexts across six backbones (131 words, two senses). Held-out evaluation sentences remain disjoint from coordinate selection.

observer states outperforms instantaneous writes by $+8.8$ to $+11.0$ points in depth-averaged accuracy across all four backbones (Table 2). Instantaneous write readouts (including crosscoders on ΔH) miss sustained semantic evidence because lexical features deposit early without continuous re-computation: for *bank*, writes to core coordinates cease after Layer 9 (Table 10), causing write classifiers to drift to foils at depths devoid of writes, whereas the carry term $f_s \odot c_{s-1}$ sustains the Layer 2 write of 3.29 through to $c_{26} = 2.54$.

Two baselines isolate the functional contribution of state memory. A gate-free running sum of writes ($c_s = \sum_{t \le s} w_t$) recovers substantial accuracy but trails gated state by 4.8 points (.5855 vs. .6335), demonstrating that adaptive retention contributes beyond unweighted accumulation. Furthermore, a nearest-centroid probe trained on the raw 2,304-dimensional residual stream achieves an accuracy of .635, matching the observer's accumulated state (.634): DOFM incurs zero loss in readout fidelity relative to dense residual representations while resolving representations into dated, coordinate-level writes and lifespans. This holds regardless of coordinate selection: under write-selected coordinates, state still outperforms write by 5.4 points on Gemma-2 (.671 vs. .617; Appendix E).

Disambiguation Dynamics: Asymmetric Writes, No Mid-Trajectory Pruning. How does intermediate computation resolve lexical competition across depth? Two competing mechanisms predict distinct lifecycles: retroactive erasure posits that intermediate layers write competing alternatives and prune losers through attenuation e_s (Swinney, 1979; Elhage et al., 2022); selective generation posits that context biases writes toward the target sense initially (Swinney & Hakes, 1976; Tenney et al., 2019). Population-level write accounting on public WSD confirms selective generation (Figure 3a; Appendix F.1). Across 2,752 targets and their 32 paired Q4 coordinates, gold-leaning coordinates activate with write probability .47 versus .23 for foil-leaning coordinates, delivering $2.5\times$ greater write mass per candidate (.79 vs. .32). Both classes receive writes at identical depths.

This write-time asymmetry governs downstream accuracy: correct instances exhibit a $4.9\times$ gold-to-foil write mass ratio, which inverts on the 770 errors. Write-time recruitment governs this asymmetry: in the persistent band, coordinates of both signs persist without erasure, a pattern preserved in the ablation without the prior where erasure remains below 0.1% across signs (Appendix F.2). In transient band Q1, where erasure removes half of written episodes, attenuation rates match across signs (53.7% gold vs. 49.7% foil), demonstrating that erasure is non-selective with respect to sense. Section 3.4 confirms that the backbone residual stream mirrors these dynamics, retaining carried directions and discarding erased ones. This recruitment bias transfers to unseen contexts across six backbones (Figure 3b): same-sense contexts re-activate a median .59–.69 of exclusive sets, versus .23–.29 for other senses. Lexical disambiguation is determined by asymmetric initial write mass (Table 12); state erasure, where present in transient bands, operates without semantic selectivity.

Ambiguity Representation: When Does Superposition Actually Occur? Because context biases initial writes without downstream pruning, genuine lexical ambiguity does sustain concurrent semantic representations in state memory. Figure 4 tests this prediction on *jam*: reference single-

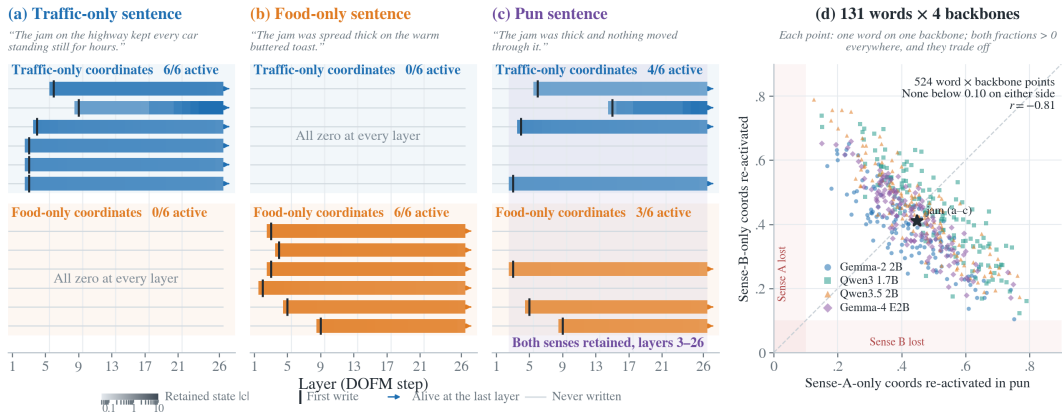


Figure 4: **Sense-exclusive Q4 coordinates sustain concurrent branches under ambiguity. (a–b)** Twelve Q4 coordinates on Gemma-2-2B in single-sense contexts (final $|c| > 0.5$ in one, $|c| < 0.05$ across all depths in the other; six per sense). Each context activates only its corresponding group. **(c)** Under puns, both groups activate by layer 3 and persist to terminal depth. **(d)** Re-activation across 131 words on four backbones: sense-A (x) versus sense-B (y) fractions under puns. Points avoid boundary zones (weakest branch ≥ 0.10; medians 0.40–0.50), trading off along a linear frontier ($r \approx -0.9$; control Jaccard 0.44–0.88).

sense contexts activate distinct traffic or food Q4 coordinates (Panels a,b), whereas the homographic pun (Panel c) co-recruits coordinates from both readings, carrying parallel semantic interpretations to layer 26. This dual-meaning preservation generalizes across 131 words on four backbones (Panel d), where both sense groups re-activate under puns (weakest branch ≥ 0.10) and trade off along a linear frontier ($r \approx -0.9$; Appendix H.1). Single-layer inspection obscures this superposition: on SemEval-2017 Task 7 puns (Miller et al., 2017) against SemCor controls (Miller et al., 1993; 1994) (Appendix F.3), instantaneous writes yield near-zero branch balance, creating an illusion of single-sense collapse. In contrast, accumulated recurrent state sustains both semantic branches (contrast $+0.127$, positive for 10 of 12 words). Intermediate layers commit to one reading when the text is unambiguous and hold two readings in state when the text demands both; superposition is a response to ambiguity, and it lasts to the output.

3.3 Address: Retention Localization and Steering

Cumulative master gates organize coordinates into a retention hierarchy. We evaluate whether this address space localizes semantic representations and enables targeted causal control.

Retention Localization: Can a Single Quartile Replace the Dictionary? Does the retention address localize semantic evidence, or does resolved meaning scatter across the overcomplete dictionary? Across all four backbones, single-band readout rises monotonically from Q1 to Q4 (Gemma-2: .577, .628, .649, .657), where Q4 alone matches full-dictionary accuracy within 0.28 percentage points, and expanding to Q4∪Q3 or Q4∪Q3∪Q2 shifts accuracy by at most 0.003 (Table 16, Appendix G); 71%–97% of top-32 sense coordinates reside in Q4 (Table 17). Without the prior, accuracy is uniform across Q1–Q4 (.658–.670): resolved sense is present, but unaddressed. The retention prior fixes these coordinates into a known subspace for readout and intervention (Appendix H.2). Within the tier, discriminability score $r_{a,j}$ dictates *what* to select while the address specifies *where* to look, enabling coordinate tracking and causal intervention across depth (Appendix B.2).

Causal Actuation: Can Address-Derived Directions Steer Ambiguity? Does semantic readout imply causal control? Can the retention address provide native intervention directions to resolve ambiguous language without external steering vectors? Prior steering frameworks rely on externally fitted contrastive directions (Rimsky et al., 2024; Turner et al., 2023). We test whether DOFM's address space supplies a self-contained causal lever. For each of polysemous words evaluated in Section 3.2, clean single-sense sentences define sense-exclusive Q4 coordinates. We construct an address-derived steering direction from the difference of their state-weighted decoder columns: $\delta =$

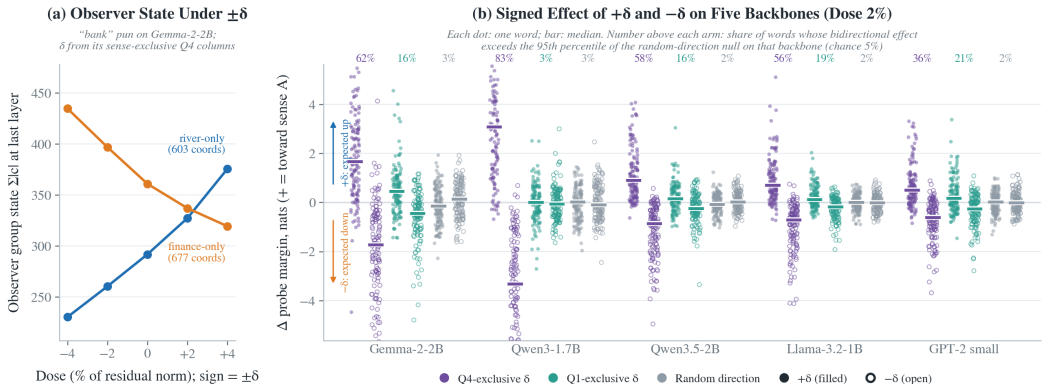


Figure 5: **The observer's own coordinates are a bidirectional lever.** The lever $\delta = D_A c_A - D_B c_B$ combines decoder columns of sense-exclusive Q4 coordinates without fitting or pun access, injected into middle-block residual streams at a fixed norm fraction. **(a)** Internal observer state: the two groups move oppositely and cross near baseline, shifting recorded sense rather than just backbone outputs. **(b)** Gloss-probe margin changes at 2% dose. Across all backbones, $\pm\delta$ steers readings bidirectionally; Q1 directions are weaker, and random vectors do nothing.

$D_A c_A - D_B c_B$, selecting 32 coordinates per sense. Injecting $\pm\delta$ at a small 2% residual dose across intermediate blocks of pun sentences evaluates bidirectional control.

Across all five tested backbones (Figure 5b), 36%–83% of words steer bidirectionally beyond the random null (chance 5%), versus 3%–21% for Q1-exclusive and 2%–3% for random directions, placing causal leverage in the persistent tier. Steered states cross with dose (Figure 5a), flipping the model's pun readings in greedy continuations (Table 3). Causal steering survives ablation: full-dictionary coordinates steer 68% of words with the prior and 69% without it. The prior's key contribution is localization: a predetermined quartile captures 76% of coordinates and 91% of steering efficacy, versus 23% in an arbitrary unconstrained quarter (70% efficacy) and 7% via lifespan sorting (Appendix H.2). The address also transfers when residual injection fails: on Gemma-4-E2B (decoder–encoder cosine -0.71), reading from encoder rows into value projections recovers steering (Appendix H.3). The observer's geometry thus dictates both direction and entry point.

Table 3: **Changing the reading of a pun.** Given only the pun, adding δ toward the other sense shifts it. Only the first clause is shown; see Appendix H.4 for more examples.

			Model Continuation	
Word (A / B)	**Model**	**Pun Sentence**	**Default Sense**	**Steered Sense**
bark (tree / dog)	Gemma-2	The bark was rough and hard to ignore.	the outer layer of the tree	the sound of the dog's voice
volume (sound / book)	Gemma-2	The volume was too much for the room.	the amount of sound	the size of the book
crane (bird / machine)	Qwen3	The crane stretched its long neck over the river.	a bird, not a machine	a big machine that lifts heavy things
suit (clothes / legal)	Qwen3	The suit was expensive and took months to prepare.	the clothes that a person wears	a legal case, not a suit of clothes

3.4 Validating Lifecycle Tracking Against Backbone Dynamics

Because the residual stream is additive, components written at depth t without opposing updates persist in residual vectors, allowing direct cross-validation between observer states and native activations. We track coordinates categorized as *carried* (Q4 episodes in Appendix F.1) and *erased* (Q1 episodes in Appendix F.2), projecting the backbone's residual H_ℓ onto unit-norm decoder directions d_j across layers. Backbone residual activations confirm observer state classifications across depth (Figure 6a). For carried coordinates, projection magnitudes rise across early write layers (depths 3–8) and plateau between $+0.9$ and $+1.3$ through layer 25. For erased coordinates, projection values peak at $+0.7$ within the write window and decay to $+0.16$ by layer 15, re-emerging only when late

Q1 writes occur. Unwritten coordinates drift between $+0.1$ and $+0.3$. The observer's write layer, carry dynamics, and erasure schedule are directly mirrored in native backbone activations.

State carry in the observer reflects native backbone persistence over repeated re-derivation. Projecting the sense-exclusive Q4 subspace of a target word out of the residual stream at single block boundaries $\ell \in \{1, \ldots, 24\}$ induces an immediate 0.4–0.7 nat loss in output sense margin, with disruption compounding across depth without downstream recovery (Figure 6b). The matched control ablates a mismatched word's Q4-exclusive subspace of identical dimension and band, which induces minimal loss (0.0–0.2 nats; paired difference -0.47 nats, with the target-word subspace producing greater degradation in 71% of cases); Q1 subspaces cost 0.1 nats, and random subspaces yield zero impact. The coordinates DOFM assigns to a word's sense are the exact directions the backbone utilizes, sustained through passive carry over repeated recomputation. This establishes what the observer adds over static probes: while probes report the presence of semantic information at depth ℓ, DOFM identifies which coordinates carry it, when they were written, that the backbone preserves them, and that targeted subspace removal degrades model output.

