When Attention Collapses: Residual Evidence Modeling for Compositional Inference

Niklas Houba
Institute for Particle Physics and Astrophysics, Department of Physics, ETH Zurich
ETH AI Center, ETH Zurich
nhouba@phys.ethz.ch

Abstract

Compositional inference – the decomposition of observations into an unknown number of latent components – is central to perception and scientific data analysis. Attention-based models perform well when components are approximately separable, as in object-centric vision. Under *additive superposition*, however – where multiple components contribute to every observation – we identify a structural failure mode we term *slot collapse*: multiple slots converge to the same dominant component while weaker ones remain unrepresented. We trace this to a general limitation: attention is *memoryless with respect to explained evidence*. All slots repeatedly operate on the same input without accounting for what has already been explained, so the strongest component dominates what every slot reads, and every slot is drawn toward it. As a result, attention fails to enforce non-redundant allocation under additive superposition. We address this by introducing *residual evidence modeling*, which tracks the remaining explanatory capacity of the input, instantiated via *evidence depletion* – a minimal modification combining multiplicative depletion with an attention bias. In controlled ablations against parallel attention, sequential processing alone, loss-based regularization, and assignment-level alternatives (sparse and optimal-transport normalizations, coverage), evidence depletion, which adds stateful residual tracking to sequential attention, attains the lowest collapse on the synthetic benchmarks (tied with optimal-transport variants on one) and is the only one of the four mechanisms run on LISA that avoids collapse. On synthetic benchmarks, evidence depletion reduces slot collapse by up to an order of magnitude, and the reduction transfers to mixtures of real recorded audio (FUSS). On gravitational-wave source inference for the ESA/NASA LISA mission, under identical architectures, data, and losses, standard attention fails while evidence depletion prevents collapse and enables multi-source posterior estimation. These results show that under additive superposition, residual evidence tracking is the key architectural ingredient – preventing collapse and enabling compositional inference. Code is available at https://github.com/nhouba/when-attention-collapses.

1 Introduction

Many inference problems require decomposing observations into an unknown number of underlying components. This setting arises across perception and scientific data analysis: object-centric visual reasoning [Locatello et al., 2020, Greff et al., 2019], audio source separation [Hershey et al., 2016], spectroscopic mixture analysis, and multi-object tracking in cluttered scenes share a common structure – observations are generated by multiple latent factors whose contributions must be disentangled.

Slot attention [Locatello et al., 2020] provides a powerful mechanism for this problem, representing inputs as a set of latent components that compete via attention [Singh et al., 2022b, Elsayed et al.,

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

2022, Burgess et al., 2019]. It has been highly successful when components are approximately separable, as in object-centric vision: different objects occupy disjoint spatial regions, and tokens can be uniquely assigned to components. However, many domains exhibit a fundamentally different structure – observations are formed through *additive superposition*, where multiple components contribute simultaneously to every token, often with large dynamic range.

In this regime, we observe a systematic failure mode of attention-based decomposition we call *slot collapse*: multiple slots converge to the same dominant component, while weaker components remain unrepresented. We show this arises from a structural limitation of attention: all slots repeatedly attend to the *same full input*, and what each slot reads is dominated by the strongest component, so every slot is drawn toward it. When that component strongly dominates every token, *different tokens* do not give a slot a *different source*: we prove that every reweighting of such tokens yields aligned readouts (Prop. 1), so dissimilar readouts require tokens the component does not dominate, which evidence depletion encourages by steering later slots away from already-claimed tokens. Sequential processing alone and longer (single-seed) training (App. B.1) leave collapse in place; what is missing is state over explained evidence. Standard attention is *memoryless with respect to explained evidence*: the input representation is unchanged across slots and iterations, so nothing prevents redundant allocation. This motivates our approach: resolving slot collapse under additive superposition by introducing an explicit notion of residual explanatory capacity, which we term *residual evidence modeling*. Proposition 1 covers any readout formed as a nonnegative weighting of shared token vectors; we validate the failure mode empirically for slot attention.

We instantiate this principle via *evidence depletion*, a minimal form of stateful attention that enforces competition over residual explanatory evidence. Each token carries a scalar representing the amount of unexplained signal. Slots are processed sequentially, and after each slot attends, the available evidence is multiplicatively reduced, restricting subsequent slots to unexplained regions. All slots observe the same underlying tokens; only the evidence weights change. As a result, early-slot errors redistribute attention without corrupting the signal, unlike classical iterative subtraction which modifies the data. A controlled ablation matched on architecture, data, and losses shows that two ingredients jointly drive the improvement: a multiplicative depletion update and an explicit attention bias toward unexplained tokens (Sec. 3.1).

We validate evidence depletion in three settings. On controlled synthetic benchmarks across three signal types, evidence depletion reduces slot collapse by up to an order of magnitude (Table 1), and, against seven alternative anti-collapse mechanisms, is lowest on two tasks and tied on the third (Table 2). On mixtures of real recorded audio from the FUSS benchmark [Wisdom et al., 2021], evidence depletion reduces collapse from 0.30 to 0.04, so failure mode and mitigation transfer beyond synthetic signals. On gravitational-wave source inference for the LISA mission [LISA Consortium, 2017] – a scientifically demanding problem with high dynamic range and a confusion foreground – we train models with identical architecture, data, and losses, varying only the attention mechanism (Table 3). Sequential processing alone is insufficient; evidence depletion prevents collapse and enables posterior estimation. Combined with conditional normalizing flows [Papamakarios et al., 2021, Rezende and Mohamed, 2015] in the SLOTFLOW framework [Houba et al., 2025], the resulting model produces approximately calibrated amortized posteriors.

Contributions. We make the following contributions:

1. **Failure mode.** We identify *slot collapse* as a structural failure of attention under additive superposition: multiple slots converge to the same dominant component because every slot reads a shared input dominated by it, and we prove that under strong uniform dominance every allocation of attention yields aligned slot readouts (Sec. 2.2, Prop. 1).
2. **Mechanism.** We introduce *evidence depletion*, a minimal modification to attention that counteracts the shared incentive behind redundant allocation via stateful residual tracking. Controlled ablations show that multiplicative depletion and an attention bias both contribute substantially to resolving collapse (Sec. 2.3, 3.1; Prop. 2; App. G).
3. **Validation.** Across synthetic benchmarks, real-world audio mixtures (FUSS), and LISA gravitational-wave inference, we show that sequential attention alone fails while linear evidence depletion lowers mean collapse to ≤ 0.07 on every main benchmark (peak overlap on synthetic tasks and FUSS, cosine overlap on LISA), matching or outperforming loss-based regularization and assignment-level alternatives (sparsemax/entmax, optimal transport, coverage) on

every synthetic task, and of the four mechanisms run at LISA scale it is the only one that avoids collapse (Sec. 3).

2 Method

2.1 Problem formulation

Consider an observation $\mathbf{x} \in \mathbb{R}^{C\times N}$ (C channels, N samples) generated by K parametric sources plus noise:

$$\mathbf{x} = \sum_{k=1}^{K} \mathbf{g}_k(\boldsymbol{\theta}_k) + \mathbf{n}\,, \quad K \sim p(K),\ \boldsymbol{\theta}_k \sim p(\boldsymbol{\theta}),\ \mathbf{n} \text{ with known statistics.} \tag{1}$$

We represent the posterior as a permutation-invariant set of per-source marginals $q_\phi(\boldsymbol{\theta}_k \mid \mathbf{c}_s(\mathbf{x}))$, matched to ground truth via Hungarian assignment, and train an amortized estimator on simulated data via

$$\phi^* = \arg\min_{\phi}\ \mathbb{E}_{p(K)\,p(\boldsymbol{\theta}|K)\,p(\mathbf{x}|\boldsymbol{\theta})}\left[\min_{\sigma}\sum_{k=1}^{K} -\log q_\phi(\boldsymbol{\theta}_k \mid \mathbf{c}_{\sigma(k)}(\mathbf{x}))\right], \tag{2}$$

where σ ranges over injective source-to-slot assignments $\{1,\dots,K\} \to \{1,\dots,S\}$ ($S \geq K$ slots), resolved by Hungarian matching (Sec. 2.4) and $\mathbf{c}_s(\mathbf{x}) \in \mathbb{R}^{d_c}$ is a learned, slot-specific context. The challenge is constructing $\{\mathbf{c}_s\}$ such that each source receives a unique, information-preserving representation. Slot attention is a natural candidate for this decomposition; we now examine why it fails under additive superposition.

2.2 Failure of attention under additive superposition

Slot attention [Locatello et al., 2020] decomposes inputs into a set of latent slots via competitive cross-attention. Each slot is a vector initialized by sampling from a shared learned Gaussian; for I iterations, every slot reads a weighted mean of the token values and is updated by a GRU, with the weights computed jointly for all slots. Given tokens $\{\mathbf{h}_\ell\}_{\ell=1}^{L}$ and slots $\{\mathbf{s}_s\}_{s=1}^{S}$, both in $\mathbb{R}^d$, with queries $\mathbf{q}_s = W_q\mathbf{s}_s$ and keys $\mathbf{k}_\ell = W_k\mathbf{h}_\ell$, attention weights are

$$\alpha_{s\ell} = \frac{\exp(\mathbf{q}_s \cdot \mathbf{k}_\ell/\sqrt{d})}{\sum_{s'}\exp(\mathbf{q}_{s'} \cdot \mathbf{k}_\ell/\sqrt{d})}\,, \tag{3}$$

with normalization over slots for each token. The softmax enforces per-token competition: slots compete to explain individual tokens, and attention weights sum to one over slots. This is a form of *assignment* – it asks "which slot should explain this token?" – and implicitly assumes that tokens can be partitioned across components. In object-centric vision, this approximately holds: different objects occupy disjoint spatial regions, and assignment is sufficient. Under *additive superposition*, however, tokens approximately take the form $\mathbf{h}_\ell \approx \sum_{k=1}^{K}\mathbf{h}_\ell^{(k)}$ (exact in the signal domain; approximate after a nonlinear encoder), with multiple sources contributing simultaneously to each token, so tokens no longer admit a partition into components (Figure 3 in Appendix K illustrates the contrast in the audio domain). Slot attention enforces competition over the assignment of tokens, but not over the allocation of explanatory capacity; under additive superposition, these are not equivalent. The competition redirects losing slots to different tokens, but when every token contains the same dominant source, different tokens do not mean different sources. The input representation remains unchanged across slots and iterations – there is no memory of explained evidence, so nothing prevents multiple slots from claiming the same component.

Definition 1 (Active slots and slot collapse)**.** *Let $\mathcal{A}(\mathbf{x})$ be the set of* active *slots on input $\mathbf{x}$, i.e., slots predicted to correspond to a real source.* Slot collapse *occurs on a multi-source input ($K \geq 2$) when two or more active slots concentrate attention on the same tokens. We measure it by the worst-case similarity of active-slot attention,*

$$\text{Collapse}(\mathbf{x}) = \max_{s\neq s'\in\mathcal{A}(\mathbf{x})} \text{sim}(\alpha_s, \alpha_{s'})\,, \tag{4}$$

averaged over multi-source inputs, with two choices of similarity: peak overlap, $\text{sim} = \mathbf{1}[\arg\max_\ell \alpha_{s\ell} = \arg\max_\ell \alpha_{s'\ell}]$ *(hard attention; the average is the fraction of inputs on which two*

active slots share a peak token), and cosine overlap, $\mathrm{sim} = \cos(\alpha_s, \alpha_{s'})$ *(its soft analogue). The active set is the first K slots by index on the synthetic and FUSS benchmarks, where K is known, and the slots whose existence-head logit (Sec. 2.4) exceeds 0 on LISA.*

We report peak overlap on synthetic/FUSS and cosine overlap on LISA (128 tokens); both give the same ordering of vanilla, sequential, and evidence-depletion attention in every domain (App. F).