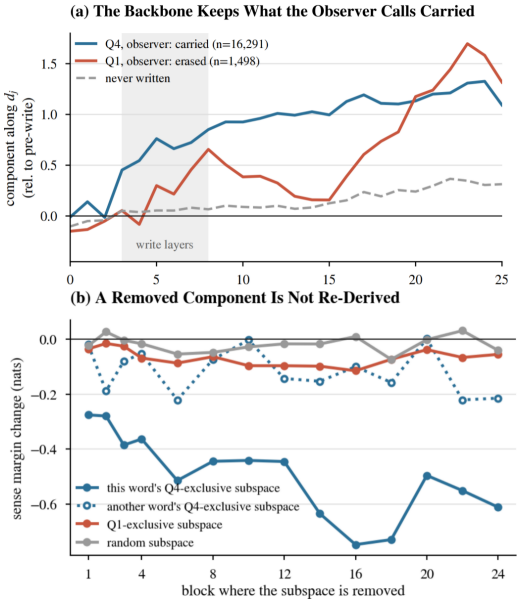


Figure 6: **The residual stream confirms the lifecycles.** (a) Backbone component along d_j for carried, erased, and unwritten coordinates. (b) Sense-margin loss after removing a subspace at one block.

4 Related Work

Sparse dictionaries at one depth and across depths. Single-layer sparse autoencoders (Huben et al., 2024; Gao et al., 2024; Lieberum et al., 2024) lack cross-depth identity, and post-hoc matching fails to form full-depth chains (Paulo & Belrose, 2025). Although crosscoders (Lindsey et al., 2024), transcoders (Dunefsky et al., 2024), and shared-basis dictionaries (Lawson et al., 2025) share identity across layers, they remain static. DOFM introduces a recurrent state that decomposes updates into writes, carries, and erasures. Models of feature timescales along the token sequence (Bhalla et al., 2026) are complementary: DOFM's lifecycle runs along depth within one position and orders the whole dictionary by retention to index readout and intervention.

Reading and structuring intermediate computation. Tools like lenses (nostalgebraist, 2020; Belrose et al., 2023) and probes (Alain & Bengio, 2017; Tenney et al., 2019) evaluate layers in isolation, failing to reveal whether representations are newly written or carried forward. While depth analyses track backbone dynamics (Lad et al., 2025; Sun et al., 2024), our semantic accounting (Section 2.4) attributes readout margins to historical writes, separating write asymmetry from late erasure (Section 3.2). Moreover, while prior ordered representations index importance or granularity (Shen et al., 2019; Bussmann et al., 2025), DOFM adapts cumulative gating to depth recurrence, creating a retention address that indexes persistence.

5 Conclusion

We introduced Depth-Ordered Feature Memory (DOFM), replacing isolated snapshots with continuous state trajectories by unifying directional *identity*, recurrent *history*, and a retention-ordered *address* space. Read through this observer, a transformer resolves a word's meaning early, carries it to the output, settles competition by how much it writes to each sense, and keeps both senses alive when the text is genuinely ambiguous. The retention prior places this resolved state in one persistent tier that can be read on its own and whose coordinates steer the backbone; remove the prior and the same coordinates scatter across the dictionary, where no post-hoc sorting finds them. Beyond lexical ambiguity, DOFM establishes the feature lifecycle as an actionable primitive for mechanistic interpretability, opening horizons for analyzing multi-step reasoning and sequence-level interventions.

AI USE DISCLOSURE

This research utilized generative artificial intelligence tools for manuscript preparation and experimental feature labeling. Automated coordinate labeling in Appendix D.1 queried `gpt-5.6-sol` on top-activating contexts, producing descriptive explanations without impacting quantitative evaluation metrics. Manuscript assistance encompassed sentence revision, section compression for conference page constraints, structural re-organization across method sections, audit of mathematical notations and tensor dimensions, and formalization of diagnostic text regarding baseline failure modes (Appendices D.2, Appendices H). Human authors derived mathematical formulations, designed experiments, executed code implementations, and processed datasets. All empirical findings, model checkpoints, code, and figures originate from physical execution logs, excluding synthetic data generation. The authors conducted human review and verification of all AI-assisted text, mathematical derivations, and arguments, assuming sole responsibility for the final submission.

REPRODUCIBILITY STATEMENT

All experiments evaluate open-weight transformer backbones (Gemma-2-2B, Qwen3-1.7B, Qwen3.5-2B, Gemma-4-E2B, Llama-3.2-1B-Instruct, GPT-2 small) on open-access data: OpenWebText for observer training, alongside the unified WSD benchmark (Senseval-2/3, SemEval-2013/2015) and SemEval-2017 Task 7 for evaluation. Section 2 and Appendix B specify observer architectures, reconstruction objectives, optimizer hyperparameters, precision, and per-backbone parameter counts. Appendix A and Appendix C document operational definitions and thresholds for lifecycle metrics, dataset partitions, repeated-lemma probe formatting, coordinate selection protocols, nearest-centroid classification, and hierarchical cluster bootstrap procedures. Appendix F details write-level replays, band-level stratifications, and unconstrained ablation controls; Appendix H specifies causal steering protocols, intervention doses, target block ranges, success criteria, and address controls; and Appendix I documents backbone residual projections and dense probe comparisons. All source code, checkpoints, and evaluation suites will be released under an open-source license upon paper acceptance.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2017. URL `https://openreview.net/forum?id=ryF7rTqgl`.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Usha Bhalla, Alex Oesterling, Claudio Mayrink Verdun, Himabindu Lakkaraju, and Flavio P. Calmon. Temporal sparse autoencoders: Leveraging the sequential nature of language for interpretability. In *International Conference on Learning Representations*, 2026. Oral presentation.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. *OpenAI Blog*, 2023.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025. URL `https://openreview.net/forum?id=m25T5rAy43`.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable llm feature circuits. *arXiv preprint arXiv:2406.11944*, 2024.

Philip Edmonds and Scott Cotton. SENSEVAL-2: Overview. In Judita Preiss and David Yarowsky (eds.), *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense*

Disambiguation Systems, pp. 1–5, Toulouse, France, July 2001. Association for Computational Linguistics. URL https://aclanthology.org/S01-1001/.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/toy_model/index.html.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. *arXiv preprint arXiv:2401.06102*, 2024.

Aaron Gokaslan and Vanya Cohen. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus, 2019.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Vedang Lad, Jin Hwa Lee, Wes Gurnee, and Max Tegmark. Remarkable robustness of LLMs: Stages of inference? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=Wxh5Xz7NpJ.

Tim Lawson, Lucy Farnik, Conor Houghton, and Laurence Aitchison. Residual stream analysis with multi-layer saes. In *International Conference on Learning Representations*, volume 2025, pp. 33464–33508, 2025.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *arXiv preprint arXiv:2408.05147*, 2024.

Jack Lindsey, Adly Templeton, Jonathan Marcus, Thomas Conerly, Joshua Batson, and Christopher Olah. Sparse crosscoders for cross-layer features and model diffing. *Transformer Circuits Thread*, 2024.

Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and José Camacho-Collados. Analysis and evaluation of language models for word sense disambiguation. *Computational Linguistics*, 47:387–443, 2021. URL https://api.semanticscholar.org/CorpusID:232278102.

George A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. A semantic concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*, 1993. URL https://aclanthology.org/H93-1061/.

George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. Using a semantic concordance for sense identification. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, 1994. URL https://aclanthology.org/H94-1046/.

Tristan Miller, Christian Hempelmann, and Iryna Gurevych. SemEval-2017 task 7: Detection and interpretation of English puns. In Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel Cer, and David Jurgens (eds.), *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 58–68, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2005. URL https://aclanthology.org/S17-2005/.

Andrea Moro and Roberto Navigli. SemEval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In Preslav Nakov, Torsten Zesch, Daniel Cer, and David Jurgens (eds.), *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 288–297, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.18653/v1/S15-2049. URL https://aclanthology.org/S15-2049/.

Roberto Navigli, David Jurgens, and Daniele Vannella. SemEval-2013 task 12: Multilingual word sense disambiguation. In Suresh Manandhar and Deniz Yuret (eds.), *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pp. 222–231, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/S13-2040/.

nostalgebraist. Interpreting GPT: the logit lens, 2020. LessWrong / AI Alignment Forum.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features, 2025. URL https://arxiv.org/abs/2501.16615.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. SemEval-2007 task-17: English lexical sample, SRL and all words. In Eneko Agirre, Lluís Màrquez, and Richard Wicentowski (eds.), *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 87–92, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/S07-1016/.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 2017.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

Yikang Shen, Shawn Tan, Alessandro Sordoni, and Aaron Courville. Ordered neurons: Integrating tree structures into recurrent neural networks. In *International Conference on Learning Representations*, 2019.

Benjamin Snyder and Martha Palmer. The English all-words task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pp. 41–43, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-0811/.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024.

David A. Swinney. Lexical access during sentence comprehension: (re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, 18(6):645–659, 1979. ISSN 0022-5371. doi: https://doi.org/10.1016/S0022-5371(79)90355-4. URL https://www.sciencedirect.com/science/article/pii/S0022537179903554.

David A. Swinney and David T. Hakes. Effects of prior context upon lexical access during sentence comprehension. *Journal of Verbal Learning and Verbal Behavior*, 15(6):681–689, 1976. ISSN 0022-5371. doi: https://doi.org/10.1016/0022-5371(76)90060-8. URL https://www.sciencedirect.com/science/article/pii/0022537176900608.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1452. URL https://aclanthology.org/P19-1452/.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

A Operational Lifecycle Metrics and Baseline Comparison

This section details the operational metrics used to quantify feature lifecycles from observer state trajectories, provides the analytical justification for numerical thresholds, and presents the contrast between DOFM and its ablation without the retention prior.

Active Indicator and Episode Tracking. For coordinate j at depth step s, an active indicator tracks presence:

$$\mathbb{I}_{s,j} = \mathbf{1}\left[|c_{s,j}| > 10^{-6}\right]. \tag{8}$$

The threshold 10^{-6} establishes a numerical zero floor: in single-precision floating-point arithmetic, roundoff noise from recurrent multiplications and additions accumulates at orders 10^{-8} to 10^{-7}; the 10^{-6} bound isolates genuine non-zero writes from hardware numerical drift. A *birth* event marks a $0 \to 1$ transition ($\mathbb{I}_{s-1,j} = 0$ and $\mathbb{I}_{s,j} = 1$). An episode begins at birth and concludes when $\mathbb{I}_{s,j}$ returns to 0 or when computation reaches terminal step S (constituting a right-censored observation). Coordinates can experience multiple distinct birth episodes across depth; this formulation omits debounce smoothing.

Three Complementary Lifespan Metrics. Multiplicative forget gates induce state decay across depth. Evaluating coordinate duration requires a distinction between absolute presence and effective computational influence:

1. **Census Lifespan (L_{census}):** The count of consecutive steps with $\mathbb{I}_{s,j} = 1$ following birth. Census lifespan captures geometric decay tails: forget gates with $f \approx 0.25$ require ten steps to fall below 10^{-6}, inflating observed durations for transient features. For slot m, episode-weighted slot lifespan is $L_m = \sum_{j \in m} \text{alive steps}_j / \sum_{j \in m} \text{births}_j$.
2. **Effective Lifespan (L_{eff}):** To isolate functional activity from geometric decay tails, effective lifespan ends upon magnitude descent below a peak-relative threshold:
$$L_{\text{eff}}(j) = \min\left\{ s > t_{\text{birth}} \;\middle|\; |c_{s,j}| < \max\left(10^{-2},\ 0.01 \cdot \max_u |c_{u,j}|\right)\right\} - t_{\text{birth}}. \tag{9}$$
The $0.01 \cdot \max_u |c_{u,j}|$ term defines a relative descent floor, truncating trajectories when coordinate magnitude decays by two orders of magnitude relative to peak activation. The absolute threshold 10^{-2} establishes an operational noise floor: because decoder columns are unit-norm ($\|D_{:,j}\|_2 = 1$), activations below 0.01 introduce residual perturbations bounded by norm 0.01, which is negligible compared to typical residual stream scales ($\|H\| \approx 10$–30).
3. **Mass-Weighted Residence (τ_{res}):** A continuous, threshold-free physical metric defining accumulated state area divided by write mass:
$$\tau_{\text{res}}(j) = \frac{\sum_{s=1}^{S} |c_{s,j}|}{\sum_{s=1}^{S} w_{s,j}}. \tag{10}$$

A.1 Construct Validity: Functional Multi-Layer Dependence

Input-layer ablation on 4,096 validation rows quantifies depth dependence. For coordinate j and source layer ℓ, perturbation mass $a_{j,\ell} = \mathbb{E}[|\sum_s |c_{s,j}| - \sum_s |c_{s,j}^{(x_\ell=0)}||]$ yields normalized drive $p_{j,\ell} = a_{j,\ell} / \sum_r a_{j,r}$. The effective driving layer count is $\mathrm{eff}_j = \exp(-\sum_\ell p_{j,\ell} \log p_{j,\ell})$.

Table 4 reports Spearman correlation coefficients between driving depth eff_j and census lifespan L_{census} (mean $\pm$ standard deviation across two initialization seeds). Trained DOFM yields correlation $\rho = +0.610 \pm 0.086$. Coordinates with lifespan $L_{\mathrm{census}} \geq 4$ maintain positive correlation $+0.528 \pm 0.090$. DOFM states integrate computation across multiple network depths.

Table 4: **Functional input-layer dependence versus census lifespan L_{census} (Gemma-2-2B).** Values report mean $\pm$ standard deviation.

Metric / Subset	DOFM (Trained)	Observer
Median $\mathrm{eff}/L_{\mathrm{census}}$ ratio	0.27 ± 0.01	1.19 ± 0.04
Fraction with ratio ≥ 1 ($L_{\mathrm{census}} \geq 4$)	0.005 ± 0.001	0.70 ± 0.05
$\rho(\mathrm{eff}, L_{\mathrm{census}})$, all coordinates	$\mathbf{+0.610 \pm 0.086}$	-0.050 ± 0.001
$\rho(\mathrm{eff}, L_{\mathrm{census}})$, $L_{\mathrm{census}} \geq 4$	$\mathbf{+0.528 \pm 0.090}$	-0.132 ± 0.021

A.2 Empirical Retention Organization and State Persistence

Table 5 reports rank correlation values between slot index and empirical duration metrics (L_{census}, τ_{res}, L_{eff}) as mean $\pm$ standard deviation across evaluation seeds. DOFM master gates produce positive rank correlation across all metrics: validation mass-weighted residence attains $\rho = 0.971 \pm 0.006$, and test effective lifespan reaches $\rho = 0.942 \pm 0.008$. The baseline architecture yields coefficients near zero ($\rho \leq 0.113$). Slot index functions as an explicit retention coordinate.