Gradient dominance and shared attention dynamics. Consider a token containing contributions from two sources with amplitudes $A_1 \gg A_2$. The attention logits $\mathbf{q}_s \cdot \mathbf{k}_\ell$ are dominated by the stronger component, and gradients with respect to slot parameters scale accordingly: $\partial\alpha_{s\ell}/\partial\mathbf{q}_s \propto \alpha_{s\ell}(1-\alpha_{s\ell}) \cdot \mathbf{k}_\ell$, where $\mathbf{k}_\ell = W_k\mathbf{h}_\ell$ is independent of s, so every slot's update is dominated by the same A_1 direction (a heuristic, for softmax over slots only; App. D). Slots start from distinct samples of the shared prior, so their *parameters* never coincide; what collapses is their *allocation*. Parallel slot attention is permutation-equivariant, and slots interact only through the softmax-over-slots denominator, which arbitrates *which* slot claims a token; Proposition 1 makes precise why redirecting losing slots to other tokens does not help when every token is dominated by the same source direction:

Proposition 1 (Readout alignment under uniform dominance)**.** *Let $\mathbf{h}$ be a unit vector and let the token vectors read by attention decompose as $\mathbf{v}_\ell = a_\ell\mathbf{h} + \mathbf{r}_\ell$ ($\ell = 1, \dots, L$) with $a_\ell > 0$ and per-token dominance ratio $\kappa := \max_\ell \|\mathbf{r}_\ell\|/a_\ell < 1$ (the maximum may be restricted to tokens with positive weight). For any two nonnegative, nonzero weight vectors $w, w' \in \mathbb{R}^L_{\geq 0}$, the readouts $\mathbf{u} = \sum_\ell w_\ell\mathbf{v}_\ell$ and $\mathbf{u}' = \sum_\ell w'_\ell\mathbf{v}_\ell$ satisfy*

$$\cos(\mathbf{u}, \mathbf{u}') \;\geq\; 1 - 2\kappa^2\,,$$

which is tight. For two sources, $\mathbf{v}_\ell = A_1 c_\ell^{(1)}\mathbf{h}^{(1)} + A_2 c_\ell^{(2)}\mathbf{h}^{(2)}$ with unit $\mathbf{h}^{(1)}, \mathbf{h}^{(2)}$, $c_\ell^{(1)} \geq c_{\min} > 0$ and $|c_\ell^{(2)}| \leq c_{\max}$, one has $\kappa \leq \rho := (A_2/A_1)(c_{\max}/c_{\min})$.

Scope and implication. Proposition 1 (proof in Appendix D) concerns readout geometry for fixed tokens, not identifiability or what training finds. Because it holds for every nonnegative weighting – softmax over slots or tokens, sparse, doubly-stochastic, or evidence-depleted – dissimilar readouts require attended tokens with no common dominant direction; evidence depletion steers later slots away from claimed tokens, whereas memoryless attention gives every slot the same incentive. Trained vanilla readouts are far more aligned than the bound requires: although it is loose on our synthetic data (κ median 0.56, bound median 0.37), even the least similar vanilla slot pair has median readout cosine 0.997–0.999 in a magnitude-spectrum surrogate space (evidence depletion: 0.81–0.92), so the alignment is learned, and the proposition describes its idealized limit (Appendix D).

2.3 Sequential attention with evidence depletion

To address this limitation, we introduce a mechanism that explicitly models residual explanatory capacity. Each token ℓ is associated with a scalar $e_\ell \in [\epsilon, 1]$, initialized to $e_\ell = 1$, representing unexplained evidence. Slots are processed sequentially; for slot s, attention is computed as

$$\alpha_{s\ell} = \mathrm{softmax}_\ell\left(\frac{\mathbf{q}_s \cdot \mathbf{k}_\ell}{\tau\sqrt{d}} + \gamma \log e_\ell\right), \qquad \mathbf{k}_\ell = W_k(e_\ell \cdot \mathbf{h}_\ell),\ \mathbf{v}_\ell = W_v(e_\ell \cdot \mathbf{h}_\ell). \tag{5}$$

Note the shift from Eq. (3): because slots are processed sequentially, the softmax now normalizes over *tokens* (each slot selects which tokens to attend to) rather than over slots. Evidence enters through two coupled pathways: the additive $\log e_\ell$ bias suppresses explained tokens, while the scaled inputs $e_\ell \cdot \mathbf{h}_\ell$ reduce information content at claimed locations. After slot s attends, evidence is updated multiplicatively:

$$e_\ell \leftarrow \max\big(e_\ell \cdot (1 - \alpha_{s\ell}^2),\ \epsilon\big)\,, \tag{6}$$

inducing monotonic reduction of available evidence at attended tokens (Eq. 6 shows the quadratic form; we also evaluate linear $(1-\alpha)$, cubic, and binary variants in Sec. 3.1).

Properties. All slots observe the same underlying signal; only the evidence weights change, so errors in early slots redistribute explanatory capacity but do not directly corrupt the input signal seen by later slots. The mechanism is related to but structurally distinct from matching pursuit [Mallat and Zhang, 1993] and CLEAN [Högbom, 1974]: those methods modify the data through subtraction,

Algorithm 1 Sequential slot attention with evidence depletion

Require: Tokens $\{\mathbf{h}_\ell\}_{\ell=1}^{L}$, S slots, I refinement iterations

$\mathbf{s}_s \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ for $s = 1, \ldots, S$

$e_\ell \leftarrow 1$ for $\ell = 1, \ldots, L$ ▷ All signal unexplained

$\pi \leftarrow \text{RandPerm}(S)$ ▷ Random order (training); fixed canonical order at inference

for $s \in \pi$ **do**

  for $i = 1, \ldots, I$ **do**

    $\mathbf{q}_s \leftarrow W_q\,\mathbf{s}_s$; $\mathbf{k}_\ell \leftarrow W_k(e_\ell \cdot \mathbf{h}_\ell)$; $\mathbf{v}_\ell \leftarrow W_v(e_\ell \cdot \mathbf{h}_\ell)$

    $\alpha_{s\ell} \leftarrow \text{softmax}_\ell\big(\mathbf{q}_s \cdot \mathbf{k}_\ell/(\tau\sqrt{d}) + \gamma \log e_\ell\big)$

    $\mathbf{s}_s \leftarrow \text{GRU}(\sum_\ell \alpha_{s\ell}\,\mathbf{v}_\ell,\ \mathbf{s}_s)$; $\mathbf{s}_s \leftarrow \mathbf{s}_s + \text{MLP}(\mathbf{s}_s)$

  end for

  $e_\ell \leftarrow \max(e_\ell(1 - \alpha_{s\ell}^2),\ \epsilon)$ ▷ Deplete (quadratic; see Sec. 3.1 for variants)

end for

return slots $\{\mathbf{s}_s\}$, attention $\{\alpha_{s\ell}\}$, evidence $\{e_\ell\}$

so estimation errors propagate into subsequent iterations as corrupted input; evidence depletion preserves the signal and modulates only attention allocation. In the hard-attention limit ($\tau \to 0$, one-hot α_s), evidence depletion reduces to progressive token masking – analogous to but distinct from classical subtraction, which removes only the estimated component while leaving other superposed components intact. The multiplicative update down-weights tokens that strongly influenced earlier slots, so later slots receive a different effective input and gradient than earlier ones (we discuss when this evidence-space formulation matters empirically in Sec. 3.1).

Proposition 2 (Monotone depletion)**.** *Let* $e_\ell^{(s+1)} = \max\big(e_\ell^{(s)}(1 - f(\alpha_{s\ell})),\ \epsilon\big)$ *with* $e_\ell^{(0)} = 1$ *and a depletion form* $f : [0,1] \to [0,1]$ *(e.g.,* $f(\alpha) = \alpha$ *or* α^2*). Then: (i)* $\epsilon \le e_\ell^{(s+1)} \le e_\ell^{(s)}$ *(monotonic decay); (ii)* $e_\ell^{(s)} \le \max\big(\epsilon, \exp(-\sum_{j<s} f(\alpha_{j\ell}))\big)$*, i.e., evidence decays at least exponentially in the cumulative depletion a token receives; (iii) if* $e_\ell^{(s)} > \epsilon$ *and* $f(\alpha_{s\ell}) > 0$*, then* $e_\ell^{(s+1)} < e_\ell^{(s)}$*; as softmax weights are positive, for any form with* $f(\alpha) > 0$ *for* $\alpha > 0$ *(all evaluated forms except binary) each later slot sees different evidence (keys, values, bias) at every token not yet at the floor.*

Proof. (i) by induction from $1 - f \in [0,1]$ and $e^{(0)} = 1 \ge \epsilon$. (ii) by induction, using $\max(\epsilon, \max(\epsilon, P)(1 - f)) \le \max(\epsilon, P(1 - f))$ and $1 - x \le e^{-x}$. (iii) $e_\ell^{(s)}(1 - f(\alpha_{s\ell})) < e_\ell^{(s)}$, and the maximum is taken against $\epsilon < e_\ell^{(s)}$. □

These properties ensure each slot operates on a different effective input at every token not yet at the floor. In contrast to standard attention, where all slots read the same input and receive gradients dominated by the same component, evidence depletion makes attention progressive: once a token has been strongly attended, its influence on subsequent slots is suppressed, so dominant-heavy tokens cannot repeatedly attract multiple slots. A qualitative visualization of this effect is shown in Figure 4 (Appendix K).

Algorithm 1 summarizes the full procedure.

Sequential processing increases the critical path from $O(I)$ (vanilla, parallel over slots) to $O(S \cdot I)$. On a GH200 this costs $\approx 4\times$ parallel attention at $S{=}5$ and $\approx 31\times$ at $S{=}40$ (Appendix I). To use the resulting slots for downstream inference, we integrate the mechanism into a standard amortized pipeline.

2.4 Integration into amortized inference

The contribution of this work is the attention mechanism; the components below are standard and the mechanism is independent of the downstream prediction head. Each slot's context $\mathbf{c}_s$ is converted into a posterior via a conditional normalizing flow [Rezende and Mohamed, 2015, Papamakarios et al., 2021]: an invertible map $f_\phi : \mathbb{R}^D \to \mathbb{R}^D$ transforms a base Gaussian into $q_\phi(\boldsymbol{\theta} \mid \mathbf{c})$ using rational-quadratic spline coupling layers [Durkan et al., 2019]. Since slot order is arbitrary, we use Hungarian matching [Kuhn, 1955, Carion et al., 2020] to assign sources to slots at each training step; the cost matrix is $C_{ks} = -\log q_\phi(\boldsymbol{\theta}_k^{\text{true}} \mid \mathbf{c}_s)$. Matched slots receive the gradient of the flow's

negative log-likelihood (NLL) of the true parameters (lower is better; we report it as *NLL* throughout); unmatched slots are trained toward the standard normal prior via the flow's base density ($\mathcal{L}_{\text{prior}}$, App. A.3). An *existence head* (MLP $\rightarrow$ logit) predicts whether each slot corresponds to a real source, trained with focal loss [Lin et al., 2020].

3 Experiments

We evaluate evidence depletion in three settings of increasing realism: (i) controlled synthetic benchmarks, (ii) real-world audio mixtures (FUSS), and (iii) LISA gravitational-wave inference. In all settings, slot attention variants share identical architecture and loss functions, varying only the attention mechanism. Figure 1 summarizes slot collapse across all five domains; detailed results follow in each subsection.

3.1 Synthetic benchmarks: isolating the attention mechanism

These experiments isolate the attention mechanism: a regression head replaces the flow (MSE on greedily matched parameters plus an existence loss), and the only auxiliary term is a weak slot-diversity penalty (w_{div}=0.5), identical for all methods. We test on three signal types with variable source count ($K \sim \mathcal{U}\{1, \ldots, 4\}$), one dominant source (amplitude $\sim\mathcal{U}(5, 10)$ vs. $\mathcal{U}(0.2, 1)$ for the others), and additive Gaussian noise: **(A)** overlapping sinusoids (frequency-domain structure), **(B)** Gaussian bumps (time-domain, no periodicity), **(C)** multi-scale mixtures (sinusoids $\times$ Gaussian envelopes). All use N=256 samples, 16 tokens, 5 slots, and are trained for 200 epochs on 5,000 samples (5 seeds each). We compare vanilla SA [Locatello et al., 2020], sequential SA (no residual tracking), and four evidence depletion forms – binary (hard mask at $\alpha > 0.5$), linear $(1-\alpha)$, quadratic $(1-\alpha^2)$, and cubic $(1-\alpha^3)$ – reporting the peak overlap rate of Definition 1. Table 1 shows the two best-performing forms; binary and cubic perform worse (Appendix G).