Table 5: **Spearman correlation ρ between slot index and empirical persistence metrics.** Values report mean $\pm$ standard deviation across evaluation seeds.

Evaluation Universe	Duration Metric	DOFM	w/o prior
Validation (4,096 rows)	Census lifespan (L_{census})	0.569 ± 0.015	0.051 ± 0.012
Validation (4,096 rows)	Mass-weighted residence (τ_{res})	$\mathbf{0.971 \pm 0.006}$	0.113 ± 0.015
TEST target tokens (2,797 items)	Census lifespan (L_{census})	0.723 ± 0.014	0.036 ± 0.010
TEST target tokens (2,797 items)	Effective lifespan (L_{eff})	$\mathbf{0.942 \pm 0.008}$	0.026 ± 0.008

Trajectory evaluation across 2,797 test instances confirms state persistence across depth in the persistent band: across 29,532 recorded Q4 sense-coordinate episodes, zero episodes undergo mid-trajectory erasure prior to step S. Target-leaning coordinates are written more often and with more mass than competing ones, and both are retained once written (Appendix F.1); the lower bands erase, but not selectively by sense (Appendix F.2).

B Training Configurations, Parameter Efficiency, and Operating Points

This section documents training configurations, architectural parameter counts across all four backbones, and observer operating points evaluating reconstruction and behavioral fidelity.

B.1 Training Configurations and Parameter Manifests

All observers train on 20M token rows extracted from OpenWebText (Gokaslan & Cohen, 2019), with contexts deterministically truncated to 256 tokens. Optimization uses PyTorch AdamW with learning rate 2×10^{-4}, betas $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$, weight decay 0.01, and no learning-rate schedule. The effective batch size is 1,024 rows, accumulated from micro-batches with exact sample weighting. Gradients are clipped to a maximum norm of 1.0, and decoder columns are renormalized to unit Euclidean norm ($\|D_{:,j}\|_2 = 1$) following each optimizer update.

To balance computational efficiency with numerical stability, residual activations are extracted from frozen backbones in bfloat16 precision and cast to float32 before normalization; observer modules, loss computation, gradients, and optimizer states execute strictly in float32. Source adapters initialize with $A_s \sim \mathcal{N}(0, 0.02^2)$ and $B_s = 0$. Table 6 documents architectural specifications and DOFM parameter counts across backbones.

Table 6: **Architectural specifications and DOFM parameter counts across backbones.** All models train on 20M token rows. The frozen language model backbone is excluded from parameter totals. Observer parameters remain compact (≤ 0.72B), matching single-layer scales rather than scaling with depth.

Backbone Model	Layers	d_{model}	Micro-Batch	DOFM Parameters
Gemma-2-2B	26	2,304	256	0.72B
Qwen3-1.7B	28	2,048	256	0.66B
Qwen3.5-2B	24	2,048	512	0.64B
Gemma-4-E2B	35	1,536	384	0.56B
Scalar EMA Control	26	2,304	256	0.72B

B.2 Reconstruction and Utilization Operating Points

Table 7 reports operating points evaluating reconstruction fidelity, dictionary utilization, and behavioral preservation on Gemma-2-2B. All models operate at an identical nominal candidate selection budget $K = 128$. Observers maintain healthy, stable reconstruction regimes across configurations (FVU .42–.48). Sequential-splice evaluations show that DOFM preserves the backbone's generation across intermediate computation (71.8% cross-entropy recovery, 74.3% for the ablation without the prior), above independent layer-wise SAEs (69.9%) and the static Crosscoder (66.4%).

Analyzing the Ordering Tax. The retention prior has a price. Removing it lowers FVU from $.4842$ to $.4360$ and raises the alive fraction from 57.65% to 98.20%; we call this difference the **ordering tax** and pay it deliberately:

- **What the ablation buys and loses:** Without the prior, all 64 slots carry free sigmoid gates and the observer spends that freedom on reconstruction. It also loses the index: slots become exchangeable, slot index no longer tracks lifespan ($\rho = +0.05$), and the coordinates that carry sense are spread over the dictionary.
- **Architectural Inductive Bias:** In contrast, Ordered DOFM imposes cumulative master gates that enforce monotonic retention across depth. This prior deliberately suppresses lower-retention slots, causing reduced alive occupancy in low slots and yielding the observed 57.65% alive fraction.
- **The Operational Payoff:** In exchange for this slight reconstruction margin, the cumulative prior organizes an otherwise unindexed feature space into a structured, queryable address hierarchy. As demonstrated in Section 3.3, this prior concentrates 77.2% of feature births and 71%–97% of sense-carrying coordinates into Q4, allowing a single quartile to recover full-dictionary readout accuracy within 0.28 percentage points and supply 91% of whole-dictionary steering efficacy (Appendix H.2). Read from its whole dictionary, the ablation reaches the same accuracy and the same steering success; what it lacks is the index, and no post-hoc sorting of its coordinates by lifespan recovers one (Appendix H.2).

DOFM trades a small share of reconstruction for an organized, queryable feature memory.

What the Prior Buys: Structured Hierarchy vs. Unindexed Representations. Evaluating the ablation without the cumulative prior isolates the empirical contribution of the retention address. Without the prior, the dictionary behaves as an unorganized collection (a "bag"): slot index carries no temporal information ($\rho = +0.05$ with lifespan), arbitrary quartiles achieve indistinguishable readout accuracy (.658–.670), and steering coordinates distribute uniformly across slots (23%–32% per quarter). Ranking unconstrained coordinates by empirical lifespan fails to construct an effective address (7% exclusive coordinate retention, 10% steering success), because unconstrained long-

lived features represent rare-firing coordinates governed by sticky forget gates rather than stable semantic concepts.

With the retention prior, the dictionary functions as an indexed hierarchy (a "shelf"): slot index predicts lifespan ($\rho = .57$ on validation, .97 for mass-weighted residence), 77% of births and 71%–97% of discriminative sense coordinates concentrate in Q4, and Q4 alone matches full-dictionary readout accuracy while delivering 91% of whole-dictionary steering efficacy. Full-dictionary readout and steering remain comparable across variants (readout .634 vs. .627; steering 68% vs. 69%); the prior contributes spatial organization rather than raw capacity, at an ordering cost of .048 FVU. This organization enables coordinate lifecycles to be addressed, inspected, and steered by design.

Table 7: **Observer operating points and behavioral preservation (Gemma-2-2B, 26 layers).** All models operate at nominal selection sparsity $K = 128$. FVU measures unexplained variance on normalized residual updates; Alive is the proportion of candidate coordinates active at least once on validation; CE Rec. evaluates sequential-splice cross-entropy recovery across 200 held-out documents (0 = zero ablation, 1 = clean backbone); Slot–life ρ is the Spearman correlation between slot index and mean lifespan; Q4 births is the proportion of birth episodes occurring in the persistent quartile (slots 48–62).

Observer Model	Val. FVU ↓	Alive Fraction	CE Rec. ↑	Slot–Life ρ	Q4 Births
Layer-wise SAEs	**.4203**	.9880	.699	–	–
Crosscoder	.4501	.9579	.664	–	–
DOFM (ours)	.4842	.5765	.718	**+.5689**	**77.2%**
w/o retention prior	.4360	**.9820**	**.743**	$+.0510$	24.8%
Scalar EMA	.5492	.4463	$-.128$	–	–

C Comprehensive Evaluation Protocols, Datasets, and Statistical Inference

This section documents the evaluation universes, probe architectures, coordinate selection formulations, and the clustered statistical inference framework used throughout the paper.

C.1 Word-Sense Disambiguation Universe and Probe Architecture

We evaluate semantic readout using the unified word-sense-disambiguation (WSD) benchmark of Raganato et al. (2017), grounded in WordNet 3.0 senses (Miller, 1995). The evaluation universe comprises three standardized functional splits:

- **Training Partition (SemCor):** Derived from the SemCor corpus (Miller et al., 1993; 1994), comprising 114,137 clean single-label instances used exclusively for coordinate selection and sense centroid estimation.
- **Validation Partition (SemEval-2007 Task 17):** Sourced from the SemEval-2007 coarse-grained English all-words task (Pradhan et al., 2007), providing 238 instances (234 single-gold instances) used for hyperparameter calibration, grid freezing, and sense-direction extraction.
- **Unified Test Suite:** A comprehensive benchmark of 2,797 eligible instances assembled from four standard international WSD evaluation campaigns: Senseval-2 (Edmonds & Cotton, 2001) (953 instances), Senseval-3 (Snyder & Palmer, 2004) (841 instances), SemEval-2013 Task 12 (Navigli et al., 2013) (565 instances), and SemEval-2015 Task 13 (Moro & Navigli, 2015) (438 instances).

A lemma–part-of-speech (POS) group enters the candidate universe only when possessing at least two distinct WordNet senses with at least five clean SemCor training instances per sense. Across the entire public manifest, 1,430 lemma–POS groups satisfy this admission criterion; 734 occur in the unified test suite.

Repeated-Lemma Probe. To extract latent semantic representations without inducing syntactic distortion (Belrose et al., 2023), input sentences embed within a repeated-lemma probe. Long contexts undergo deterministic cropping to 256 tokens centered on the target word. The probe appends

a repeated instance of the target lemma separated by a delimiter token. Observer state trajectories are recorded across depth, and semantic classification readouts evaluate strictly at the final subtoken of the repeated lemma. Across all 2,797 test instances, zero tokenization failures or out-of-context truncations occurred.

C.2 Two Feature Selection Protocols: Raw vs. Standardized

To eliminate post-hoc selection bias, coordinate selection rules freeze on training instances prior to test evaluation:

1. **Raw-Difference Selection Protocol:** Designed to evaluate state versus instantaneous write representations prior to test evaluation. For candidate sense a, observations pool across depths and training instances to rank coordinates by pooled mean difference: $\bar{c}_j(a) - \bar{c}_j(\neg a)$. The top $k = 32$ coordinates per sense are selected, and their union J is standardized using training-set channel means μ_j and standard deviations σ_j.
2. **Standardized Discriminability Selection Protocol:** Applied in retention-region queries and band-restricted analyses (Section 2.5). Coordinates are selected by absolute standardized mean difference:
$$d_j(a) = \frac{\bar{c}_j(a) - \bar{c}_j(\neg a)}{\sqrt{\mathrm{var}_j + 10^{-6}}}. \tag{11}$$
Selecting coordinates via $|d_j(a)|$ eliminates cross-slot activation scale disparities, ensuring that feature recruitment reflects intrinsic discriminability over raw magnitude bias.

Nearest-Centroid Readout and Metric. For selected coordinate union J ($m = |J|$), training states are normalized: $u_{s,j} = (c_{s,j} - \mu_j)/\sigma_j$. Sense centroids $p_a \in \mathbb{R}^m$ are computed by averaging u across training instances belonging to sense a. At test depth s, classification assigns the nearest centroid under standardized Euclidean distance: $\widehat{a}_s = \arg\min_a \|u_s - p_a\|_2$. The primary evaluation metric is depth-averaged accuracy (accuracy-depth AUC):

$$\mathrm{AUC}_{\mathrm{WSD}} = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{N}\sum_{n=1}^{N}\mathbf{1}\left[\widehat{a}_{n,s} = a_n\right]. \tag{12}$$

C.3 Statistical Inference and Clustered Resampling Protocols

All reported confidence intervals and hypothesis tests represent non-parametric bootstrap distributions across $B = 5{,}000$ resamples, constructed to respect the hierarchical dependence structures inherent in linguistic data (Miller et al., 1994).

Hierarchical Cluster Bootstrap (WSD and Steering). Standard independent and identically distributed (i.i.d.) resampling at the token or instance level violates exchangeability assumptions when multiple ambiguous words originate from the same sentence, as intra-sentence tokens share syntactic structure, topic distributions, and residual context. Naive resampling understates sampling variance and inflates Type I error rates.

To ensure conservative and asymptotically valid inference, we define the resampling unit at the source-sentence level. Let the evaluation universe partition into M disjoint sentence clusters $\mathcal{C}_1, \ldots, \mathcal{C}_M$, where cluster $\mathcal{C}_m = \{(x_{m,i}, y_{m,i})\}_{i=1}^{n_m}$ contains all n_m ambiguous target instances embedded within source sentence m. A bootstrap replicate draws M sentence clusters with replacement, preserving intra-cluster token dependencies intact:

$$\mathcal{D}^{*(b)} = \bigcup_{m=1}^{M} \mathcal{C}_{k_m}, \qquad k_m \sim \mathrm{Uniform}(\{1, \ldots, M\}). \tag{13}$$

Furthermore, resampling is stratified across the four constituent benchmark corpora (Senseval-2, Senseval-3, SemEval-2013, SemEval-2015), guaranteeing that each bootstrap resample preserves the empirical dataset mixture proportions. For estimator $\widehat{\theta}$, the $B = 5{,}000$ replicates define the empirical distribution $\{\widehat{\theta}^{*(b)}\}_{b=1}^{B}$, yielding the 95% percentile interval $[\widehat{\theta}^{*}_{(0.025)}, \widehat{\theta}^{*}_{(0.975)}]$.

D IDENTITY VERIFICATION, COORDINATE SELECTIVITY, AND CHAIN SURVIVAL DIAGNOSTICS

This section documents the lexical ambiguity evaluation benchmarks, presents qualitative case studies of coordinate selectivity using verified checkpoint features, and details the geometric breakdown of post-hoc matching across independent layer-wise SAEs.

D.1 CROSS-CONTEXT SELECTIVITY: AMBIGUITY BENCHMARKS AND QUALITATIVE CASES

Benchmark Sources and Evaluation Rationale. The full public WSD benchmark comprises 1,430 admitted lemma–POS groups across fine-grained WordNet synsets (evaluated comprehensively in Section 3.2 and Appendix C). However, fine-grained WordNet senses frequently exhibit semantic overlap (e.g., distinguishing a financial firm from a physical bank building), where residual representations naturally share geometric subspaces. To evaluate directional coordinate identity without semantic confounding, we evaluate polysemous words drawn from established lexical ambiguity and homonymy benchmark suites: the SemEval-2017 Task 7 homographic ambiguity benchmark (Miller et al., 2017) and coarse-grained lexical disambiguation suites (Loureiro et al., 2021; Miller, 1995).