Three findings:

1. **Sequential SA collapses consistently** (0.81–0.97 across these tasks), confirming that sequential processing alone does not resolve slot collapse.
2. **Evidence depletion reduces collapse**: every depletion form improves on sequential SA, and linear depletion improves on vanilla on every task (Appendix G).
3. **The effect is consistent across domains** – frequency, time, and heterogeneous signal types.

Linear depletion $(1-\alpha)$ achieves the lowest collapse on these benchmarks. Our main LISA experiments use quadratic $(1-\alpha^2)$ for its softer exploration–commitment trade-off; a matched comparison shows that linear depletion also prevents collapse there (Sec. 3.4), so we recommend linear as the default.

Why some depletion forms fare better. Since attention is normalized over tokens, typical weights are well below one, and $(1 - \alpha^p)$ removes less evidence as p grows, pushing later slots less strongly away from claimed tokens; linear depletes most. The binary mask discards the graded residual altogether, so partially explained tokens cannot be shared, and it performs worst (Appendix G). The trend is not strict: cubic beats quadratic on sinusoids.

How the two baselines fail differently. Sequential processing without residual tracking exacerbates collapse on synthetic benchmarks (0.81–0.97 vs. 0.22–0.32 for vanilla) but partially reduces it on LISA (Sec. 3.4). On small problems, vanilla SA's parallel competition can find non-collapsed solutions through soft specialization, whereas sequential SA degenerates to early slots dominating allocation ("first slot wins"). At scale, parallel competition fails entirely (NLL diverges, angular correlation ≈ 0; Fig. 2 in Appendix B.1), while sequential ordering gives each slot an independent gradient step – but neither resolves collapse.

Evidence space vs. data space. We test an architecture-matched data-space variant – identical sequential structure, identical quadratic $(1-\alpha^2)$ depletion, only the variable carrying the residual differs. Tokens are depleted directly, $\mathbf{h}_\ell \leftarrow \mathbf{h}_\ell \cdot (1 - \alpha_{s\ell}^2)$, rather than via a separate evidence variable (Table 1, bottom rows). Without an attention bias, data-space depletion is unstable (std 0.39–0.40, with seeds collapsing entirely). Adding a token-norm bias ($\gamma \log(\|\mathbf{h}_\ell\| / \max_{\ell'} \|\mathbf{h}_{\ell'}\|)$ in

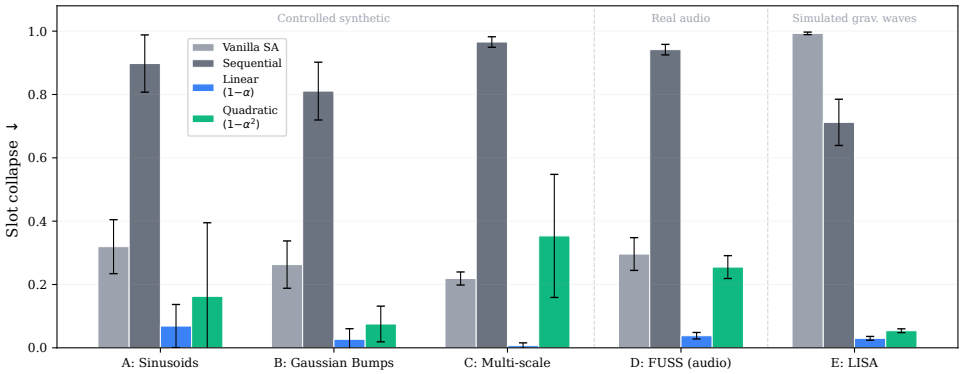

Figure 1: Slot collapse across all five domains (mean ± std over 5 seeds; LISA linear: 3 seeds). Synthetic and FUSS report peak overlap; LISA reports cosine overlap of active slots (Definition 1) on the matched reruns of Table 3, so absolute values are not comparable across the LISA boundary. Linear evidence depletion (blue) has the lowest collapse in every domain; quadratic depletion (green) is used in the main LISA experiments. Data-space variants in Table 1.

Table 1: Slot collapse rate (peak overlap rate; mean ± std, 5 seeds). Synthetic tasks use 200 epochs; FUSS uses 200 epochs on real audio. Data-space variants deplete tokens directly; evidence-space maintains a separate evidence variable.

Method	Sinusoids	Gaussians	Multi-scale	FUSS
Vanilla SA	0.32 ± 0.09	0.26 ± 0.07	0.22 ± 0.02	0.30 ± 0.05
Sequential SA	0.90 ± 0.09	0.81 ± 0.09	0.97 ± 0.02	0.94 ± 0.02
Evidence linear $(1-\alpha)$	$\mathbf{0.07 \pm 0.07}$	$\mathbf{0.03 \pm 0.03}$	$\mathbf{0.01 \pm 0.01}$	$\mathbf{0.04 \pm 0.01}$
Evidence quadratic $(1-\alpha^2)$	0.16 ± 0.23	0.08 ± 0.06	0.35 ± 0.19	0.26 ± 0.04
Data-space quad. (no bias)	0.41 ± 0.40	0.23 ± 0.39	0.36 ± 0.20	—
Data-space quad. (with bias)	0.18 ± 0.30	0.08 ± 0.04	0.35 ± 0.20	—

the attention logits) recovers performance to within noise of evidence-space depletion (e.g., 0.18 vs. 0.16 on sinusoids, 0.08 vs. 0.08 on Gaussians). Thus *multiplicative depletion together with an explicit attention bias is what reduces collapse*, whichever variable carries the residual: the update tracks what has been explained, and the bias makes this state shape allocation. On these tasks the separate evidence variable is therefore not needed to reduce collapse; we keep it because it leaves token representations intact for downstream heads that re-read them (e.g., the LISA source encoder, Appendix A) and makes residual capacity an explicit, interpretable state.

Loss-based regularization is insufficient. We evaluate whether stronger loss-based regularization alone can resolve collapse, testing diversity losses (w_{div}=5), attention orthogonality losses (w_{ortho}=3), and their combination on both vanilla and sequential SA (Appendix H). On vanilla SA, the best loss combination partially mitigates collapse (0.06/0.11/0.06 on sinusoids/Gaussians/multi-scale); evidence depletion (linear) matches or improves on every task (0.07/0.03/0.01) without these additional losses. On sequential SA, loss-based regularization does not resolve collapse: even combining both losses, it remains at 0.30–0.43. Losses treat the symptom: they help the vanilla backbone but not the sequential one, and on LISA vanilla collapses despite diversity, orthogonality, and entropy losses (0.99; Sec. 3.4), whereas evidence depletion reaches equal or lower collapse without them.

Table 2: Alternative mechanisms: peak overlap (5 seeds, same harness; evidence depletion is linear).

Method	Sinus.	Gauss.	Multi-sc.
Vanilla SA	0.32 ± 0.09	0.26 ± 0.07	0.22 ± 0.02
Sparsemax	0.22 ± 0.14	0.19 ± 0.03	0.21 ± 0.04
Entmax-1.5	0.36 ± 0.11	0.26 ± 0.08	0.18 ± 0.05
SA-SH	0.07 ± 0.04	0.13 ± 0.03	0.12 ± 0.03
MESH	$\mathbf{0.06 \pm 0.02}$	0.18 ± 0.02	0.12 ± 0.03
Coverage (mech.)	0.78 ± 0.06	0.57 ± 0.13	0.70 ± 0.12
Coverage (loss)	0.23 ± 0.06	0.30 ± 0.04	0.24 ± 0.04
Stick-breaking	0.47 ± 0.20	0.76 ± 0.32	0.43 ± 0.19
Evidence depletion	0.07 ± 0.07	$\mathbf{0.03 \pm 0.03}$	$\mathbf{0.01 \pm 0.01}$

Sensitivity to γ and τ. Collapse is low in a broad region around τ=0.3 for all γ; τ=1.0 washes out the bias, and τ=0.1 is high-variance unless γ is large (Appendix E).

Alternative anti-collapse mechanisms. We compare against seven mechanisms that reshape attention without tracking explained evidence, implemented as slot-attention modules in the same harness (Appendix H): sparsemax [Martins and Astudillo, 2016], entmax-1.5 [Peters et al., 2019], MESH [Zhang et al., 2023] with its entropy-minimizing transport cost, SA-SH (Sinkhorn slot attention with symmetry-breaking noise, the baseline MESH builds on), a coverage mechanism and a coverage loss [Tu et al., 2016], and stick-breaking scopes after Genesis-V2 [Engelcke et al., 2021]. Evidence depletion is clearly lowest on Gaussians and multi-scale; on sinusoids, MESH, SA-SH, and evidence depletion are within seed variability (Table 2). Cosine overlap agrees: evidence depletion reaches 0.27/0.14 on Gaussians/multi-scale, every alternative ≥ 0.53. The sparse and transport-based mechanisms change *how* the same fixed input is divided among slots – sparse maps sharpen the division, doubly-stochastic ones balance mass across slots, which resolves ties between partitionable objects – while every slot still reads the same undepleted tokens, so dominant-heavy tokens can attract all of them; coverage, which accumulates attention additively, raised collapse as a mechanism (Table 2).

3.2 Real-world audio mixtures (FUSS)

To test whether evidence depletion generalizes beyond synthetic signals, we evaluate on the Free Universal Sound Separation (FUSS) benchmark [Wisdom et al., 2021]: real-world audio mixtures of 1–4 everyday sounds (speech, music, environmental noise) summed additively at 16 kHz. We crop to 4 seconds (64,000 samples), tokenize with 32 ChunkFFT tokens (d=128, 5 slots), and use spectral parameters (dominant frequency, RMS amplitude, spectral centroid) as ground truth.

Evidence depletion (linear) reduces peak-overlap collapse from 0.30 ± 0.05 (vanilla) to 0.04 ± 0.01 – a $\approx 8\times$ reduction on real audio, reaching the same range as the synthetic benchmarks (Table 1), while cosine overlap shows a more modest $\approx 1.6\times$ reduction ($0.79 \rightarrow 0.50$; Appendix F). Sequential SA collapses completely (0.94 ± 0.02) despite identical architecture, hyperparameters, and training budget. Collapse persists on these spectrally complex, heterogeneous real mixtures and evidence depletion reduces it, so FUSS shows collapse transfer, not separation: an attention-to-mask readout gives SI-SDR improvement -2.1 to $+1.1$ dB for every variant, whereas an oracle mask from the same readout family gains $+7.7$ dB, because attention trained for parameter inference is not separation-optimal (Appendix J). Quadratic depletion (0.26 ± 0.04) improves only marginally over vanilla on this benchmark, consistent with its softer update. Full per-seed results are in Appendix J.

3.3 Validation at scale: LISA inference

To test whether evidence depletion scales beyond simple benchmarks to a real scientific inference pipeline, we apply it to parameter estimation of overlapping galactic binary (GB) gravitational-wave sources for the Laser Interferometer Space Antenna [LISA Consortium, 2017]. The setting is well suited to test the mechanism: GB signals contribute additively to every observation, exhibit large dynamic range across sources ($\sim$2 orders of magnitude in strain amplitude), and overlap in frequency. Each source is described by 8 parameters (frequency f, frequency derivative $\dot{f}$, amplitude A, inclination ι, right ascension α, declination δ, polarization ψ, initial phase φ_0). The signal model uses three noise-orthogonal detector channels [Tinto and Dhurandhar, 2021] sampled at N=524,288 points over 1 year, colored noise per SciRDv1 [LISA Science Study Team, 2018], $K_{\text{resolve}} \sim \mathcal{U}\{1, \ldots, 4\}$ resolvable sources, and a Poisson(λ=3) confusion foreground of sub-threshold sources. Architecture and training details are in Appendix A; key attention hyperparameters are γ=3 (γ/τ=10 in Eq. (5) terms, App. A), τ=0.3, I=3 refinement iterations.