These benchmarks focus on mutually exclusive, coarse-grained semantic concepts spanning distinct categories:

- **Lexical Homographs (Miller et al., 2017):** Words possessing mutually exclusive phonetic or lexical senses (e.g., *bank*, *bat*, *charge*, *crane*, *date*, *draft*, *jam*, *light*, *match*, *pitch*, *scale*, *spring*).
- **Coarse Polysemous Entities (Loureiro et al., 2021):** Words contrasting distinct conceptual categories (e.g., *apple* [technology company vs. fruit], *java* [programming platform vs. geographical island], *mole* [espionage agent vs. animal species], *seal*, *palm*, *bark*, *club*, *bass*).

Across all evaluation words, target senses represent coarse, non-overlapping semantic concepts with paired discriminative evaluation contexts.

Coordinate Selection and Precision Formulation. For target word w and candidate sense a, training instances select up to 24 coordinates $F_{w,a}$ active in at least half of own-sense sentences and zero other-sense sentences, ranked by mean peak state magnitude $|c|$. In held-out test sentence r, a coordinate activates when its write term exceeds 10^{-6} at any depth ($B(r) = \{j : \max_s |w_{s,j}^{(r)}| > 10^{-6}\}$). Precision evaluates on own-sense instances (p_{own}) against cross-sense instances (p_{cross}):

$$p_{\text{own}}(r) = \frac{|B(r) \cap F_{w,a}|}{|F_{w,a}|}, \qquad p_{\text{cross}}(r) = \frac{|B(r) \cap F_{w,b}|}{|F_{w,b}|}. \tag{14}$$

We report the median word-level ratio $p_{\text{own}}/p_{\text{cross}}$.

Automated Semantic Labeling via Frontier Language Models. To characterize identified coordinates without subjective manual interpretation, we adapt the automated feature explanation paradigm of Bills et al. (2023):

1. **Context Extraction:** We scan 4,096 held-out validation documents (OpenWebText) and isolate the top-20 maximal activating text snippets per coordinate (± 15 tokens centered on the active token).
2. **Summary Statistics:** We calculate overall write frequency, token entropy (bits), and the top-10 most frequent activating vocabulary tokens.
3. **Deterministic Explanation Generation:** These 20 exemplar snippets and summary statistics feed into an automated explainer model (`gpt-5.6-sol` via an OpenAI-compatible API endpoint at deterministic temperature 0.0) using a standardized template. The model outputs a functional description (e.g., *"Activates on commercial banking institutions, monetary policy, and interest rate adjustments"*), a structural category, and an explanation confidence score.

Qualitative Analysis with Verified Feature Coordinates. Table 8 presents qualitative case studies using frozen benchmark sentences from the evaluation corpus, tracking verified coordinate indices from our checkpoint audits. In matching semantic contexts, target DOFM coordinates activate during intermediate computation and persist to terminal depth $S = 26$ ($c_{26} > 0$). In competing contexts, the identical coordinate remains unwritten or stays below $|c| < 0.05$ across all depths. These sentence pairs serve illustrative purposes; population evaluations report quantitative selectivity ratios in Table 1 and held-out validation in Figure 3b.

Table 8: **Qualitative case studies of cross-context coordinate selectivity (Gemma-2-2B).** Feature indices represent verified coordinates from checkpoint audits. Each coordinate activates in matching contexts while maintaining silence in competing contexts. In contrast, post-hoc matching across independent layer-wise SAEs terminates within 3–5 layers at standard similarity thresholds (Table 9).

Lemma / Sense	Context Sentence	DOFM Coordinate Response
bank (Finance)	*It accuses 18 major international **banks** of having taken in US lenders on mortgage...*	**Coord #110420** Written at L2, carried to L26 ($c_{26} = 2.54$)
bank (River)	*The **bank** of the river was covered in soft green moss.*	**Coord #110420** Silent ($\|c\| < 0.05$ across all depths)
class (Students)	*Decades of educational research tell us unequivocally that even smaller **classes** have...*	**Coord #119578** Written at L4, carried to L26 ($c_{26} = 2.08$)
class (Category)	*Scientists call the new **class** of genes tumor-suppressors, or simply anti-cancer...*	**Coord #119578** Never written ($w_s = 0$ across all depths)
jam (Traffic)	*The **jam** on the highway kept every car standing still for hours.*	**Coord #121512** Written at L6, carried to L26 ($c_{26} = 3.81$)
jam (Food)	*The **jam** was spread thick on the warm buttered toast.*	**Coord #121512** Never written ($w_s = 0$ across all depths)

D.2 Failure of Post-Hoc Matching Across Independent Layer-Wise SAEs

To examine whether independent dictionaries can recover continuous trajectories post-hoc, we evaluate adjacent feature pairs across 26 source-local SAEs on Gemma-2-2B (5,041 features per block, matching DOFM's total width of 131,072). Across 126,025 adjacent feature pairs across 25 layer boundaries, trained greedy cosine similarity yields a median of 0.0955 and a 95th percentile of only 0.2462 (versus $0.0758/0.0887$ for isomorphic random decoders); Hungarian one-to-one matching yields median 0.0846 against 0.0735.

Chain Survival Across Similarity Thresholds. Table 9 reports surviving 26-layer tracking chains across cosine similarity thresholds. At the prespecified 0.70 operating point—the cosine criterion under which independently trained dictionaries are held to share a latent (Paulo & Belrose, 2025)—**zero full 26-layer chains survive** ($0/5{,}041$ for both greedy and Hungarian matching).

Relaxing the threshold to permissive values (≤ 0.30) reveals two structural failure modes:

1. **Rapid Chain Termination (Hungarian Matching):** Strict one-to-one matching eliminates tracking chains within 3–5 layers due to continuous representation rotation across depth. Even at threshold 0.10, only a single chain survives 26 layers across the entire dictionary (0.02%).
2. **Many-to-One Collapse (Greedy Matching):** Relaxing greedy matching to threshold 0.10 retains 36.16% of chains nominally, but induces severe representational collapse: multiple shallow

features converge onto identical high-norm deep features, reducing target-layer feature coverage to 47.1%.

In contrast, DOFM enforces directional identity through a single unit-norm decoder ($D_{:,j}$). Coordinate j remains invariant across all 26 layers by construction, maintaining 100% continuous trajectory tracking across all thresholds without matching heuristics.

Table 9: **Full-depth tracking chain survival across cosine similarity thresholds (Gemma-2-2B, 26 layers).** Evaluates 5,041 candidate coordinates from Layer 0 across 25 adjacent layer links. At standard thresholds (≥ 0.70), zero complete chains survive under either matching rule (isomorphic random decoders yield zero surviving chains across all thresholds).

Cosine Threshold	Greedy Chains	Greedy (%)	Hungarian Chains
≥ 0.10 (permissive)	1,823	36.16%	1 (0.02%)
≥ 0.20	220	4.36%	0 (0.0%)
≥ 0.30	33	0.65%	0 (0.0%)
≥ 0.50	0	**0.0%**	0 (0.0%)
≥ 0.70 (standard)	**0**	**0.0%**	**0** (0.0%)

Fine-Step Adjacent Update Resolution. To confirm that tracking all 26 adjacent steps captures independent physical computation over coarse sampling, we evaluate adjacent update sign accuracy. Moving from two-layer to one-layer steps incurs a modest $+0.026$ FVU on common targets. The 26-step observer recovers calibration-centered adjacent-update norm contrast signs with 77.8% accuracy (versus 50.0% random chance). The shared no-memory arm attains 78.6%. This confirms that single-layer residual updates carry independent local contrast information, satisfying the depth-resolution prerequisite for reading step-by-step feature trajectories.

E Public WSD State–Write Readout, Sub-Corpora Breakdown, and History Baselines

This section provides a concrete mechanistic walkthrough illustrating the state–write disparity, documents the sub-corpora breakdown across all four test collections for Ordered DOFM, details simple history baselines, and evaluates the physical dominance of the recurrent carry term.

Mechanistic Walkthrough: Layer-Wise Traces on Concrete Instances. To illustrate the layer-wise traces behind the state–write disparity, Table 10 tracks the layer-wise trajectories of sense coordinates across two syntactic categories: the noun *bank* (Coord #110420; financial context) and the verb *stand* (Coord #127225; spatial location context).

The writes are early and few. In the *bank* context every write to the 32 sense lanes falls between layers 2 and 9, and after layer 9 no sense lane is written again. Because an instantaneous write readout evaluates only w_s, it has nothing to read at those layers: it falls back to a default and drifts, flipping *bank* to the river foil at layers 18, 21 and 26, and never landing on the located sense of *stand* at any of the 26 layers. In contrast, DOFM's carry term $f_s \odot c_{s-1}$ sustains the coordinate across the intermediate computation without re-writing ($c_{26} = 2.54$ from a single write of 3.29 at layer 2), preserving the correct state readout at every layer for *bank* and from layer 4 onward for *stand*.

Per-Corpus Breakdown across Test Sets (Ordered DOFM). Table 11 reports State and Write accuracy-depth AUC for Ordered DOFM across all four constituent benchmarks of the unified WSD test universe. Across all test collections—from early Senseval benchmarks to newer SemEval evaluations—the state advantage remains consistent, confirming that depth accumulation is a robust property across distinct text domains.

Selection-Representation Control. Table 2 selects readout coordinates by state discriminability. We repeated the paired gold/foil readout of Appendix F.1 with the 32 coordinates per target selected on the write representation instead (top-32 by $|d|$ within Q4 from write statistics), reading each

Table 10: **Mechanistic walkthrough of state accumulation versus write volatility across depth** Layer-by-layer traces of primary sense coordinates in three frozen contexts: *bank* (Coord #110420, financial), *stand* (Coord #127225, located), and *class* (Coord #119578, students). "Sense lanes written" tracks write activity within the core 32 discriminative coordinates. At L1 (prior to writes), the write readout reflects the baseline centroid distance under unactivated features (Default). At deeper layers (e.g., L18), core sense coordinates receive zero writes ($w_s = 0$); without accumulated memory, the instantaneous write readout is captured by incidental background activations within the broader readout coordinate union J, causing it to fluctuate and drift to competing foils (**Wrong**). In contrast, DOFM's carry term ($f_s \odot c_{s-1}$) sustains the committed early write across intermediate depths ($c_s \geq 1.92$ across all three cases), dominating background activations and preserving the correct semantic classification throughout depth (**Correct**).

Depth	Sense lanes written	Write	Carry	State	Write Readout	State Readout
Case 1: *bank* (noun, financial)		*"It accuses 18 major international **banks** of having taken in US lenders . . . "*				
L1	0 / 32	0.00	0.00	0.00	Default (finance)	Default (finance)
L2	2 / 32	**3.29**	0.00	3.29	**Correct**	**Correct**
L6	3 / 32	0.00	**3.13**	3.13	**Correct**	**Correct**
L12	0 / 32	0.00	**2.96**	2.96	**Correct**	**Correct**
L18	0 / 32	0.00	**2.84**	2.84	**Wrong** (River)	**Correct**
L26	0 / 32	0.00	**2.54**	2.54	**Wrong** (River)	**Correct**
Case 2: *stand* (verb, located)		*"An ancient stone church **stands** amid the fields . . . "*				
L1	0 / 32	0.00	0.00	0.00	**Wrong** (Upright)	**Wrong** (Upright)
L2	1 / 32	**2.08**	0.00	2.08	**Wrong** (Upright)	**Wrong** (Upright)
L6	0 / 32	0.00	**2.07**	2.07	**Wrong** (Other)	**Correct**
L12	0 / 32	0.00	**2.05**	2.05	**Wrong** (Upright)	**Correct**
L18	0 / 32	0.00	**2.02**	2.02	**Wrong** (Other)	**Correct**
L26	0 / 32	0.00	**1.92**	1.92	**Wrong** (Upright)	**Correct**
Case 3: *class* (noun, students)		*"Decades of educational research tell us unequivocally that even smaller **classes** have zero effect on the academic performance of the pupils . . . "*				
L1	0 / 32	0.00	0.00	0.00	Default (students)	Default (students)
L4	5 / 32	**2.56**	0.00	2.56	**Correct**	**Correct**
L6	3 / 32	0.00	**2.51**	2.51	**Correct**	**Correct**
L12	0 / 32	0.00	**2.36**	2.36	**Correct**	**Correct**
L18	0 / 32	0.00	**2.24**	2.24	**Wrong** (Category)	**Correct**
L26	0 / 32	0.00	**2.08**	2.08	**Correct**	**Correct**

representation out with its own training centroids. On Gemma-2-2B (2,752 targets) state leads write in both settings: depth-averaged .689 vs. .595 on state-selected coordinates and .671 vs. .617 on write-selected ones; at layer 26, .720 vs. .564 and .730 vs. .611. The multi-class readout of the main protocol, with each representation given its own per-sense top-32, gives the same ordering (.657 vs. .594 on Q4; .659 vs. .593 on the whole dictionary).

Simple History Baselines: Why Adaptive Gating is Necessary. To test whether simple unweighted accumulation can match gated recurrent state, we evaluate three baseline history mechanisms:

1. **Fixed-Write Cumulative Sum ($c_s = \sum_{t \leq s} w_t$):** This mechanism accumulates writes across depth without forget gates ($f \equiv 1$). While preserving historical additions, the absence of attenuation allows transient syntactic noise to accumulate without bound, degrading downstream readout accuracy by 4.8 percentage points relative to gated state.
2. **Fixed-Write Exponential Moving Average ($c_s = \lambda c_{s-1} + w_s$):** Applying a uniform scalar decay parameter ($\lambda = 0.756$, tuned on validation data) applies uniform treatment across all coordinates. Because it cannot simultaneously sustain persistent semantic concepts across 26

Table 11: **Per-corpus breakdown of public WSD accuracy-depth AUC for Ordered DOFM.** Evaluations report State AUC / Write AUC across all four constituent benchmarks of the unified WSD test universe (2,797 total instances). All paired state-minus-write differences (Δ) exclude zero under 5,000 source-sentence cluster bootstrap iterations ($p < 0.001$).

Test Benchmark (Instances)	Gemma-2-2B	Qwen3-1.7B	Qwen3.5-2B	Gemma-4-E2B
Senseval-2 (953)	.619 / .520	.608 / .521	.625 / .513	.631 / .520
Senseval-3 (841)	.593 / .500	.573 / .487	.555 / .476	.589 / .482
SemEval-2013 (565)	.669 / .568	.662 / .564	.634 / .530	.679 / .553
SemEval-2015 (438)	.697 / .609	.681 / .600	.674 / .584	.685 / .592
Overall AUC (2,797)	**.634 / .538**	**.620 / .532**	**.614 / .516**	**.637 / .527**
State − Write Gap (Δ)	$+.096$	$+.088$	$+.097$	$+.110$

layers while purging local syntactic tokens within 2 layers, uniform decay fails to capture distinct feature lifecycles.

3. **Trained Scalar EMA Observer:** Learning a global scalar retention coefficient within the full training loop (Table 7) produces severe underfitting (FVU .5492, sequential splice CE recovery $-.128$).

These comparisons establish that intermediate feature tracking requires input-adaptive, coordinate-differentiated gating.

Physical Dominance of the Carry Term. Integrating feature mass across depth on the 2,797 WSD test instances confirms that the carry term $f_s \odot c_{s-1}$ accounts for 91.4% of total state mass on Gemma-2-2B. Mean coordinate residence times span 12.16 layers (Gemma-2), 12.44 layers (Qwen3), 9.65 layers (Qwen3.5), and 15.06 layers (Gemma-4). Features persist in background state over active layer-wise re-writing.

F Disambiguation Dynamics: Coordinate Fate Analysis and Counterexample

This section details the origin of the population lifecycle counts, documents the absence of mid-trajectory erasure in the persistent band on the public WSD test suite, presents the *air* counterexample bounding state tracking, and reconciles these observations with write-level accounting (Section F.1), and evaluates identical accounting across transient bands and the ablation without the prior (Section F.2).

Derivation of the 29,532 Observed Lifecycle Episodes. Across the 2,797 public WSD test instances, each target token evaluates $k = 32$ candidate sense coordinates identified from training statistics. Across the 2,752 instances possessing complete sense mappings, this defines a denominator of $2{,}752 \times 32 = 88{,}064$ potential coordinate channels. A coordinate enters the active population only upon an observed birth event ($|c_{s,j}| > 10^{-6}$). Across all test instances, this identifies exactly 29,532 active coordinate episodes:

$$N_{\text{episodes}} = \underbrace{18{,}151}_{\text{Gold-Leaning Episodes}} + \underbrace{11{,}381}_{\text{Foil-Leaning Episodes}} = 29{,}532. \tag{15}$$

Tracking these 29,532 active lifecycles across depth steps $s = 1, \ldots, S$ reveals that Within Q4, the cumulative prior enforces high retention (per-step erasure rates .03–.09, Table 17); that active competitors persist rather than suffer downstream pruning is confirmed by the ablation without the prior, transient bands (Appendix F.2), and native residual stream projections (Section 3.4).

Sense Coordinate Fate Classification. For any target word token, sense-carrying coordinates partition into gold-leaning and foil-leaning subsets based on the sign of their standardized discriminability $d_j(\text{gold})$. We track coordinate trajectories under three mutually exclusive operational outcomes:

1. **Never Written:** Coordinate magnitude remains below active threshold ($|c_{s,j}| \leq 10^{-6}$) across all depths ($w_s = 0$ for all s).
2. **Written and Effectively Erased:** Coordinate activates ($|c_{s,j}| > 10^{-6}$) but subsequently drops below effective lifespan threshold ($\max(10^{-2}, 0.01 \max_u |c_{u,j}|)$) prior to terminal depth S.
3. **Written and Retained:** Coordinate activates and maintains magnitude above effective threshold through to terminal step S ($c_S > 0$).

Table 12 examines coordinate lifecycles across five representative evaluation instances. In unambiguous contexts, competing foil coordinates are rarely recruited: most are never written (2–27 per instance), while the few that activate (1–4) persist alongside target features. Conversely, target representations can operate with extreme sparsity: in the located *stand* context, only one of 29 target coordinates activates (#127225, written 2.08 at layer 2), yet this single persistent feature sustains the correct state readout from layer 4 to terminal depth (Table 10). Crucially, across all instances—including the *air* failure case—the intermediate category (*written and effectively erased*) is strictly zero. Differences across instances thus stem entirely from initial write-time recruitment, never from retroactive pruning (Appendix F.1).

Table 12: **Fate of signed sense-carrying coordinates across depth (Gemma-2-2B).** For each context, coordinates partition by their training discriminability sign into Target-leaning (Gold) and Competing-leaning (Foil). Counts report coordinates that were: *Never Written* ($w_s \equiv 0$) / *Written then Erased* ($c_s \to 0$) / *Written and Retained* ($c_{26} > 0$). In these contexts competing senses receive few or no writes and no coordinate of either sign is erased.

Lemma	Evaluation Context	Target Coordinates (Gold)			Competing Coordinates (Foil)		
		Never	Erase	Retain	Never	Erase	Retain
stand	Located (*church stands amid fields...*)	28	0	**1**	2	0	1
	Upright (*people standing around...*)	0	0	**3**	27	0	2
class	Category (*new class of genes...*)	5	0	**11**	14	0	2
	Students (*smaller classes have...*)	0	0	**16**	12	0	4
air	Atmosphere (*cloud of black smoke into the air*)	4	0	**7**	7	0	14

The *air* Counterexample. The fate census documents the boundary instance where accumulated state underperforms instantaneous writes: the noun *air* in "... sent an evil-looking cloud of black smoke into the **air**" (Senseval-3, gold *air* as atmosphere, foil *air* as gas). Intermediate layers recruit competing coordinates across both senses: out of 32 paired coordinates, 7 of 11 gold-leaning and 14 of 21 foil-leaning coordinates receive writes, zero undergo erasure, and all 21 active coordinates persist to layer 26 (Table 12). Consequently, state readout achieves accuracy at 1 of 26 layers while write readout succeeds at 21 of 26 layers, yielding negative state-minus-write differences across all eight recurrent observer configurations ($-.80$ to $-.20$).

Because intermediate computation possesses no retroactive erasure mechanism to suppress active coordinates once integrated into state, both representations persist to the terminal depth. The accumulated state margin drifts toward the spatial foil, pulling state readout accuracy below instantaneous write accuracy on this single instance. This counterexample establishes a clear boundary: the absence of retroactive erasure means that early write-time ambiguities persist across depth. Early writes to both senses are carried to the output, and the state readout reports exactly that; the instance is the population rule of Appendix F.1 seen from its far end.

F.1 Foil Coordinates Sustain Reduced Writes Without Downstream Erasure: Population Accounting and Held-Out Validation

The lifecycle census in Appendix F records coordinate activation and erasure events without logging explicit write dynamics. Two critical questions require resolution. First, because 11,381 of

the 29,532 active sense-coordinate episodes (38.5%) lean toward the foil sense, the absence of mid-trajectory erasure does not inherently establish write-time exclusion. Second, panels (a,b) in Figure 4 evaluate the identical sentence pair used to define sense-exclusive coordinates, displaying the selection criterion rather than validating generalizability. This section resolves both questions through a comprehensive write-level replay and an independent held-out sentence evaluation.

Population Write Accounting. We replay all 2,797 public WSD test instances on Gemma-2-2B under the frozen evaluation protocol, recording layer-wise write ($w_{s,j}$), erase ($e_{s,j}$), and state ($c_{s,j}$) sequences across the 32 paired Q4 coordinates per target (the top-32 coordinates by $|d_j|$, partitioned by the sign of d into gold- and foil-leaning sets). A coordinate registers a write event when $|w_{s,j}| > 0$ at any depth. Because writes remain zero outside candidate selection, this matches the active threshold $|c_{s,j}| > 10^{-6}$ and reproduces archived census counts (18,151 gold-leaning and 11,381 foil-leaning episodes, zero effective erasures). The nearest-centroid readout (Section 2.4) evaluates over these identical 32 coordinates, enabling exact margin decomposition into individual coordinate contributions, stratified by layer-26 classification accuracy.

Table 13 documents population write statistics across all 2,752 complete-map targets. Gold-leaning coordinates activate with double the write probability of foil-leaning coordinates (.468 vs. .231) and deliver 2.5× greater write mass per candidate (.787 vs. .319). Both coordinate classes receive initial writes at identical depths (mean initial write at layer 6.0 vs. 5.9, with over 92% of write mass preceding layer 12). Once written, both groups persist identically to the final layer (mean lifespan 20.6 vs. 21.0 layers, with zero mid-trajectory erasures).

This write asymmetry dictates downstream readout accuracy: across the 1,982 correctly classified targets, gold coordinates command 4.9× greater write mass than foil coordinates (1.101 vs. .226). Conversely, across the 770 incorrect instances, this ratio inverts (.309 gold vs. .721 foil), providing population-scale confirmation of the mechanism observed in the *air* counterexample. Across all targets, the median foil write mass share is .35 (IQR .10–.69), with 15.5% of targets exhibiting zero foil writes across all 26 depths. The single-sense instances in Table 12 and Figure 4a,b represent the extreme non-activation boundary of this continuous recruitment bias.

Table 13: **Gold- versus foil-leaning sense coordinates on the public WSD test set (Gemma-2-2B, ordered observer).** Per-candidate quantities report totals divided by the candidate coordinate count of that sign; lifespans evaluate over written episodes. Results are stratified by layer-26 nearest-centroid state readout correctness.

	All targets (2,752)		**Read correctly** (1,982)		**Read wrongly** (770)	
	Gold	Foil	Gold	Foil	Gold	Foil
Candidate coordinates	38,761	49,303	23,396	40,028	15,365	9,275
P(written at least once)	.468	.231	.625	.171	.230	.488
Mean first-write layer (written)	6.0	5.9	5.9	6.0	6.3	5.7
Write mass $\sum_s \lvert w_{s,j}\rvert$ per candidate	.787	.319	1.101	.226	.309	.721
Write mass per written coordinate	1.68	1.38	1.76	1.32	1.34	1.48
Final state $\lvert c_{26,j}\rvert$ per candidate	.654	.267	.913	.189	.259	.603
Written episodes / effectively erased	18,151 / 0	11,381 / 0	–	–	–	–

Held-Out Sentence Generalization. To evaluate whether coordinate exclusivity transfers beyond the exemplar sentences used for selection, we authored two additional single-sense evaluation contexts for each of the 131 polysemous words in Appendix H.1 (`words_heldout.json`). Sense-exclusive coordinate sets remain fixed from initial selection; evaluations inspect coordinate activation ($\max_s |c_{s,j}| > 10^{-6}$) across unseen sentences without retraining or reselection.

Table 14 and Figure 3b report re-activation distributions across six backbones. Unseen same-sense sentences re-activate a median .59–.69 of sense-exclusive coordinates, other-sense sentences re-activate .23–.29, and homographic puns span intermediate rates (.41–.49). Same-sense re-activation exceeds cross-sense re-activation across 91%–99% of the 262 coordinate sets on all backbones. Furthermore, the strict exclusivity criterion ($|c| < 0.05$ across all depths) holds on other-sense sentences for a median .72–.77 of coordinates. Feature exclusivity transfers to unseen contexts: write-time exclusion operates as a pervasive probabilistic recruitment bias across model families.

Table 14: **Generalization of sense-exclusive coordinates across held-out evaluation sentences.** Values report median [IQR] across 262 exclusive coordinate sets (131 words $\times$ two senses) for coordinate fractions active in unseen same-sense sentences, homographic puns, and other-sense sentences; "Silent" denotes coordinate fractions maintaining $|c| < 0.05$ across all depths in other-sense contexts; the final column reports the proportion of sets where same-sense activation strictly exceeds other-sense activation.

Backbone	Words	Same sense	Pun	Other sense	Silent	Same > other
Gemma-2 2B	131	0.59 [0.51, 0.64]	0.41 [0.32, 0.48]	0.23 [0.18, 0.29]	0.77	99%
Qwen3 1.7B	131	0.67 [0.59, 0.74]	0.47 [0.36, 0.57]	0.26 [0.20, 0.33]	0.74	97%
Qwen3.5 2B	131	0.63 [0.54, 0.69]	0.44 [0.36, 0.53]	0.28 [0.23, 0.37]	0.72	97%
Gemma-4 E2B	131	0.59 [0.51, 0.66]	0.43 [0.35, 0.50]	0.28 [0.21, 0.33]	0.72	98%
Llama-3.2 1B	131	0.66 [0.58, 0.72]	0.49 [0.37, 0.60]	0.29 [0.22, 0.35]	0.72	97%
GPT-2 small	131	0.69 [0.55, 0.79]	0.48 [0.34, 0.66]	0.28 [0.17, 0.40]	0.72	91%

F.2 Write Asymmetry Across Bands and Without the Retention Prior

The zero-erasure count of Section F.1 was measured inside Q4, the band the prior reserves for retention. To see what the prior contributes to it, we repeat the write-level replay of Section F.1 with the sense coordinates selected, by the same paired top-32 $|d|$ rule, inside each of the other three bands of the ordered observer, and on the ablation without the prior (Table 7), with the coordinates selected from its whole dictionary and from two arbitrary quarters of it. Everything else (test rows, probe, crop, thresholds, the effective-erasure criterion of Appendix A) is unchanged. Figure 7 documents the write probability, write mass per candidate, and the proportion of written episodes erased prior to layer 26 across selections.

The write asymmetry belongs to the backbone's computation as the observer reads it, and the erasure belongs to the prior. Gold-leaning coordinates are written 1.8–$2.4\times$ as often as foil-leaning ones and carry 2.0–$3.0\times$ the write mass in every band and with or without the prior. Erasure follows the band the prior assigns: in Q1 it removes 54% of written gold-leaning and 50% of written foil-leaning episodes before the final layer, in Q2 3.6% and 2.5%, and in Q3, Q4 and without the prior fewer than 0.1% of either sign; within every band both signs have the same lifespan and reach layer 26 at the same rate. The prior therefore separates what is kept from what is discarded, and it is this separation that makes the persistent band readable and steerable (Section 3.3); the sense decision itself is made by the writes.