3.4 LISA ablation

Unlike the minimal synthetic harness (Sec. 3.1), this experiment tests the mechanism in a complex inference pipeline in which only attention varies. We train three models on one fixed dataset (same simulator version), architecture, loss functions, budget, and evaluation, varying only the slot attention mechanism: **Vanilla SA** (simultaneous competition, softmax over slots, no temperature), **Sequential SA** (slots processed one at a time, softmax over tokens with τ=0.3, no evidence masking or bias), and **Evidence Depletion** (sequential + evidence masking + log-evidence bias).

Table 3 quantifies the comparison across 5 seeds using the cosine overlap of Definition 1. Vanilla SA collapses: overlap 0.99 ± 0.00, flow NLL diverges to $+7.4$, and CRPS remains at or above the

Table 3: LISA ablation (10^5 samples, 30 epochs, mean $\pm$ std over 5 seeds, quadratic depletion). Overlap = cosine overlap of active slots (Definition 1). Prior-predictive CRPS over resolvable (SNR ≥ 7) sources: $f \approx 0.33$, $A \approx 0.19$, $\iota \approx 0.13$, $\alpha \approx 0.17$ (App. B). Seed-42 CRPS read at epoch 31 (App. B.1).

Method	Overlap↓	NLL↓	CRPS f↓	CRPS A↓	CRPS ι↓	CRPS α↓
Vanilla SA	$.99_{\pm.00}$	$+7.4_{\pm0.3}$	$.33_{\pm.01}$	$.38_{\pm.09}$	$.35_{\pm.08}$	$.33_{\pm.02}$
Sequential	$.71_{\pm.07}$	$-4.3_{\pm0.1}$	$.14_{\pm.00}$	$.14_{\pm.00}$	$.10_{\pm.00}$	$.12_{\pm.00}$
Evid. Depl.	$\mathbf{.05}_{\pm.01}$	$\mathbf{-6.0}_{\pm0.3}$	$\mathbf{.04}_{\pm.01}$	$\mathbf{.09}_{\pm.01}$	$\mathbf{.08}_{\pm.00}$	$\mathbf{.09}_{\pm.00}$

prior, indicating failure to learn useful posteriors. Sequential processing partially reduces overlap (0.71 ± 0.07) but slots do not fully specialize.

Evidence depletion prevents collapse in this setting. Overlap drops to 0.05 ± 0.01, and the model achieves the best CRPS on all reported parameters and the best flow NLL (-6.0); full per-parameter results at convergence are in Appendix B. The effect is consistent across all seeds, and held-out peak overlap gives the same ordering (Appendix F). Across the 15 runs, overlap rank-correlates with flow NLL (Spearman $r_s = 0.85$) and with the four CRPS columns ($r_s = 0.85$–0.89): in this controlled ablation, lower attention overlap is strongly associated with better posteriors.

Further controls (Appendix B.1). Linear depletion also prevents collapse on LISA (linear vs. quadratic: overlap 0.030 vs. 0.054, NLL -5.49 vs. -5.94; 3 seeds). SA-SH collapses like vanilla (0.993 ± 0.006): of the four attention mechanisms run at LISA scale (vanilla, sequential, SA-SH, evidence), only evidence depletion avoids collapse (full MESH was run on synthetic tasks only). All arms share the full loss. Evidence depletion needs none of the three explicit anti-collapse regularizers ($\mathcal{L}_{\text{div}}$, $\mathcal{L}_{\text{ortho}}$, $\mathcal{L}_{\text{ent}}$; overlap 0.056/0.036 with them disabled, 2 seeds) but uses the Hungarian-matched frequency loss on attention centroids, which vanilla also receives; with that loss removed as well, it collapses too (0.87/0.97; App. B.1). The gap persists through 100 epochs (evidence 0.018 vs. vanilla 0.99, one seed per arm). The full system trained to convergence on 10^6 samples produces approximately calibrated posteriors (coverage within 5 pp of nominal for constrained parameters; Table 6, App. B). A qualitative visualization (Appendix K, Figure 4) shows the distinct attention patterns of evidence depletion on simulated LISA inputs (overlap 0.07 vs. 0.98 for vanilla).

4 Related work

Object-centric learning. Object-centric methods [Greff et al., 2019, Burgess et al., 2019] and slot attention [Locatello et al., 2020] with its extensions [Singh et al., 2022b, Elsayed et al., 2022] decompose visual scenes into object representations. SLATE [Singh et al., 2022a] and Invariant Slot Attention [Biza et al., 2023] target representation fidelity and symmetry invariance respectively, and are orthogonal to slot collapse under additive superposition. Krimmel et al. [2024] show that softmax normalization causes collapse under cardinality shift (more objects than seen during training); our finding under amplitude dominance is a distinct trigger with the same symptom, suggesting that slot attention's normalization is fragile under multiple conditions the original formulation did not anticipate. MESH [Zhang et al., 2023] recasts slot attention as optimal transport with an entropy-minimizing cost to break ties between objects. The symptom is shared but the cause differs: MESH resolves *assignment* ambiguity, whereas under additive superposition no token partition matches the source decomposition. Genesis-V2 [Engelcke et al., 2021] uses a stochastic stick-breaking process; we include a stick-breaking-scope variant as a baseline (Table 2). Slot collapse under superposition can be viewed as a disentanglement failure due to insufficient inductive bias [Locatello et al., 2019]; evidence depletion provides the additional bias needed when observations are additive mixtures rather than spatial compositions.

Iterative subtraction and matching pursuit. CLEAN [Högbom, 1974] and matching pursuit [Mallat and Zhang, 1993] decompose signals by greedily subtracting the best-matching component. Unlike evidence depletion, they modify the data and so propagate estimation errors (Sec. 2.3).

Set prediction and iterative inference. Hungarian matching for end-to-end set prediction predates DETR [Rezatofighi et al., 2018, Zhang et al., 2019]; DETR [Carion et al., 2020] popularized it for detection. We adopt it for variable-K source assignment with flow posteriors; iterative refinement

of amortized posteriors [Marino et al., 2018] is complementary to our sequential slot processing, in which each slot captures a different source.

Attention memory: sinks, inhibition of return, and coverage. In LLMs, *attention sinks* [Xiao et al., 2024] describe concentration on uninformative tokens; slot collapse is conceptually related, where attention concentrates on dominant components across *all* slots. Evidence depletion has a structural analogue in *inhibition of return* [Klein, 2000, Itti and Koch, 2001], where previously attended locations are suppressed on a saliency map; unlike IOR, evidence depletion is trained end-to-end. Coverage in neural machine translation [Tu et al., 2016] also keeps state over prior attention, but accumulates it *additively* ($\sum_t \alpha_t$) over decoder steps assuming near one-to-one alignment; evidence depletion *multiplicatively* decays per-token explanatory capacity across slots in a many-to-many mixture, entering both the logit bias and the scaled keys/values.

Source separation and amortized inference. ICA [Hyvärinen and Oja, 2000] decomposes observations into independent sources under a linear mixing model; SLOTFLOW handles variable source count and produces approximately calibrated posteriors. Waveform separators such as Hybrid Transformer Demucs [Rouard et al., 2023] and speaker diarization address related set problems via reconstruction or clustering rather than amortized variable-K posteriors; our model reconstructs no waveforms, so we do not compare on separation quality. DINGO [Dax et al., 2021] and related work [Green and Gair, 2021] perform neural posterior estimation for single compact binary coalescence events; SLOTFLOW extends this to multiple sources with variable K. Existing LISA source-separation pipelines such as `GBMCMC` [Cornish and Crowder, 2005, Littenberg and Cornish, 2023] and global-fit approaches [Babak et al., 2008] target the full catalog with non-amortized reversible-jump MCMC under different modeling assumptions; SLOTFLOW is complementary, providing amortized variable-K posteriors rather than a full-catalog fit.

5 Discussion and conclusion

Standard attention is memoryless with respect to explained evidence. Under additive superposition, this induces slot collapse: a shared input draws multiple slots toward the same dominant component, and any reweighting of uniformly dominated tokens yields aligned readouts (Prop. 1). Evidence depletion addresses the allocation side by introducing a minimal form of state – tracking residual explanatory capacity – which steers later slots toward unexplained tokens and enforces progressive specialization. Across synthetic signals, real audio mixtures, and LISA inference, linear evidence depletion lowers mean collapse to ≤ 0.07 on every main benchmark (peak overlap; cosine on LISA) under identical architectures, data, and losses, where parallel and sequential attention fail. More broadly, attention-based decomposition of inputs that admit no partition into components benefits from explicit state over what has already been explained. *When is it needed?* Vanilla SA is partially collapsed on the synthetic tasks (0.22–0.32) but fully on LISA (0.99); amplitude ratio alone does not explain this (Appendix I), and which LISA property (128 tokens, dense overlap, $\sim$100$\times$ amplitude range, confusion foreground) drives full collapse remains open.

Limitations. (i) Collapse metrics are attention-level proxies; on LISA they rank-correlate with posterior quality across mechanisms (Spearman $r_s = 0.85$–0.89), and on FUSS they do not translate into separation. (ii) Identifiability and recovery guarantees remain open. (iii) The benefit requires dominance: at amplitude ratios $\leq$1.5 evidence depletion is not reliably better than vanilla SA. (iv) Cost is linear in S; at $S \geq 8$ the shared cardinality readout fails for all variants, so large-S scalability is open; synthetic training is unstable at L=64 tokens. (v) Depletion form, γ, τ are chosen empirically; adaptive selection and hierarchical decomposition are open. (vi) LISA is simulated and compute-heavy (5-seed ablation; convergence runs are single-seed); $\dot{f}$, φ_0 stay prior-limited.

Acknowledgments and Disclosure of Funding

This research was funded by the Gravitational Physics Professorship at ETH Zurich. Computations were performed at the Swiss National Supercomputing Centre (CSCS). The author thanks Michele Vallisneri for the insightful discussions and for his valuable contributions to editing the manuscript. The author declares no competing interests.

References

Stanislav Babak et al. The mock LISA data challenges: from challenge 1B to challenge 3. *Classical and Quantum Gravity*, 25:184026, 2008.

Jean-Baptiste Bayle, Maude Le Jeune, and Jonathan Menu. JAXGB (0.1.0), 2025.

Jean-Baptiste Bayle et al. lisaorbits: LISA orbit computation. https://pypi.org/project/lisaorbits/, 2024.

Ondrej Biza, Sjoerd van Steenkiste, Mehdi S. M. Sajjadi, Gamaleldin F. Elsayed, Aravindh Mahendran, and Thomas Kipf. Invariant slot attention: Object discovery with slot-centric reference frames. In *ICML*, 2023.

Christopher P. Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. MONet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

Neil J. Cornish and Jeff Crowder. LISA data analysis using Markov chain Monte Carlo methods. *Physical Review D*, 72:043005, 2005.

Neil J. Cornish and Louis J. Rubbo. LISA response function. *Physical Review D*, 67:022001, 2003.

Maximilian Dax, Stephen R. Green, Jonathan Gair, Jakob H. Macke, Alessandra Buonanno, and Bernhard Schölkopf. Real-time gravitational wave science with neural posterior estimation. *Physical Review Letters*, 127:241103, 2021.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. In *NeurIPS*, volume 32, 2019.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. nflows: normalizing flows in pytorch. https://github.com/bayesiains/nflows, 2020.

Gamaleldin F. Elsayed, Aravindh Mahendran, Sjoerd van Steenkiste, Klaus Greff, Michael C. Mozer, and Thomas Kipf. SAVi++: Towards end-to-end object-centric learning from real-world videos. In *NeurIPS*, 2022.

Martin Engelcke, Oiwi Parker Jones, and Ingmar Posner. GENESIS-V2: Inferring unordered object representations without iterative refinement. In *NeurIPS*, 2021.

Stephen R. Green and Jonathan Gair. Complete parameter inference for GW150914 using deep learning. *Machine Learning: Science and Technology*, 2:03LT01, 2021.

Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher P. Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In *ICML*, 2019.

John R. Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *ICASSP*, 2016.

J. A. Högbom. Aperture synthesis with a non-regular distribution of interferometer baselines. *Astronomy and Astrophysics Supplement Series*, 15:417, 1974.

Niklas Houba, Giovanni Giarda, and Lorenzo Speri. SlotFlow: Amortized trans-dimensional inference with slot-based normalizing flows. *arXiv preprint arXiv:2511.23228*, 2025.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13:411–430, 2000.

Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2:194–203, 2001.

Raymond M. Klein. Inhibition of return. *Trends in Cognitive Sciences*, 4:138–147, 2000.

Markus Krimmel, Jan Achterhold, and Jörg Stückler. Attention normalization impacts cardinality generalization in slot attention. *Transactions on Machine Learning Research*, 2024.

Harold W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:318–327, 2020.

LISA Consortium. Laser Interferometer Space Antenna. *arXiv preprint arXiv:1702.00786*, 2017.

LISA Science Study Team. LISA science requirements document. Technical Report ESA-L3-EST-SCI-RS-001, ESA, 2018.

Tyson B. Littenberg and Neil J. Cornish. Prototype global analysis of LISA data with multiple source types. *Physical Review D*, 107:063004, 2023. doi: 10.1103/PhysRevD.107.063004.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *ICML*, 2019.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *NeurIPS*, volume 33, pages 11525–11538, 2020.

Stéphane G. Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41:3397–3415, 1993.

Joseph Marino, Yisong Yue, and Stephan Mandt. Iterative amortized inference. In *ICML*, 2018.

André F. T. Martins and Ramón Fernandez Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *ICML*, 2016.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22:1–64, 2021.

Adam Paszke, Sam Gross, Francisco Massa, et al. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*, volume 32, 2019.

Ben Peters, Vlad Niculae, and André F. T. Martins. Sparse sequence-to-sequence models. In *ACL*, 2019.

S. Hamid Rezatofighi, Roman Kaskman, Farbod T. Motlagh, Qinfeng Shi, Daniel Cremers, Laura Leal-Taixé, and Ian Reid. Deep Perm-Set Net: Learn to predict sets with unknown permutation and cardinality using deep neural networks. *arXiv preprint arXiv:1805.00613*, 2018.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *ICML*, pages 1530–1538, 2015.

Simon Rouard, Francisco Massa, and Alexandre Défossez. Hybrid transformers for music source separation. In *ICASSP*, 2023.

Gautam Singh, Fei Deng, and Sungjin Ahn. Illiterate DALL-E learns to compose. In *ICLR*, 2022a.

Gautam Singh, Yi-Fu Wu, and Sungjin Ahn. Simple unsupervised object-centric learning for complex and naturalistic videos. In *NeurIPS*, 2022b.

Massimo Tinto and Sanjeev V. Dhurandhar. Time-delay interferometry. *Living Reviews in Relativity*, 24(1):1, 2021. doi: 10.1007/s41114-020-00029-6.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. Modeling coverage for neural machine translation. In *ACL*, 2016.

Scott Wisdom, Hakan Erdogan, Daniel P. W. Ellis, Romain Serizel, Nicolas Turpault, Eduardo Fonseca, Justin Salamon, Prem Seetharaman, and John R. Hershey. What's all the FUSS about free universal sound separation data? In *ICASSP*, 2021.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *ICLR*, 2024.

Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. Deep set prediction networks. In *NeurIPS*, 2019.

Yan Zhang, David W. Zhang, Simon Lacoste-Julien, Gertjan J. Burghouts, and Cees G. M. Snoek. Unlocking slot attention by changing optimal transport costs. In *ICML*, 2023.

A LISA implementation details

This appendix consolidates the architecture, flow context, and training details for the LISA experiments (Sec. 3.3, 3.4). All LISA data is simulated: the LISA mission is scheduled for launch in the mid-2030s and no observational data yet exists. Waveforms are generated with JaxGB [Bayle et al., 2025] using LISA orbital configurations from lisaorbits [Bayle et al., 2024], with instrument noise following the Science Requirements Document v1 [LISA Science Study Team, 2018].

Ablation validity. The LISA architecture is complex (21.7M parameters, dual-path encoder, 10-stream flow context, 10 loss terms), but the ablation in Sec. 3.4 is designed to isolate the attention mechanism from this complexity. All three variants (vanilla SA, sequential SA, evidence depletion) share *exactly* the same tokenizer, source encoder, normalizing flow, loss function (including all auxiliary terms and weights), training data, hyperparameters, and random seed. The only difference is the slot attention module: vanilla uses simultaneous softmax-over-slots competition (no temperature), sequential processes slots one at a time with softmax over tokens (τ=0.3, as for evidence depletion), and evidence depletion adds the multiplicative update and log-evidence bias. Any performance difference is therefore attributable to the attention mechanism under matched architecture, data, losses, and initialization. The qualitative visualization in Appendix K illustrates this: the attention patterns themselves differ markedly across variants (overlap 0.98 vs. 0.07). The architectural complexity reflects the difficulty of the underlying problem: LISA will observe millions of overlapping galactic binary signals in a single data stream (our 1-year simulation alone has 524,288 samples per channel), each source described by 8 correlated physical parameters with large dynamic range. Extracting source-specific information from this mixture requires physics-informed feature extraction, temporal modulation encoding, and structured posterior estimation – all orthogonal to the attention mechanism under study.

A.1 Architecture

Time-frequency tokenization. The signal ($3 \times 524{,}288$ detector-channel samples: A, E, T) is split into 4 overlapping segments, Hann-windowed, and FFT'd. The CNN operates on 4 real-valued channels $[\mathrm{Re}(\tilde{A}), \mathrm{Im}(\tilde{A}), \mathrm{Re}(\tilde{E}), \mathrm{Im}(\tilde{E})]$ (the T channel is noise-dominated and omitted); Hz-based patch pooling produces 128 tokens ($\in \mathbb{R}^{128}$) with learned positional encodings (Hz + seconds).

Sequential slot attention. $S = K_{\max} + 1 = 5$ slots with evidence depletion ($\gamma = 3$, $\tau = 0.3$, $I = 3$ iterations, $\epsilon = 10^{-4}$). In this implementation the bias is added before the temperature scaling, $\mathrm{softmax}_\ell\big((\mathbf{q}_s \cdot \mathbf{k}_\ell/\sqrt{d} + \gamma \log e_\ell)/\tau\big)$, so these settings correspond to a bias strength $\gamma/\tau = 10$ in Eq. (5); the synthetic and FUSS models use Eq. (5) as written.

Dual-path source encoder. Given each slot's frequency estimate $\hat{f}_s$ (produced by a differentiable regression head), a 64-bin spectral window is extracted using a Gaussian-weighted soft crop centered on $\hat{f}_s$ from 16 re-segmented time windows. **Path 1** (physics-informed): 21 hand-crafted features per segment (power ratios, Stokes parameters, cross-phase, detrended phase) $\rightarrow$ temporal CNN $\rightarrow$ trunk MLP ($2048 \rightarrow 256$). **Path 2** (learned): A 2D CNN on a 10-channel encoding of the raw spectral window (log-magnitudes, phases, and normalized cross-spectra of A and E) with *frequency-only pooling* – preserving the 16-step temporal orbital modulation that encodes sky position and polarization (Table 4).

Normalizing flow. 12 RQ-spline coupling layers [Durkan et al., 2019] (nflows [Durkan et al., 2020], PyTorch [Paszke et al., 2019]), 64 bins, hidden dim 640, physics-informed split: odd layers transform $(A, \iota, \psi, \varphi_0)$, even layers transform $(f, \dot{f}, \alpha, \delta)$. Flow context: 778 dimensions from 10 feature streams (Table 5). Total: 21.7M parameters.

Auxiliary 1D polarization flow. The polarization angle ψ exhibits a 4-fold symmetry under $\psi \rightarrow \psi + \pi/2$ in LISA's antenna response, encoded by harmonics of period $\pi/2$. A standard joint flow over all 8 parameters tends to average over the symmetric modes, yielding overly broad posteriors on ψ. We therefore train a small auxiliary 1D normalizing flow on ψ alone, conditioned on the same slot context, with the explicit harmonic encoding $(\sin(4\psi), \cos(4\psi))$ that respects this symmetry. This auxiliary head is used only as a training-time regularizer through the $\mathcal{L}_{\text{pol}}$ loss term;

the main posterior at evaluation comes from the primary 8-dimensional flow. In our development runs, removing this term caused ψ recovery to degrade to AngCorr ≈ 0.

Table 4: Encoder ablation: temporal-preserving (frequency-only pooling) vs. baseline (global pooling), matched epochs.

Metric	Baseline	Freq-only pool	Δ
RA CRPS	0.125	0.054	-57%
Dec CRPS	0.178	0.085	-52%
Inc CRPS	0.069	0.047	-32%
Freq MAE (μHz)	0.51	0.26	-49%

A.2 Flow context assembly

The flow context vector concatenates 10 feature streams totaling 778 dimensions (Table 5). SBI dropout is applied to the slot attention projection and frequency resolution streams to regularize against overconfidence.

Table 5: Flow context streams (778-dim total).

Stream	Dim	Notes
Slot attention proj.	128	SBI dropout 30%
Source features proj.	192	SourceEncoder trunk
Modulation stats	64	Temporal statistics MLP
Anchor proj.	32	$\hat{f}_{\text{norm}}$, $\log P$
Freq. resolution	16	SBI dropout 30%
Frequency emb.	32	Frequency value
Angular emb.	64	AngularHead MLP
Angular skip	10	Raw sin/cos
Pol temporal	48	Stokes U + cross-phase
Spectro features	192	SpectrogramEncoder

A.3 Training

Loss function. $\mathcal{L} = \mathcal{L}_{\text{flow}}(1) + \mathcal{L}_{\text{exist}}(3) + \mathcal{L}_{\text{div}}(0.5) + \mathcal{L}_{f,\text{attn}}(5) + \mathcal{L}_{f,\text{direct}}(50) + \mathcal{L}_{\text{ang}}(3) + \mathcal{L}_{\text{prior}}(0.1) + \mathcal{L}_{\text{ortho}}(0.3) + \mathcal{L}_{\text{ent}}(0.1) + \mathcal{L}_{\text{pol}}(1)$. Weights in parentheses. $\mathcal{L}_{\text{flow}}$: negative log-likelihood of the normalizing flow on matched (existing) slots. $\mathcal{L}_{\text{exist}}$: focal loss [Lin et al., 2020] on existence-head logits, $\text{FL}(p_t) = -\alpha_t(1-p_t)^{\gamma_f} \log p_t$ with γ_f=2, α_t=0.5, label smoothing 0.01. $\mathcal{L}_{f,\text{direct}}$: MSE between predicted and true frequency in relative coordinates. $\mathcal{L}_{f,\text{attn}}$: MSE on attention-weighted frequency estimates. $\mathcal{L}_{\text{ang}}$: weighted MSE on angular sin/cos harmonic encodings (see below). $\mathcal{L}_{\text{div}}$: mean squared cosine similarity between active slot embedding pairs, $\mathcal{L}_{\text{div}} = \frac{1}{|\mathcal{P}|}\sum_{(i,j)\in\mathcal{P}}(\hat{\mathbf{s}}_i^\top \hat{\mathbf{s}}_j)^2$, where $\hat{\mathbf{s}}_i = \mathbf{s}_i/\|\mathbf{s}_i\|$ and $\mathcal{P}$ is the set of active slot pairs. $\mathcal{L}_{\text{ortho}}$: mean absolute cosine similarity between active slot attention vectors, $\frac{1}{|\mathcal{P}|}\sum_{(i,j)\in\mathcal{P}} |\hat{\boldsymbol{\alpha}}_i^\top \hat{\boldsymbol{\alpha}}_j|$, optionally weighted by per-token signal power. $\mathcal{L}_{\text{ent}}$: penalises high-entropy attention for active slots, $\mathcal{L}_{\text{ent}} = \frac{1}{|\mathcal{A}|}\sum_{s\in\mathcal{A}} H(\boldsymbol{\alpha}_s)/H_{\max}$, where $H_{\max} = \log L$. $\mathcal{L}_{\text{prior}}$: flow NLL evaluated on standard-normal samples for unmatched (non-existing) slots, encouraging the flow to map unused contexts to a simple prior. $\mathcal{L}_{\text{pol}}$: NLL of auxiliary 1D normalizing flow on ψ alone, conditioned on slot context. The auxiliary terms ($\mathcal{L}_{\text{div}}$, $\mathcal{L}_{\text{ortho}}$, $\mathcal{L}_{\text{ent}}$) discourage redundant slot allocation, as does $\mathcal{L}_{f,\text{attn}}$, which is Hungarian-matched per slot; they are held fixed across all ablation variants in Sec. 3.4 so that only the attention mechanism varies.

Angular encoding. Harmonic multipliers: RA (m=1), Inc (m=2), Dec (m=2), Pol (m=4), Phase (m=1). The $4\times$ harmonic for polarization matches the 4-fold symmetry of LISA's antenna pattern under $\psi \to \psi + \pi/2$; in preliminary experiments, $\sin(\psi)$ and $\sin(2\psi)$ encodings consistently failed to recover ψ.

Hyperparameters and resources. 10^5 samples for the ablation comparison (30 epochs); 10^6 samples for the full model (150 epochs). Adam optimizer, batch 128, lr 10^{-4}, ReduceLROnPlateau (patience 10, factor 0.5), gradient clip 5.0, 4–8$\times$NVIDIA GH200 (120 GB), 10^5 validation samples, 21.7M parameters.

B LISA results at convergence

CRPS is computed in the flow's normalized coordinates: f (offset from the window centre over the half-width), $\log_{10} \dot{f} + 16$, $\log_{10} A$ (affinely) and $\delta/(\pi/2)$ lie in $[-1, 1]$; ι/π, $\alpha/2\pi$, ψ/π, $\varphi_0/2\pi$ in $[0, 1]$. Under the sampling prior the prior-predictive CRPS is ≈ 0.333 for f, $\dot{f}$, A, 0.25 for δ, 0.125 for ι, and 0.167 for α, ψ, φ_0; because only sources with SNR ≥ 7 are resolvable targets, the evaluated population favours high A and face-on/off ι (near 0 or π), for which it is ≈ 0.19 (A) and ≈ 0.13 (ι). Table 6 reports parameter recovery for the evidence depletion model trained to convergence on 10^6 samples (150 epochs). Frequency achieves CRPS $= 0.015$ (95% below the prior); amplitude, inclination, and right ascension reach CRPS ≤ 0.05, declination 0.073. The initial phase φ_0 returns a prior-width posterior (CRPS $= 0.168 \approx 0.167$), as expected for an unconstrained parameter. The frequency derivative $\dot{f}$ likewise returns a prior-width posterior (CRPS $= 0.335 \approx 0.333$, nominal coverage): encoder compression discards its information (Appendix C).

Polarization (ψ) achieves partial recovery (AngCorr 0.72), limited by a physical degeneracy with sky position through LISA's orbital modulation [Cornish and Rubbo, 2003]. Here, AngCorr denotes the angular correlation: the Pearson correlation between the predicted and true $(\sin(m\theta), \cos(m\theta))$ harmonic encodings, where m is the parameter-specific harmonic multiplier (App. A.3).

Existence prediction and completeness. With $S = 5 \geq K$ slots, the Hungarian matching assigns every true source to a slot by construction, so recovery quality is measured by CRPS and coverage (Table 6). *Existence-head accuracy* of 79% measures how often the head correctly predicts which slots correspond to real sources. Errors here skew toward false positives: the head occasionally flags an unmatched "empty" slot as active. This is consistent with the focal-loss training and the practice of running with $S > K_{\max}$. For downstream use, this bias is the desirable direction – false positives can be filtered by posterior width or amplitude threshold; false negatives would hide sources. Frequency MAE degrades gracefully from 0.13 μHz (K=1) to 0.27 μHz (K=4).

Table 6: LISA parameter recovery at convergence (10^6 samples, 150 epochs, realistic noise + confusion foreground). Prior-predictive CRPS per parameter in the text (e.g., f, $\dot{f}$ $\approx$0.333; φ_0 0.167).

Parameter	CRPS	AngCorr	Cov50%	Cov90%	Status
Frequency f	0.015	—	0.45	0.87	Learned
Freq. deriv. $\dot{f}$	0.335	—	0.49	0.89	Prior
Amplitude A	0.050	—	0.49	0.89	Learned
Inclination ι	0.042	0.80	0.48	0.89	Learned
Right asc. α	0.043	0.83	0.47	0.88	Learned
Declination δ	0.073	0.81	0.47	0.88	Learned
Polarization ψ	0.147	0.72	0.51	0.89	Partial
Phase φ_0	0.168	—	0.51	0.91	Prior

Freq. MAE: 0.13 μHz (K=1) to 0.27 μHz (K=4). Exist. head accuracy 0.79.

B.1 Additional LISA ablations

Auxiliary losses. All ablation models share the same loss function (App. A.3), including auxiliary terms ($\mathcal{L}_{\text{div}}$, $\mathcal{L}_{\text{ortho}}$, $\mathcal{L}_{\text{ent}}$, and the Hungarian-matched $\mathcal{L}_{f,\text{attn}}$) that also discourage redundant slot allocation, so the comparison in Table 3 isolates the attention mechanism *given* these losses. Evidence depletion's improvements on f, A, ι, and α are attributable to the attention mechanism under the shared losses (these parameters use only the primary 8D flow); the ψ improvement may additionally benefit from the auxiliary 1D polarization flow, which is shared across all variants.

Further controls. Table 3 uses 5-seed matched reruns, which replace the 3-seed table of the submitted version (overlap .99/.67/.08); for seed 42 (a 100-epoch run) CRPS is read at epoch 31, as it is logged every 10 epochs, and overlap and NLL at epoch 30. All runs below use the matched protocol of Table 3 (10^5 samples, 30 epochs, same dataset, architecture, losses, and evaluation), with the seeds stated.

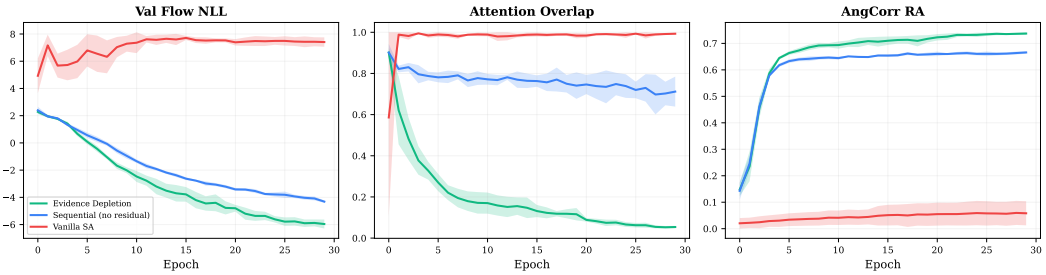

Figure 2: LISA ablation: training dynamics (10^5 samples, 30 epochs; mean $\pm$ std over the 5 seeds of Table 3). Vanilla SA (red) diverges with flow NLL > 0; sequential (blue) learns but maintains high attention overlap; evidence depletion (green) achieves low overlap and best metrics across all three panels.

- *Depletion form* (seeds 42/123/456). Linear depletion: overlap 0.030 ± 0.006, NLL -5.49; quadratic on the same seeds: 0.054 ± 0.004, -5.94. Both prevent collapse; linear gives lower overlap, quadratic slightly better NLL.
- *SA-SH* (Sinkhorn slot attention with symmetry-breaking noise; seeds 42/123/456): overlap 0.993 ± 0.006, NLL $+2.96 \pm 0.34$ (vanilla on the same seeds: NLL $+7.32$). SA-SH and vanilla both collapse in attention but differ in NLL. A one-seed sweep over Sinkhorn iterations $\{5, 10, 20\}$ and noise scales $\{0.01, 0.05, 0.10\}$, evaluated at epoch 15 (default SA-SH 0.998, evidence depletion $\approx$0.15 at the same epoch), stays substantially collapsed in every setting (0.79–1.00, lowest at 20 Sinkhorn iterations). Full MESH adds an inner entropy-minimizing optimization of the transport cost, which we evaluated in the synthetic harness (Table 2) but did not integrate into the LISA pipeline; we make no claim about full MESH at LISA scale.
- *Anti-collapse regularizers off* (2 seeds). With $\mathcal{L}_{\text{div}}$, $\mathcal{L}_{\text{ortho}}$, and $\mathcal{L}_{\text{ent}}$ disabled, keeping only the task-supervision losses that the collapsing vanilla baseline also uses, evidence depletion reaches overlap 0.056 and 0.036. When the attention-supervised frequency loss $\mathcal{L}_{f,\text{attn}}$ is also disabled (same 2 seeds), evidence depletion collapses as well (overlap 0.87 and 0.97; one run ended after 24 epochs), so on LISA the mechanism relies on this frequency loss, which supervises each matched slot's attention centroid and which the collapsing vanilla baseline also receives.
- *Convergence* (1 seed per arm, 100 epochs). Overlap/NLL: evidence $0.018/-8.61$, sequential $0.713/-6.92$, vanilla $0.991/+7.80$. In this single-seed extension, insufficient training time does not explain the gap.
- *Overlap and posterior quality.* Across the 15 runs of Table 3, Spearman correlation of overlap with flow NLL is 0.85, and with CRPS on f, A, ι, α it is 0.87, 0.88, 0.89, 0.85. Within a single mechanism the seed-to-seed spread is too small for a meaningful correlation.

C Encoder compression and the limits of amortization

A general property of any amortized inference system that compresses observations through an encoder is that information distributed thinly across many samples can be lost. This appendix uses the gravitational-wave frequency derivative $\dot{f}$ as a concrete instance to illustrate the point; the observation generalizes beyond the LISA setting to any encoder-based simulation-based inference (SBI) system.

For sources with $\dot{f} \sim 10^{-16}$ Hz/s, the phase accumulation over a 1-year observation is $\Delta\Phi = \pi \dot{f} T^2 \approx 0.3$ rad, which is detectable in coherent template matching but is distributed essentially uniformly across $N = 524{,}288$ samples. Our encoder reduces this to 16 detrended phase measurements per slot before passing to the source encoder; a subsequent 8:1 compression discards the subtle quadratic phase curvature amid stronger angular gradients. The model returns a prior-width posterior on $\dot{f}$ (CRPS $= 0.335$, versus 0.333 for a uniform prior on its normalized range $[-1, 1]$, with nominal coverage 0.49/0.89), indicating that encoder compression discards the $\dot{f}$ information: the posterior is uninformative rather than miscalibrated.

The observation suggests a design principle for amortized multi-source inference: *parameters whose information is distributed across many samples may require either matched-filter features or uncompressed representations*. Resolving this is orthogonal to the slot collapse problem and applies

to any encoder-based SBI architecture, especially when scaling to long signals or settings with weak, distributed evidence.