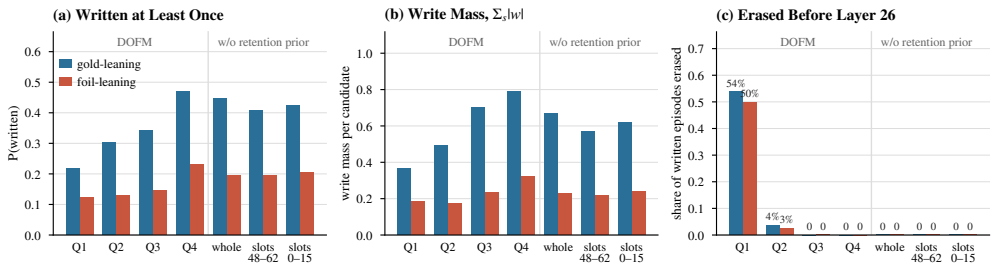


Figure 7: **Write asymmetry holds in every band and without the prior; erasure follows the band and treats both senses alike.** Evaluations track gold- versus foil-leaning coordinates (32 paired coordinates per target, public WSD test set, Gemma-2-2B, 2,752 targets) selected within each band of the ordered observer (Q1–Q4) or from the whole dictionary / two arbitrary quarters of the ablation without the retention prior (slots 48–62, slots 0–15). **(a)** Probability of activation via writes ($P(\text{written})$). **(b)** Write mass $\sum_s |w_{s,j}|$ per candidate. **(c)** Proportion of written episodes erased prior to layer 26 under the effective lifespan criterion. Target coordinates exhibit $1.8\times$–$2.4\times$ higher write probabilities and $2.0\times$–$3.0\times$ greater write mass across all configurations; state erasure is confined to lower bands (Q1, Q2) and attenuates both semantic signs at matched rates.

F.3 Ambiguity Superposition: Puns vs. Single-Sense Controls

To evaluate whether recurrent state sustains concurrent meanings under intentional ambiguity, we evaluate human-written homographic puns from SemEval-2017 Task 7 against single-sense control contexts from SemCor.

Five-Stage Filtering Pipeline for the Core 12-Word Benchmark. SemEval-2017 Task 7 contains 1,607 homographic pun contexts across several hundred candidate lemmas. To construct an unconfounded benchmark for dual-meaning superposition, candidate words undergo a strict five-stage screening pipeline:

1. **Homographic Restriction:** Heterographic puns (exploiting phonological similarity across distinct spellings, e.g., *prophet/profit*) are excluded, as spelling divergence resolves ambiguity at embedding input. Puns must exploit homographs possessing identical surface tokens.
2. **WordNet Sense Grounding:** The target lemma must possess two unambiguous gold sense keys in SemEval-2017 Subtask 3, filtering the pool to approximately 150 well-annotated homographic candidates.
3. **Training Representation Floor:** To ensure reliable identification of sense-specific coordinates ($F_{w,a}$ and $F_{w,b}$), both competing senses must have at least 10–15 clean single-label instances in SemCor, filtering the candidate pool down to 20 lemmas.
4. **Multiple Pun Occurrences:** To prevent single-joke evaluation noise, each selected lemma must appear in multiple independent pun sentences within SemEval-2017 (averaging 3.8 puns per word, isolating 46 total puns).
5. **Balanced Control Pairing:** For each of the 12 resulting lemmas, exactly 8 unambiguous single-sense control sentences are extracted from SemCor (4 clean instances for Sense A and 4 clean instances for Sense B), establishing an exact $12 \times 8 = 96$ fully paired control suite.

This pipeline isolates 12 canonical homographs: *bank*, *bat*, *charge*, *crane*, *date*, *draft*, *jam*, *light*, *match*, *pitch*, *scale*, and *spring*.

Mathematical Formulations of Branch Balance. For sentence r, candidate sense a, and observer depth s, branch activation strength aggregates selected coordinates: $g_{r,a,s} = \sum_{j \in F_{w,a}} |c_{r,s,j}|$. Branch balance evaluates the ratio between the minority and majority branch at the ambiguous target token position:

$$b_r = \mathrm{mean}_{s:\max_a g_{r,a,s}>0} \frac{\min_a g_{r,a,s}}{\max_a g_{r,a,s}}. \tag{16}$$

The ambiguity balance contrast measures the difference between pun sentences and single-sense controls: $\Delta_{\text{ambiguity}} = \text{median}(b_{\text{puns}}) - \text{median}(b_{\text{controls}})$. When evaluated on recurrent state c_s, DOFM attains $\Delta = +0.127$ (Ordered) and $+0.133$ (Free-gated), remaining positive across 10 of 12 words. Replacing state with instantaneous writes collapses balance contrast to $+0.007$ (an 18-fold reduction), while static Crosscoders achieve only $+0.024$.

Relaxed Version with SemEval-2017 Subtask 3. In SemEval-2017 Subtask 3, approximately 150 candidate homographic lemmas possess verified, stable dual WordNet sense keys. Across this broader universe, DOFM recurrent state preserves a robust ambiguity contrast ($\Delta \approx +0.10$), while instantaneous writes remain collapsed near zero ($\Delta_{\text{write}} \approx +0.01$). This confirms that stateful preservation of dual-meaning superposition is an invariant property across broader vocabularies.

G Retention Address Profiles: Functional Band Roles and Readout Staircases

This section documents the single-band and cumulative semantic readout staircases across all four backbones, presents the physical profile of each retention tier, details the functional role of coordinates across the address spectrum, and closes with a frequency-stratified control on what the retention axis does and does not index.

G.1 Single-Band and Cumulative Readout Staircases Across Backbones

Two selections test how semantic readout scales with the retention hierarchy ($k = 32$ coordinates per sense in both cases). Table 15 restricts candidate selection to a single quartile band: on all four backbones, state accuracy-depth AUC rises monotonically from Q1 and saturates at Q4, where one quartile matches selection from the complete 131k dictionary. Table 16 then widens the region from Q4 to the cumulative unions Q4∪Q3 and Q4∪Q3∪Q2: the AUC changes by at most 0.003, so the coordinates that carry resolved sense are already inside Q4.Without the prior the profile is flat across Q1–Q4: the prior decides where resolved sense sits, and the whole-dictionary accuracy is the same with or without it.

Table 15: **Single-band restricted public WSD state readout across backbones (TEST accuracy-depth AUC).** $k = 32$ coordinates per sense selected within each band on training data. Readout rises monotonically from Q1 to a plateau at Q4 across all backbones; without the prior (last row) every band reads alike.

Backbone Model	Q1 (Lowest)	Q2	Q3	Q4 (Highest)	ALL (131k)
Gemma-2-2B (26 layers)	.577	.628	.649	**.657**	.659
Qwen3-1.7B (28 layers)	.473	.562	.605	**.657**	.660
Qwen3.5-2B (24 layers)	.489	.567	.612	**.653**	.650
Gemma-4-E2B (35 layers)	.369	.568	.616	**.672**	.673
w/o retention prior (Gemma-2-2B)	.658	.670	.662	.661	–

Table 16: **Cumulative retention-region selection for semantic readout (TEST accuracy-depth AUC).** Restricting candidate selection to the persistent quartile band (Q4) matches full-pool (131k) readout within 0.28 percentage points across all backbones ($k = 32$ coordinates per sense). Differences report Q4 minus ALL in percentage points.

Backbone Model	Q4	Q4 + Q3	Q4 + Q3 + Q2	ALL (131k)	Difference (pp)
Gemma-2-2B	.6568	.6582	.6586	.6586	−0.180
Qwen3-1.7B	.6574	.6597	.6600	.6600	−0.266
Qwen3.5-2B	.6527	.6504	.6499	.6499	+0.277
Gemma-4-E2B	.6723	.6729	.6729	.6729	−0.060

G.2 Cross-Model lifecycle profiles and Functional Tier Demarcation

This subsection documents the comprehensive physical and semantic profiles across all four evaluated transformer backbones (Gemma-2-2B, Qwen3-1.7B, Qwen3.5-2B, and Gemma-4-E2B). We detail the mathematical formulations for all diagnostic metrics, analyze the physical division of labor across the address spectrum, and establish the foundational interpretation of the retention axis.

Consolidated Cross-Model Profile. Table 17 compiles empirical measurements across the four retention quartiles (Q1–Q4) evaluated on 4,096 held-out validation documents for each backbone. Across disparate model families, layer counts (24 to 35 layers), and hidden dimensions (1,536 to 2,304), the cumulative master gating prior induces identical functional stratification.

Mathematical Derivation of Diagnostic Metrics. Each metric in Table 17 measures a distinct lifecycle dimensions of the recurrent state dynamics:

1. **Token Vocabulary Entropy (Bits):** Measures the lexical specificity of a coordinate. For coordinate j, we identify all token positions across validation text where it undergoes non-zero write activity ($\max_s w_{s,j} > 0$). Let $p_j(v)$ denote the empirical categorical distribution of vocabulary tokens $v \in \mathcal{V}$ observed at these write positions. Vocabulary entropy evaluates:

$$H_j = -\sum_{v \in \mathcal{V}} p_j(v) \log_2 p_j(v). \tag{17}$$

Table 17: **Comprehensive physical and semantic profiles across all four backbones.** Evaluated over active coordinates across 4,096 held-out validation documents for each respective model (Q1 lowest retention to Q4 highest retention). Vocabulary entropy samples 1,024 coordinates per band; erasure rates, update mass shares, and decoded norm shares evaluate across all coordinates. Decoded norm shares are non-additive due to non-orthogonal decoder columns. Sense coordinate metrics evaluate the top-32 standardized discriminability selections from public WSD.

Diagnostic Metric	Tier	Gemma-2	Qwen3	Qwen3.5	Gemma-4
Token vocabulary entropy (bits) Lexical generality	Q1	3.4	2.2	1.0	0.0
	Q2	3.8	2.8	3.0	2.2
	Q3	4.0	2.9	2.8	4.1
	Q4	**5.6**	**5.2**	**5.7**	**6.1**
Mean per-step erasure rate ($1 - f_s$) State clearance speed	Q1	**.75**	**.95**	**.88**	**.91**
	Q2	.33	.41	.36	.38
	Q3	.16	.21	.18	.19
	Q4	**.03**	**.09**	**.05**	**.04**
Erase share of update mass Attenuation vs. carry	Q1	.45	.50	.48	.50
	Q2	.38	.42	.39	.40
	Q3	.32	.31	.30	.29
	Q4	**.11**	**.14**	**.12**	**.10**
Relative decoded norm share Residual update energy	Q1	.10	.08	.07	.06
	Q2	.25	.21	.22	.20
	Q3	.39	.35	.36	.34
	Q4	**.80**	**.87**	**.87**	**.86**
Top-32 sense coordinates (unique) Semantic concentration	Q1	**0.0%**	**0.0%**	**0.0%**	**0.0%**
	Q2	7.1%	0.8%	0.9%	1.1%
	Q3	21.5%	1.9%	2.1%	2.9%
	Q4	**71.4%**	**97.3%**	**97.0%**	**96.0%**
Early-born coordinates in Q4 Share born at steps 1–10 (G_1 enrichment)		**87.0%** (2.2×)	**62.7%** (3.2×)	**87.5%** (2.1×)	**62.4%** (2.3×)

Low entropy (0.0–2.2 bits in Q1) indicates token-specificity: coordinates activate on single surface tokens or punctuation marks. High entropy (5.2–6.1 bits in Q4) indicates breadth of lexical usage rather than semantic abstraction: entropy scales with write frequency ($\rho = 0.885$ across all 20,712 active coordinates on Gemma-2). Stratifying coordinates by write-frequency quintiles shrinks the Q1-to-Q4 entropy divergence from 2.5 bits to 0.2–0.6 bits, with the highest-frequency quintile populated by Q4 coordinates. We interpret the entropy profile as operational generality, consistent with the frequency-controlled context-dependence analysis in Section G.3.

2. **Mean Per-Step Erasure Rate ($1 - f_s$):** Quantifies temporal turnover speed. The forget gate $f_{s,m} \in [0, 1]$ governs state carry. Erasure rate $1 - f_{s,m}$ measures the fraction of carried state attenuated per depth step, averaged across observer steps $s = 1, \ldots, S$ and validation rows. In Q1, erasure rates reach 75%–95%, producing rapid state clearance within 1–2 layers. In Q4, erasure rates drop to 3%–9% (retention $f_s \geq 0.91$), enabling features introduced in early layers to sustain substantial magnitude across dozens of layers without rapid dissipation.

3. **Erase Share of Update Mass:** Measures whether coordinate activity is dominated by new writes or state attenuation. Computed as:

$$\text{EraseShare}_j = \frac{\sum_{s=1}^{S} |e_{s,j}|}{\sum_{s=1}^{S} (|w_{s,j}| + |e_{s,j}|)}. \tag{18}$$

In Q1, attenuation accounts for 45%–52% of total event mass, reflecting rapid write-then-erase cycles. In Q4, erasure accounts for only 10%–14%, demonstrating that coordinate mass is overwhelmingly sustained through passive retention rather than active attenuation.

4. **Relative Decoded-Update Norm Share:** Measures the energetic contribution of a band to the model's residual stream update. For slot subset $B \in \{\text{Q1}, \text{Q2}, \text{Q3}, \text{Q4}\}$, we compute the band-restricted decoded update vector:

$$\Delta \widehat{y}_s^{(B)} = D_{:,B} \left(w_{s,B} - e_{s,B} \right), \tag{19}$$

and report the median ratio $\|\Delta \widehat{y}_s^{(B)}\|_2 / \|\widehat{y}_s^{\text{full}}\|_2$. Because decoder columns are non-orthogonal, these ratios are non-additive. Across all backbones, Q4 commands 80%–87% of total update norm energy, while Q1 contributes less than 10%.