D Allocation under additive superposition: analysis and proof

Gradient intuition. Consider the attention logit for slot s at token ℓ containing two additive sources: $\mathbf{h}_\ell = A_1\mathbf{h}^{(1)} + A_2\mathbf{h}^{(2)}$ with $A_1 \gg A_2$. Through the key projection $\mathbf{k}_\ell = W_k\mathbf{h}_\ell$, the logit becomes $\mathbf{q}_s \cdot \mathbf{k}_\ell = A_1(\mathbf{q}_s \cdot W_k\mathbf{h}^{(1)}) + A_2(\mathbf{q}_s \cdot W_k\mathbf{h}^{(2)})$. For softmax over slots (Eq. 3), the gradient of the attention with respect to the slot query is (up to the scaling constant $1/\sqrt{d}$):

$$\frac{\partial \alpha_{s\ell}}{\partial \mathbf{q}_s} \propto \alpha_{s\ell}(1-\alpha_{s\ell}) \cdot \mathbf{k}_\ell = \alpha_{s\ell}(1-\alpha_{s\ell}) \cdot (A_1 W_k\mathbf{h}^{(1)} + A_2 W_k\mathbf{h}^{(2)}) \,.$$

When $A_1 \gg A_2$, the gradient *direction* is dominated by the A_1 term, and this direction is the same for all slots; for the loss gradient $\sum_\ell g_\ell \, \partial\alpha_{s\ell}/\partial\mathbf{q}_s$ this holds unless the per-token loss sensitivities g_ℓ cancel. (This is a heuristic for softmax over slots; for softmax over tokens the Jacobian is $\alpha_{s\ell}(\mathbf{k}_\ell - \sum_m \alpha_{sm}\mathbf{k}_m)/(\tau\sqrt{d})$, in which a component common to all keys cancels.) Distinct samples from the shared slot prior yield distinct parameter trajectories – slots never become literally equal – so the relevant question is not whether slot states coincide but whether their *allocations* do, which is the level at which Definition 1 measures collapse. Proposition 1 answers it for the readouts.

Proof of Proposition 1. Write $\mathbf{u} = \sum_\ell w_\ell \mathbf{v}_\ell = A\,\mathbf{h} + \mathbf{R}$ with $A = \sum_\ell w_\ell a_\ell > 0$ and $\mathbf{R} = \sum_\ell w_\ell \mathbf{r}_\ell$. By the triangle inequality and $\|\mathbf{r}_\ell\| \le \kappa a_\ell$, $\|\mathbf{R}\| \le \sum_\ell w_\ell \|\mathbf{r}_\ell\| \le \kappa A$. Hence $\mathbf{u}$ lies in the closed ball of radius κA around $A\mathbf{h}$; since $\kappa < 1$ this ball excludes the origin, and every point in it makes an angle at most $\arcsin\kappa$ with $\mathbf{h}$. The same holds for $\mathbf{u}'$, so the angle between $\mathbf{u}$ and $\mathbf{u}'$ is at most $2\arcsin\kappa < \pi$, and since cosine is decreasing on $[0,\pi]$, $\cos(\mathbf{u},\mathbf{u}') \ge \cos(2\arcsin\kappa) = 1 - 2\kappa^2$. The bound is attained: with $\mathbf{n} \perp \mathbf{h}$ a unit vector and $\theta = \arcsin\kappa$, the two tokens $\mathbf{v}_\pm = \mathbf{h} + \kappa(-\sin\theta\,\mathbf{h} \pm \cos\theta\,\mathbf{n})$ have $\|\mathbf{r}_\pm\| = \kappa a_\pm$ and one-hot readouts at angle exactly 2θ. For two sources with unit $\mathbf{h}^{(1)}, \mathbf{h}^{(2)}$, take $\mathbf{h} = \mathbf{h}^{(1)}$, $a_\ell = A_1 c_\ell^{(1)}$, $\mathbf{r}_\ell = A_2 c_\ell^{(2)}\mathbf{h}^{(2)}$; then $\|\mathbf{r}_\ell\|/a_\ell \le A_2 c_{\max}/(A_1 c_{\min}) = \rho$. □

For any $\eta \ge \kappa$ with $\eta < 1/2$, $1-2\kappa^2 \ge 1-2\eta^2 \ge (1-2\eta)/(1-2\eta+2\eta^2) \ge 1-2\eta$ (the middle step is $4\eta^3(1-\eta) \ge 0$); this applies in particular to the cross-token ratio $\eta = \max_\ell \|\mathbf{r}_\ell\| / \min_\ell a_\ell \ge \kappa$ stated (as δ) during the discussion period. The proof only uses tokens with positive weight, so κ may be taken over $\mathrm{supp}(w) \cup \mathrm{supp}(w')$. In our implementation the depletion factor multiplies the layer-normalized token $\tilde{\mathbf{h}}_\ell$ and the value map is affine, $\mathbf{v}_\ell = W_v(e_\ell\tilde{\mathbf{h}}_\ell) + \mathbf{b}_v$; the proposition and the corollary below apply to the pre-projection readouts $\sum_\ell w_\ell e_\ell \tilde{\mathbf{h}}_\ell$, whose value-space images differ by a fixed linear map and, for weights on the simplex, the common offset $\mathbf{b}_v$.

Corollary 1 (Invariance under token reweighting)**.** *The bound of Proposition 1 is unchanged if each token is rescaled by any factor $e_\ell > 0$, since $\|e_\ell\mathbf{r}_\ell\|/(e_\ell a_\ell) = \|\mathbf{r}_\ell\|/a_\ell$ (equivalently, $\sum_\ell w_\ell e_\ell \mathbf{v}_\ell$ is just another nonnegative weighting). In particular, evidence depletion – which rescales tokens by $e_\ell \ge \epsilon$ (Eq. 5) and shifts the logits by $\gamma \log e_\ell$ – produces readouts obeying the same bound on uniformly dominated tokens.*

Corollary 1 sharpens how evidence depletion relates to the bound. A cross-token ratio such as η above does grow under depletion (claimed tokens lose a_ℓ while unclaimed ones keep $\|\mathbf{r}_\ell\|$), but this only loosens one sufficient condition; the per-token κ, and hence the bound, is invariant. Evidence depletion therefore does not defeat the geometric constraint. What it changes is *which* tokens later slots read: the bias steers them away from claimed tokens, and dissimilar readouts become possible only when the attended tokens admit no common dominant direction (κ not small on their support). This is why we describe the mechanism as redirecting explanatory capacity, and why we make no separation guarantee: whether such tokens exist and are found is an empirical question. This corrects our discussion-period argument, which read the growth of η under depletion as removing the premise of the bound, and which identified the measured ratio $\hat{\rho}$ with the proposition's ratio.

Empirical check on trained models. We test the readout geometry on Synthetic-A (sinusoids), 1024 test mixtures with $K \in \{2,3,4\}$, using three vanilla and three evidence-depletion (linear) checkpoints and the trained attention weights extracted exactly from each model. Readouts are

computed in a surrogate token space: the trained attention rows applied to per-token magnitude spectra $|\mathrm{rFFT}|$, in which each token is the dominant source's magnitude profile plus a perturbation no larger than the weaker sources and noise (entrywise, $\big||X_1{+}X_2| - |X_1|\big| \le |X_2|$). In the model's own value space, after the nonlinear encoder, vanilla readouts show no comparable alignment, and in the raw signed space the dominant direction rotates with phase; we therefore make no claim about those spaces. The bound is weak here as well: the dominant source's own profile changes direction across tokens, so the smallest achievable κ has median 0.56 and the bound $1 - 2\kappa^2$ has median 0.37, equal here to the geometric floor (per input, the smallest pairwise token cosine, which lower-bounds the cosine of any two nonnegative readouts when positive); κ is computed with an approximate minimum-enclosing-cap iteration. The cross-token energy ratio $\hat{\rho}$ we quoted during the discussion period (median 0.21) does not bound κ in this space. The check therefore tests the trained *allocation*, not the bound. Pair statistics are over all five slots.

- *Vanilla SA.* Even the least similar slot pair, with attention cosine 0.24–0.55, has readouts with median cosine 0.997–0.999 across seeds, far above the geometric floor: the alignment reflects the learned allocation, not the geometric constraint.
- *Evidence depletion.* Attention rows overlap little (median mean-pair cosine 0.05–0.19), and readouts are markedly less aligned (median 0.81–0.92 across seeds). Computing the same readouts with or without the depletion factors e_ℓ changes the median by less than 0.01, in line with Corollary 1: the reduction comes from *where* later slots attend, not from rescaling.

E Sensitivity to γ and τ

We sweep the bias strength $\gamma \in \{1, 3, 10\}$ and softmax temperature $\tau \in \{0.1, 0.3, 1.0\}$ on Synthetic-A (sinusoids), reporting peak overlap rate (Definition 1, mean $\pm$ std over 3 seeds).

Table 7: Sensitivity to γ and τ on Synthetic-A (sinusoids), using quadratic depletion $(1{-}\alpha^2)$. Peak overlap rate (mean $\pm$ std, seeds 0–2 of the Table 1 runs; seed 3 collapses there, so the default cell is lower than in Table 1). Default values used in the main paper: $\gamma{=}3$, $\tau{=}0.3$ (LISA: effective $\gamma{=}10$, App. A).

	$\tau = 0.1$	$\tau = 0.3$	$\tau = 1.0$
$\gamma = 1$	0.35 ± 0.45	$\mathbf{0.03 \pm 0.01}$	0.76 ± 0.17
$\gamma = 3$	0.31 ± 0.42	0.06 ± 0.05	0.69 ± 0.33
$\gamma = 10$	$\mathbf{0.02 \pm 0.01}$	0.03 ± 0.01	0.26 ± 0.25

The dominant trend is the temperature τ: moderate values ($\tau{=}0.3$) yield consistently low collapse across all γ, while high temperature ($\tau{=}1.0$) washes out the bias and causes collapse in most settings. At low temperature ($\tau{=}0.1$), the method works well with strong bias ($\gamma{=}10$, collapse 0.02) but becomes unstable at weaker bias (std > 0.4), likely because sharp attention concentrates on noise fluctuations before the bias can redirect. The default configuration ($\gamma{=}3$, $\tau{=}0.3$) lies in the stable low-collapse regime; strong bias ($\gamma{=}10$) at $\tau{=}0.3$ performs comparably (0.03). Evidence depletion is therefore stable in a broad region around $\tau{=}0.3$, but not across the whole grid: only four of the nine settings are reliably low-collapse, $\tau{=}1.0$ is at or above vanilla collapse unless $\gamma{=}10$, and $\tau{=}0.1$ is high-variance unless γ is large. The temperature should be chosen with this in mind; adaptive selection of γ and τ remains open.

F Collapse metrics

All collapse numbers in the paper are instances of Eq. (4): the maximum similarity over pairs of active slots, averaged over multi-source inputs. Let $\alpha_s \in \mathbb{R}^L$ denote the attention vector of slot s over L tokens.

Peak overlap. With $\mathrm{sim}(\alpha_s, \alpha_{s'}) = \mathbf{1}[p_s = p_{s'}]$, $p_s = \arg\max_\ell \alpha_{s\ell}$, the maximum over pairs is 1 iff two active slots share a peak token, so the average is the fraction of inputs exhibiting collapse:

$$\mathrm{PeakOverlap} = \frac{1}{|\{i : K_i \ge 2\}|} \sum_{i:\, K_i \ge 2} \mathbf{1}\Big[\exists\, s \ne s' \in \mathcal{A}_i : p_s^{(i)} = p_{s'}^{(i)}\Big] .$$

Cosine overlap. With $\mathrm{sim}(\alpha_s, \alpha_{s'}) = \alpha_s \cdot \alpha_{s'} / (\|\alpha_s\| \, \|\alpha_{s'}\|)$,

$$\text{CosOverlap}(\mathbf{x}) = \max_{\substack{s,s' \in \mathcal{A}(\mathbf{x}) \\ s \neq s'}} \frac{\alpha_s \cdot \alpha_{s'}}{\|\alpha_s\| \, \|\alpha_{s'}\|},$$

averaged over inputs with $|\mathcal{A}| \geq 2$. This captures the worst-case pair of collapsing slots and is more informative than peak overlap when L is large (128 tokens for LISA) and attention patterns can be similar without sharing a single peak.

Active set. On the synthetic and FUSS benchmarks, $\mathcal{A}_i$ is the first K_i slots by index (ground-truth K_i); because evidence depletion processes slots in index order at inference, these are the first-processed slots, while for vanilla SA they are an arbitrary subset of exchangeable slots. On LISA, $\mathcal{A} = \{s : \sigma(\text{exist-logit}(s)) > 0.5\}$, the predicted active slots. Within each domain, the active-set rule is identical for all methods.

Both metrics on all five domains. Table 8 reports both metrics for the core three-way comparison. They agree on the ordering in every domain; cosine overlap is less extreme than peak overlap on the low-dimensional tasks, and on FUSS the cosine reduction is $\approx 1.6\times$ rather than $\approx 8\times$.