5. **Top-32 Sense Coordinate Share (Unique):** Measures semantic concentration. Using standardized discriminability $|d_j(a)|$ on training data (SemCor), $k = 32$ coordinates are selected per candidate sense. We evaluate the proportion of unique selected coordinates falling within each quartile band. Across all backbones, Q4 captures 71.4%–97.3% of sense-carrying coordinates, while Q1 contains zero (0.0%).

6. **Early-Born Sense Coordinates (Steps 1–10) and Enrichment:** Evaluates the computational origin of persistent state. Q4 coordinates partition by dominant birth depth: G_0 (embedding step 0), G_1 (early intermediate steps 1–10), and G_2 (late steps > 10). We calculate the fraction of top-32 sense selections born in G_1, and compute its enrichment ratio over the background Q4 population:

$$\text{Enrichment} = \frac{P(\text{Sense} \in G_1)}{P(j \in G_1 \mid j \in \text{Q4})}. \tag{20}$$

Across all four backbones, sense coordinates show $2.1\times$–$3.2\times$ enrichment in early intermediate depths (steps 1–10), indicating that persistent sense state is written by early intermediate blocks rather than carried from the embedding (G_0 contributes $< 4\%$); this identifies the computational origin of the state without implying semantic abstraction.

Mechanistic Analysis Across the Retention Hierarchy. Integrating these metrics across the four backbones demonstrates a consistent functional stratification induced by the cumulative gating prior:

- **Q1 (Slots 0–15, Transient Local Scratchpad):** Displays aggressive per-step erasure (75%–95%), negligible write frequencies, and near-zero vocabulary entropy (0.0–2.2 bits on transfer runs). It functions as a rapid scratchpad for local syntax and token-level punctuation, completely devoid of lexical semantic information (0.0% sense features).
- **Q2 (Slots 16–31, Short-Range Syntactic Buffer):** Functions at moderate decay ($1 - f_s \approx .33$–$.41$), integrating clause-level syntax, agreement dependencies, and phrase boundaries while contributing 20%–25% of update energy.
- **Q3 (Slots 32–47, Contextual Integration and Plastic Reasoning Engine):** Retains 80%–84% of state across steps, contributing 34%–39% of update energy and hosting 20% of sense coordinates on Gemma-2. Q3 captures contextualized representational transformations prior to final lexical commitment.
- **Q4 (Slots 48–62, Persistent Semantic Repository):** Achieves near-perfect retention (91%–97% per step), broad lexical coverage (5.2–6.1 bits entropy), and commands over 80%–87% of decoded update energy across all model families. Sense coordinates are overwhelmingly born in intermediate layers 1–10 and sustained silently to terminal depth.

G.3 Frequency-Stratified Control: Context Dependence Across Bands

To check that the rise in vocabulary entropy from Q1 to Q4 reflects generality rather than contextual abstraction, we measure a context-dependence score for every coordinate active at least 20 times on validation text: one minus the share of write-magnitude variance explained by token identity,

$$\text{CD}_j = 1 - \frac{\sum_{v \in \mathcal{V}} N_v \left(\bar{w}_{j,v} - \bar{w}_j \right)^2}{\sum_n \left(w_j^{(n)} - \bar{w}_j \right)^2}, \tag{21}$$

where $\bar{w}_{j,v}$ is the mean write magnitude when token v is present and $\bar{w}_j$ the global mean; $\text{CD}_j \approx 0$ is a rigid token detector, $\text{CD}_j \approx 1$ a coordinate modulated by context independently of the surface

token. Aggregated over bands, median CD rises from Q1 to Q4 on Gemma-2 ($.158 \rightarrow .216$) and Gemma-4 ($.100 \rightarrow .231$) but not on Qwen3 ($.425 \rightarrow .234$) or Qwen3.5 ($.192 \rightarrow .184$), and the apparent gradient is confounded with activation frequency: partitioning coordinates into within-band write-frequency quintiles removes it (on Gemma-2, Q3 exceeds Q4 in the intermediate quintiles, .203 vs. .177 and .239 vs. .200; on the Qwen backbones Q4 has the lowest context-dependence in the first four quintiles). Context modulation therefore tracks activation frequency, not slot index, which is the basis for the first demarcation above. This does not contradict the sense-selectivity of Q4: the sense-exclusive coordinates used for readout and steering are a small subset chosen for their selectivity, and what the persistent tier contributes is that their state, once written, is carried to the output.

H Pun Coexistence and Address-Selected Steering

This appendix supports Figures 4 and 5. It first describes the word set and the rule that selects sense-exclusive coordinates (Section H.1), then the steering protocol and its population statistics (Section H.2), then the one backbone on which residual injection fails and what recovers it (Section H.3), and finally example sentences and continuations (Section H.4).

H.1 Word Set and Sense-Exclusive Coordinates

Words and sentences. The words are polysemous nouns with two clearly separable senses. For each, two single-sense sentences and one pun sentence were authored for this study and were not used to fit any observer or readout. One further word (*spade*) was dropped because the Qwen tokenizers split it.

Sense-exclusive coordinates. A Q4 coordinate is sense-A-exclusive if its final state satisfies $|c| > 0.5$ in sentence A and $|c| < 0.05$ at every layer in sentence B, and symmetrically for B. The pun is never used for selection; it is only read out afterwards. Coordinates are ranked by cumulative state $\sum_s |c_s|$ in their own sentence; Figure 4a–c shows the top six per sense. The two single-sense sentences also share most of their active Q4 coordinates (Jaccard 0.44–0.88), so exclusivity is a property of the selected subset, not of the sentences. Table 18 gives the statistics behind Figure 4d for all six backbones: both senses are re-activated in the pun for essentially every word, and the two fractions trade off across words with $r \approx -0.8$ to -0.9.

Table 18: **Pun coexistence across backbones.** Number of sense-exclusive Q4 coordinates per sense (median), Jaccard overlap of the two single-sense sentences' active Q4 sets, median fraction of each exclusive group re-activated in the pun, share of words with both fractions > 0, weakest side over words, and across-word correlation $r(\mathrm{A}, \mathrm{B})$.

Backbone	Excl. coords A / B	Jaccard	Re-act. A / B	Both > 0	Weakest side (min)	$r(\mathrm{A}, \mathrm{B})$
Gemma-2-2B	542 / 539	0.58	0.42 / 0.40	100%	0.10	-0.85
Qwen3-1.7B	156 / 152	0.74	0.50 / 0.44	100%	0.12	-0.87
Qwen3.5-2B	306 / 314	0.71	0.45 / 0.44	100%	0.12	-0.91
Gemma-4-E2B	380 / 370	0.64	0.43 / 0.43	100%	0.15	-0.89
Llama-3.2-1B-Inst.	66 / 66	0.78	0.52 / 0.43	100%	0.18	-0.85
GPT-2 small	52 / 53	0.66	0.50 / 0.45	99%	0.00	-0.79

H.2 Steering Protocol and Population Statistics

Directions. The lever is $\delta = D_A c_A - D_B c_B$: the decoder columns of the top 32 sense-A-exclusive coordinates weighted by their final state, minus the same for sense B, normalised to unit length. The Q1 control applies the same rule to slots 0–14 (15 slots, matching the 15 writable slots of Q4; elsewhere Q1 denotes slots 0–15) but ranks by peak rather than final state, because Q1 states are erased before the last layer (median 1–12 exclusive coordinates per sense; none for a third of the words on Qwen3 and for most words on Gemma-4). The random control is a Gaussian unit vector with a per-word seed.

Injection and endpoint. $\pm\,\text{dose} \cdot \|h_t\|\,\hat{\delta}$ is added to the residual input of the blocks spanning 15–70% of depth (Gemma-2: 4–18; Qwen3: 4–20; Qwen3.5: 4–17; Llama-3.2-1B: 2–11; GPT-2 small: 2–8), at every position from the target word to the end of the sequence. The endpoint is the mean token log-probability of the sense-A label minus that of the sense-B label after the probe `<pun>\nThe word "<w>" here is about`, computed from fp32 logits. The bidirectional effect is $E = [\Delta(+\delta) - \Delta(-\delta)]/2$.

Success criterion. Because the response is close to linear at these doses, $+v$ and $-v$ have opposite effects for almost any v, so requiring only the right sign pattern is met by 37–43% of random directions. A word therefore counts as steered when E exceeds the 95th percentile of $|E|$ over random directions on the same backbone, which fixes chance at 5% by construction; requiring instead $|\Delta| \geq 0.5$ nats on both sides gives the same ordering of arms.

Results. Table 19 gives the statistics behind Figure 5b, and Figure 8 the dose response and coordinate budget: the effect grows monotonically with dose on every backbone and saturates by 16–32 coordinates per sense. A second view of the same data asks whether the lever can reverse the reading the model commits to. Given only the pun, the sign of the clean probe margin defines that reading; steering against it ($-\delta$ when the model reads sense A, $+\delta$ when it reads sense B) reverses the sign for the share of words in Table 20, while steering with it reverses the sign for 0–1% of words on the residual-steerable backbones.

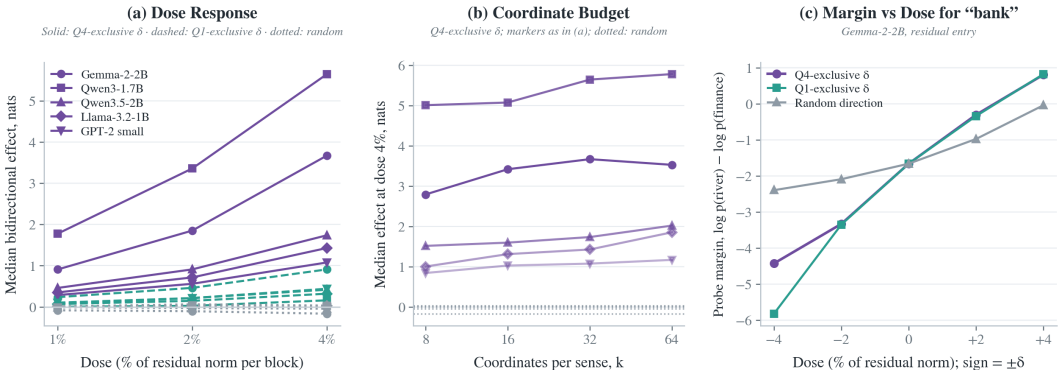


Figure 8: **Dose response, coordinate budget, and a single-word margin curve** (residual entry; the five backbones of Figure 5). (a) Median bidirectional effect at doses 1/2/4% for Q4-exclusive, Q1-exclusive and random directions. (b) Median effect at dose 4% as a function of the number of coordinates per sense; dotted lines are the random references. (c) Probe margin of *bank* on Gemma-2-2B against signed dose; for this word the Q1 direction is as effective as the Q4 direction.

Table 19: **Steering statistics at dose 2% (residual entry).** Median bidirectional effect E (nats) and, in parentheses, the share of words with E above the random-direction null (95th percentile; chance 5%). Residual injection does not steer Gemma-4-E2B (Section H.3); the ablation without the prior and the post-hoc address are in Figure 9. CAA is the unit residual contrast between the two single-sense sentences at the target word, averaged over the injected blocks; it uses the dense residual and serves as the upper bound. On Gemma-4-E2B CAA is evaluated through the value entry (Section H.3).

Backbone	Q4-exclusive δ	Q1-exclusive δ	Random direction	CAA (dense)
Gemma-2-2B	+1.85 (62%)	+0.46 (16%)	−0.10 (3%)	+4.82 (97%)
Qwen3-1.7B	+3.36 (83%)	+0.03 (3%)	+0.06 (3%)	+2.91 (85%)
Qwen3.5-2B	+0.91 (58%)	+0.21 (16%)	−0.03 (2%)	+2.41 (92%)
Llama-3.2-1B-Inst.	+0.71 (56%)	+0.15 (19%)	−0.00 (2%)	+2.08 (97%)
GPT-2 small	+0.56 (36%)	+0.21 (21%)	+0.04 (2%)	+0.87 (52%)
Gemma-4-E2B	−0.05 (0%)	−0.17 (0%)	+0.09 (2%)	–

Table 20: **Share of words whose reading is reversed by steering against it** (residual entry; Q4-exclusive / Q1-exclusive / random directions).

Backbone	Dose 1%	Dose 2%	Dose 4%
Gemma-2-2B	15 / 3 / 2%	27 / 8 / 2%	47 / 24 / 5%
Qwen3-1.7B	33 / 3 / 1%	55 / 4 / 2%	76 / 9 / 11%
Qwen3.5-2B	11 / 3 / 1%	21 / 9 / 3%	40 / 16 / 5%
Llama-3.2-1B-Inst.	8 / 5 / 2%	15 / 9 / 3%	27 / 11 / 4%
GPT-2 small	13 / 7 / 3%	20 / 18 / 6%	37 / 26 / 6%
Gemma-4-E2B	37 / 29 / 40%	39 / 29 / 40%	44 / 14 / 40%

Address controls: what the prior contributes to the lever. The lever is built from sense-exclusive coordinates; the address says where to find them. Three questions follow. Is the lever a property of the retained coordinates themselves, so that it survives removing the prior? Does the prior localize those coordinates, so that one predetermined quarter does the work of the whole dictionary? And can the address be recovered after the fact by ranking an unordered dictionary on measured lifespan? We answer them on Gemma-2-2B with the same 131 words, sentences, probe, blocks, doses, coordinate budget ($k = 32$ per sense) and success criterion as Figure 5, using the ablation without the prior (Table 7) and, for the post-hoc address, its coordinates alive at least once on 512 held-out OpenWebText rows ranked by mean census lifespan, top quarter kept (32,619 coordinates, spread evenly over slots). Figure 9 gives the result.

The lever survives the ablation: exclusive coordinates drawn from the whole dictionary steer 68% of words above the random null with the prior and 69% without it, and reverse the model's reading of 35% and 39% of words at dose 2%. The prior localizes them: in DOFM 76% of the exclusive coordinates lie in Q4, so Q4 alone steers 62% of words, 91% of the whole-dictionary figure, while Q1 holds a median of 13 per sense and steers 16%; without the prior the exclusive coordinates spread evenly, and an arbitrary quarter holds 23–32% of them and steers 48–55%. The address cannot be recovered after the fact: the lifespan-ranked quarter holds 7% of the exclusive coordinates, leaves 19 words with fewer than 32 on one side, and steers 10% of words, because without the prior the longest-lived coordinates are the rarest-firing (12% of them fire on fewer than 0.1% of tokens, against 3% of the rest) and are seldom sense-selective. The cumulative prior, applied during training, is what collects the retained, sense-selective coordinates into one known quarter.

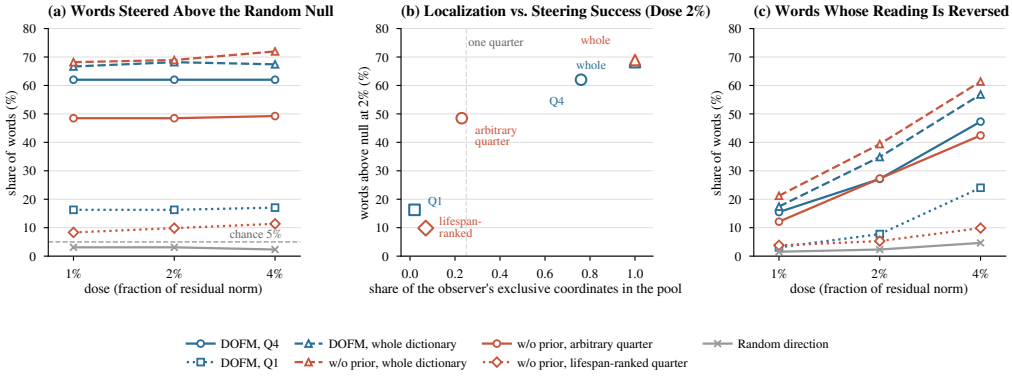


Figure 9: **Address controls for causal steering (Gemma-2-2B, residual injection, $k = 32$ per sense).** Sense-exclusive coordinates are drawn from DOFM (Q4, Q1, whole dictionary) and from the ablation without the retention prior (whole dictionary, an arbitrary quarter, and the top quarter of a post-hoc lifespan ranking); random directions give the null. **(a)** Proportion of words steered above the random-direction null (chance 5%) across doses 1%, 2%, and 4%. **(b)** Steering success at dose 2% plotted against the proportion of the observer's exclusive coordinates contained within the candidate pool: ordered Q4 isolates 76% of coordinates within a single quarter, an arbitrary free-gated quarter captures 23%, and post-hoc lifespan sorting isolates only 7%. **(c)** Proportion of words whose committed default reading is flipped by opposing steering.

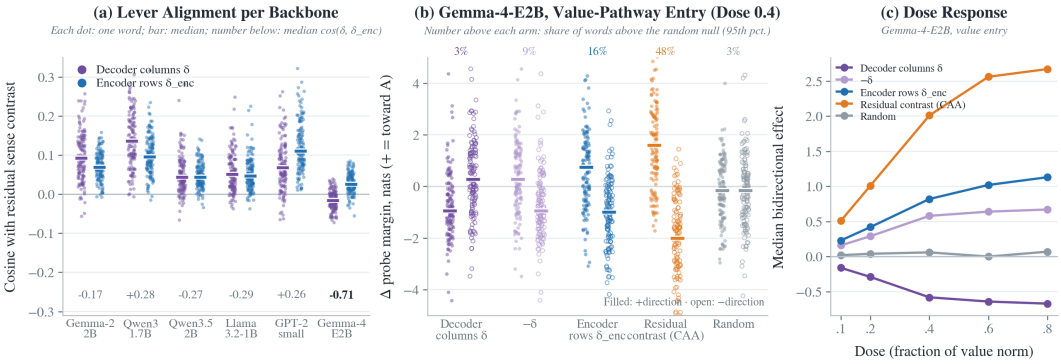


Figure 10: **Gemma-4-E2B.** (a) Cosine between each word's lever and the backbone's residual sense contrast, for the decoder-column lever δ and the encoder-row lever δ_{enc}, on all six backbones; the number below each backbone is the median cosine between the two levers, strongly negative only on Gemma-4-E2B. (b) Signed margin change on Gemma-4-E2B at dose 0.4 of the value norm, injected through the value projections of blocks 5–14, for five direction variants (filled: +direction, open: −direction); numbers give the share of words above the random null. (c) Median bidirectional effect against dose.

Table 21: **Ten examples of the pun words.** Sense A / sense B, the two single-sense reference sentences, and the designed pun sentence; all sentences were authored for this study.

Word (A / B)	Sentence A	Sentence B	Pun
jam (traffic / food)	The jam on the highway kept every car standing still for hours.	The jam was spread thick on the warm buttered toast.	The jam was thick and nothing moved through it.
bank (river / finance)	The bank of the river was covered in soft green moss.	The bank approved the loan after checking her credit history.	The bank was flooded by the end of the week.
bat (animal / sports)	The bat hung upside down in the dark cave all day.	The bat cracked when it hit the fastball over the fence.	The bat swung wildly through the dark.
match (fire / game)	The match burned down to his fingers before he blew it out.	The match ended in a draw after extra time.	The match went out sooner than anyone expected.
charge (electric / fee)	The charge in the battery drained overnight in the cold.	The charge on the bill was higher than the quoted price.	The charge was too much for anyone to handle.
crane (bird / machine)	The crane stood on one leg in the shallow marsh.	The crane lifted the steel beam onto the roof.	The crane stretched its long neck over the river.
suit (clothing / lawsuit)	The suit was tailored in grey wool.	The suit was filed in federal court on Monday.	The suit was expensive and took months to prepare.
pitch (throw / voice)	The pitch sailed high over the catcher's glove.	The pitch of her voice rose as she got excited.	The pitch was perfect and everyone noticed.
litter (trash / animals)	The litter on the beach was collected by volunteers.	The litter of puppies was born in the barn.	The litter was everywhere by morning.
stock (finance / soup)	The stock fell ten percent after the earnings call.	The stock simmered for six hours with bones and herbs.	The stock was rich and slow to reduce.

H.3 GEMMA-4-E2B: WHERE THE LEVER MUST BE READ FROM AND WHERE IT MUST ENTER

Gemma-4-E2B is a dense model whose blocks add a gated per-layer token embedding to the residual stream and share key/value projections across the last 20 blocks; its per-block update norm is of the same order as the residual norm (ratio 0.7–1.8 over the injected blocks, versus 0.4–0.7 on Gemma-2). Two consequences follow (Figure 10). First, additive residual injection is not a usable intervention on this backbone: neither δ nor a CAA-style contrast direction gives sign-consistent effects at any dose (0.25–8%) or block range, random directions at 1% already move the margin by ± 1 nat, and even a gradient-optimised direction only produces degenerate text. Second, the observer's decoder columns for the sense-exclusive coordinates are anti-parallel to its own encoder rows on this backbone (median cosine -0.71, versus -0.17, $+0.28$, -0.27, -0.29 and $+0.26$ on Gemma-2, Qwen3, Qwen3.5, Llama-3.2-1B and GPT-2 small; Figure 10a): the observer detects a feature along a direction u and decodes an update close to $-u$, as expected when blocks overwrite rather than accumulate. Accordingly δ is weakly anti-aligned with the backbone's residual sense contrast (-0.02; $+0.04$ to $+0.14$ elsewhere), whereas the encoder-row lever $\delta_{\text{enc}} = E_A c_A - E_B c_B$ is aligned ($+0.03$, positive for 88% of words).

Delivering directions through the value projections of the K/V-producing blocks (5–14), which the last 20 blocks share, is a well-posed intervention on Gemma-4: an optimised direction gives monotone, coherent steering there, and the CAA contrast mapped by W_v steers 48–65% of words above the random null at doses 0.4–0.8 of the value norm. Through this entry the encoder-row lever recovers a modest but real effect (Figure 10b,c): 16%, 18% and 23% of words above the random null at doses 0.4, 0.6 and 0.8 (random 0–3%), reversing the model's reading for 23–38% of words (random 8–18%), with coherent continuations at 0.4 (*charge*: "the same as tax or fee"; *crane*: "the same as the word bird"; *litter*: "the same as garbage" vs. "a litter of puppies"). The decoder-column lever stays at chance (3%). On the other backbones the encoder-row lever matches or surpasses the decoder-column lever (above the null at dose 2%: 63/62% on Gemma-2, 81/83% on Qwen3, 58/58% on Qwen3.5, 52/56% on Llama-3.2-1B, and 53/36% on GPT-2 small, where it is stronger), so "the directions the observer reads" represents a robust definition across backbones: where decoder and encoder agree the two coincide, and where the architecture makes them disagree the read direction is the one that actuates. Readout on Gemma-4 is unaffected (Table 16).

H.4 EXAMPLE SENTENCES AND CONTINUATIONS

Stimulus design and authoring principles. The 131 stimulus triplets were authored specifically to ensure that the pun sentence genuinely activates both meanings under zero context prior to steering, rather than strongly biasing toward one dominant interpretation. Table 21 provides ten representative examples illustrating this construction across diverse semantic domains (concrete objects, social interactions, and abstract attributes). For each word:

- **Sentences A and B** serve as non-overlapping controls. They contain unambiguous lexical triggers (e.g., *highway/car* vs. *warm buttered toast* for *jam*) designed to isolate the target coordinate subspaces without residual overlap.
- **The Pun sentence** synthesizes syntactic constraints from both readings while omitting definitive resolving tokens (e.g., in *"The jam was thick and nothing moved through it"*, the property of being *thick* applies naturally to spreadable food, whereas *nothing moved through it* satisfies the dynamic condition of vehicular gridlock).

Context representation and probe formatting. To prevent recency artifacts, each backbone processes the entire sequence up to the token prior to classification before the probe prompt is introduced. The baseline generations demonstrate that unsteered backbones typically commit to a single deterministic reading, whereas our address-selected steering reliably retrieves the suppressed semantic branch without destabilizing syntax.

Table 21 summarizes the design for ten illustrative cases. The subsequent Table 22 demonstrates how steering along the identified Q4-exclusive coordinates flips the model's committed interpretation across multiple model families.

Table 22: **Changing the reading of a pun: further examples.** Same construction as Table 3: given only the pun, each backbone commits to one sense (**orange**); adding δ toward the other sense shifts it (**blue**; dose in gray). Words are chosen not to overlap with the main text; first clause shown.

			Model Continuation	
Word (A / B)	**Model**	**Pun Sentence**	**Default Sense**	**Steered Sense**
bass (fish / music)	Gemma-2	The bass was deep and hard to catch.	the size of the fish	the sound of the bass guitar
star (celebrity / celestial)	Gemma-2	The star was bright and everyone was looking.	the star of the show, the person who is the center of attention	the star of Bethlehem
punch (hit / drink)	Gemma-2	The punch was strong and knocked people out.	the action of hitting someone with a fist	the drink
calf (animal / leg)	Gemma-2	The calf was sore after the long walk.	the animal, not the human	the muscle, not the animal
mole (animal / spy)	Qwen3	The mole had been hiding underground for years.	a type of animal, not a unit of measurement	a person who is a spy or a traitor
pen (writing / enclosure)	Qwen3	The pen was left open all night.	a writing instrument, not a person	a place where animals are kept
port (harbor / wine)	Qwen3	The port was ready by evening.	a place where ships come and go	the wine, not the port city
row (line / argument)	Qwen3	The row was long and nobody wanted to be in it.	a line of people or things arranged in a line	a disagreement or argument between two or more people
bass (music / fish)	Qwen3.5	The bass was deep and hard to catch.	the sound, not the instrument	the fish, not the musical instrument
trunk (car / tree)	Qwen3.5	The trunk was old and covered in scratches.	the trunk of the car	the trunk of the tree

I Backbone-Side Validation: Protocols

Residual projections (Figure 6a). For every public WSD test target we store the residual H_ℓ at the probe position for $\ell = 0, \ldots, 26$ (hidden-state index; the observer's step s writes into H_s) and subtract the per-layer mean over the 2,797 test probes. For each of the 32 paired sense coordinates of the write-level replays (Q4 selection, Appendix F.1; Q1 selection, Appendix F.2) we compute $\langle H_\ell - \bar{H}_\ell,\, d_j/\|d_j\| \rangle$, subtract the mean over the three layers before the observer's birth step, and average the curves over episodes whose birth lies at layers 3–8: 16,291 carried Q4 episodes, 1,498 erased Q1 episodes, and 20,000 never-written coordinates given a random birth drawn from the same distribution. The residual norm rises from 68 at layer 8 to 256 at layer 25 and drops to 116 at layer 26; the component along d_j is stable in absolute terms through layer 25 and drops with the norm at layer 26, which is why the figure ends at 25.

Single-block subspace removal (Figure 6b). For each word and sense we take the single-sense sentence, the top-32 sense-exclusive Q4 coordinates (Appendix H.1) and the Q1-exclusive coordinates of the steering control, orthonormalise the corresponding decoder columns, and at the input of one block $\ell \in \{1, 2, 3, 4, 6, \ldots, 24\}$ project the residual of every position from the target word onward out of that subspace. The endpoint is the fp32 gloss-probe margin of Section 3.3. The random control uses a 32-dimensional Gaussian subspace with a per-word seed; the matched control takes the sense-exclusive Q4 coordinates of the next word in the list (same band, same size, disjoint from the word's own set). Medians over 262 cases at blocks 8–24: the word's own Q4 subspace -0.44 to -0.75 nats, another word's -0.00 to -0.22, Q1 -0.04 to -0.11, random -0.07 to $+0.03$; the paired difference between the word's own and another word's subspace is -0.47 nats and favours the word's own in 71% of cases.

Raw-residual probe (Table 2). For each of the 734 test lemma groups we fit a standardized nearest-centroid readout on the raw 2,304-dimensional residual at 13 layers ($\ell = 2, 4, \ldots, 26$), using all single-label SemCor training rows of the group (per-dimension mean and scale over the group's rows, one centroid per sense), and evaluate it on the 2,752 test targets with a complete sense map. Depth-averaged multi-class accuracy is .635 (.636 excluding the final layer; peak .678 at layer 10), against .634 for the observer's state readout on the same targets.