Table 8: Peak / cosine overlap for the core comparison (5 seeds; evidence depletion is linear on synthetic and FUSS, quadratic on LISA). LISA: cosine on the validation set from the training logs, as in Table 3; held-out evaluation (2–3 seeds per arm) gives peak 0.80/0.47/0.03 and cosine 0.82/0.61/0.06 (one vanilla seed predicts fewer than two active slots on most held-out inputs, so its cosine covers a small, atypical subset). Means over seeds.

Method	Sinusoids	Gaussians	Multi-scale	FUSS	LISA (cos.)
Vanilla SA	0.32 / 0.90	0.26 / 0.87	0.22 / 0.90	0.30 / 0.79	0.99
Sequential SA	0.90 / 0.95	0.81 / 0.91	0.97 / 1.00	0.94 / 0.97	0.71
Evidence depletion	**0.07 / 0.39**	**0.03 / 0.27**	**0.01 / 0.14**	**0.04 / 0.50**	**0.05**

G Additional depletion forms

Table 1 in the main text reports the two best-performing depletion forms (linear and quadratic). For completeness, Table 9 includes binary and cubic variants.

Table 9: All depletion forms: slot collapse rate (mean $\pm$ std, 5 seeds, 200 epochs).

Method	Sinusoids	Gaussians	Multi-scale
Binary (hard mask at $\alpha > 0.5$)	0.30 ± 0.27	0.41 ± 0.19	0.71 ± 0.26
Linear $(1-\alpha)$	$\mathbf{0.07 \pm 0.07}$	$\mathbf{0.03 \pm 0.03}$	$\mathbf{0.01 \pm 0.01}$
Quadratic $(1-\alpha^2)$	0.16 ± 0.23	0.08 ± 0.06	0.35 ± 0.19
Cubic $(1-\alpha^3)$	0.11 ± 0.05	0.26 ± 0.32	0.52 ± 0.13

Binary depletion performs worst: the hard threshold discards the graded residual, so a token is either fully available or removed and partially explained tokens cannot be shared. Among the continuous forms, which deplete a moderately attended token less as the exponent grows, the trend (linear $>$ quadratic $>$ cubic) on Gaussians and multi-scale suggests that stronger depletion generally helps, though cubic outperforms quadratic on sinusoids (0.11 vs. 0.16). We swept four shapes to check that the result is not specific to one functional form; all reduce collapse relative to sequential SA.

H Loss-based and alternative baselines

All synthetic models are trained with Adam (learning rate 3×10^{-4}, cosine annealing over the 200 epochs, batch 128, gradient clipping at 1.0) on signals with additive Gaussian noise of standard deviation 0.1; FUSS models use the same optimizer with batch 32. We test whether stronger loss-based regularization can resolve slot collapse without modifying the attention mechanism. Six baselines combine vanilla or sequential SA with stronger diversity losses (w_{div}=5.0, vs. the default 0.5), attention orthogonality losses (w_{ortho}=3.0), or both. All use identical architecture, data, and training budget as the main ablation (200 epochs, 5 seeds).

Table 10: Loss-based baselines: slot collapse rate (mean ± std, 5 seeds, 200 epochs). Reference: evidence depletion (linear) achieves 0.07/0.03/0.01 on these tasks with the default weak diversity term (w_{div}=0.5, as for all methods in Table 1) and no other auxiliary loss.

SA variant	Loss config	Sinusoids	Gaussians	Multi-scale
Vanilla	w_{div}=5	0.14 ± 0.04	0.19 ± 0.04	0.10 ± 0.02
Vanilla	w_{ortho}=3	0.08 ± 0.02	0.10 ± 0.03	0.07 ± 0.02
Vanilla	Both	0.06 ± 0.03	0.11 ± 0.06	0.06 ± 0.01
Sequential	w_{div}=5	0.74 ± 0.16	0.68 ± 0.05	0.85 ± 0.07
Sequential	w_{ortho}=3	0.45 ± 0.06	0.50 ± 0.08	0.49 ± 0.04
Sequential	Both	0.30 ± 0.04	0.36 ± 0.03	0.43 ± 0.01
Evidence linear (w_{div}=0.5)		$\mathbf{0.07 \pm 0.07}$	$\mathbf{0.03 \pm 0.03}$	$\mathbf{0.01 \pm 0.01}$

Loss-based regularization partially reduces collapse on vanilla SA (e.g., $0.32 \rightarrow 0.06$ on sinusoids with both losses) but does not resolve it on sequential SA ($0.90 \rightarrow 0.30$ at best). Evidence depletion achieves lower collapse *without these additional losses*, indicating that on these tasks the mechanism, not the added regularization, drives the reduction.

Alternative anti-collapse mechanisms (Table 2). All alternatives replace the attention normalization or add a state/penalty inside the same slot-attention module, with identical tokenizer, slot count, iterations, losses, data, and 200-epoch budget (5 seeds per task). *Sparsemax* and *entmax-1.5* replace the softmax over slots by the corresponding sparse map (exact implementations). *SA-SH* normalizes the slot–token attention with 5 log-domain Sinkhorn iterations and adds Gaussian noise (σ=0.05) to the logits for symmetry breaking; this is the Sinkhorn slot attention that MESH builds on. *MESH* [Zhang et al., 2023] additionally optimizes the transport cost to minimize the entropy of the Sinkhorn plan (normalized-gradient steps on the cost, initialized with small noise), following the published algorithm; we verified that assignment entropy decreases monotonically over its inner iterations and that attention sharpens. *Coverage (mechanism)* adds $-\beta \cdot \text{cov}_\ell$ to the logits, where cov_ℓ accumulates attention received by token ℓ (β=3, matched to γ); *coverage (loss)* instead penalizes token-normalized cumulative attention. *Stick-breaking* assigns each slot a sigmoid scope over the remaining tokens, in the spirit of Genesis-V2's stick-breaking process (our variant is deterministic). Evidence and sequential variants, and the two sequential alternatives (coverage mechanism, stick-breaking), use τ=0.3; the parallel alternatives (sparsemax, entmax, SA-SH, MESH, coverage loss) use their native formulations without this temperature, which affects attention sharpness and therefore cosine overlap more than peak overlap; we report peak overlap as the primary metric for this comparison. Under cosine overlap, evidence depletion reaches 0.27/0.14 on Gaussians/multi-scale, whereas every alternative stays at or above 0.53 (MESH: 0.64/0.89/0.84 on sinusoids/Gaussians/multi-scale). Classical matching pursuit [Mallat and Zhang, 1993] is collapse-free by construction – it performs explicit residual accounting, the principle evidence depletion instantiates in attention space – and is accurate on well-separated sinusoids, but it requires an exact parametric template per source, returns point estimates rather than posteriors, and its subtraction premise does not hold under a confusion foreground; we therefore do not treat it as an architecture-matched baseline. A non-slot set predictor (learned queries, a single cross-attention pass) likewise stays at or above 0.53 cosine overlap.

Scope of comparisons. This paper studies slot collapse under *additive superposition*, where every token contains contributions from all sources simultaneously. This regime differs from object-centric vision (CLEVR, Multi-dSprites, MOVi), where different objects occupy disjoint spatial regions and tokens can be uniquely assigned to components. Methods designed for the spatial regime – such as MESH [Zhang et al., 2023], which modifies the assignment problem via optimal transport – target a different failure mode (tie-breaking under spatial ambiguity), but they are directly applicable to the additive setting, and we evaluate them as baselines: full MESH ties evidence depletion on sinusoids (Table 2) but is markedly worse on Gaussians and multi-scale, and SA-SH collapses on LISA (Sec. 3.4). The architecture-matched comparisons – vanilla and sequential SA, loss-based regularization, and the alternative attention mechanisms – isolate the contribution of the attention mechanism itself, which is the paper's core contribution.

I Stress tests, robustness, and cost

All experiments in this appendix use the synthetic harness of Sec. 3.1 (sinusoids, peak overlap, linear depletion) with the seed counts given.

Amplitude ratio. We vary the dominant-to-secondary amplitude ratio with all else fixed (5 seeds; Table 11). At equal amplitudes, evidence depletion is bimodal across seeds (per-seed collapse 0.066/0.736/0.179/0.150/0.736; mean 0.37) and does not improve over vanilla (0.30); the same holds at $1.5\times$ (0.41 vs. 0.33). From $2\times$ on, its mean collapse is below vanilla at every ratio, with the clearest gain under moderate dominance (0.21 and 0.12 at $3\times$ and $5\times$, vs. 0.34); seed variance remains high (std 0.14–0.22 except at $5\times$), and at $10\times$ the gap narrows (0.19 vs. 0.24). Vanilla collapse is roughly flat across ratios (0.24–0.34). The method thus targets dominance-driven redundant allocation rather than all mixture regimes.

Table 11: Amplitude-ratio sweep (sinusoids, peak overlap, mean over 5 seeds).

Ratio A_1/A_2	1	1.5	2	3	5	10
Vanilla SA	0.30	0.33	0.30	0.34	0.34	0.24
Evidence depletion	0.37	0.41	0.26	0.21	0.12	0.19

Slot over-provisioning and cardinality. With $K \leq 4$ and $S \in \{5, 8, 12, 16\}$ slots (3 seeds), attention-level collapse stays low for evidence depletion (0.035/0.082/0.037/0.092) versus vanilla (0.28/0.33/0.32/0.42), so we see no attention-level signature of premature masking. However, the fraction of evidence depleted at matched slots declines with S ($0.88 \to 0.71$ from S=5 to 16; unmatched $0.91 \to 0.56$), and the shared cardinality readout fails at $S \geq 8$: exact-K accuracy falls to $\approx$0 for *both* evidence depletion and vanilla. Since the collapse metric scores the first K (first-processed) slots, it cannot by itself detect damage to later slots; end-to-end scalability under substantial over-provisioning therefore remains unresolved. At higher cardinality ($K \leq 8$, S=9), evidence depletion keeps collapse lower than vanilla (0.24 vs. 0.54).

Token count. The architecture is stable up to L=32 tokens but unstable at L=64 (0.484 ± 0.396, 3 seeds). At N=256 samples, L=64 corresponds to 4-sample chunks with only three rFFT bins per token, so this instability is at least partly a tokenizer degeneracy; we nonetheless list it as a limitation.

Slot ordering. At training time, slot processing order is randomized via `RandPerm` in Algorithm 1 so that the model does not specialize slots to a particular index. At inference time we use a fixed canonical order (slot index $1, \ldots, S$). On LISA, Hungarian matching (Sec. 2.4) makes the loss permutation-invariant over slots, so randomization only diversifies the gradient signal each slot receives over the depletion sequence. The synthetic and FUSS harness instead uses an index-based existence target (the first K slots) with greedy parameter matching among the top-K slots by existence logit; there, randomization additionally decouples a slot's index from its position in the depletion sequence.

Early errors and slot order. Forcing the first slot to a corrupted allocation at inference raises collapse ($0.039 \to 0.144$ under uniform corruption; seeds 0–2 of the Table 1 linear runs, retrained, with 0.039 their unperturbed collapse in this evaluation) but leaves the fraction of remaining sources recovered by later slots essentially unchanged ($0.189 \to 0.195$; 0.208 under wrong-peak corruption): the perturbation degrades allocation locally but does not cascade. Randomizing the inference order changes collapse by less than 0.05 within a seed. Ordering slots greedily by remaining evidence at inference is slightly worse than the canonical order (0.070 vs. 0.039, identical parameter error); evidence-guided ordering *during training* remains an interesting direction.

Latency. Table 12 reports per-forward-pass inference time of the attention module on one NVIDIA GH200 (L=128, d=64, batch 32, I=3). Parallel attention is flat in S; evidence depletion grows linearly at $\approx$2.5 ms per slot ($\approx 4\times$ vanilla at the operating point S=5, $\approx 31\times$ at S=40). This linear growth is the real scaling cost of sequential processing and motivates hierarchical decomposition at large slot counts.

Table 12: Attention-module latency (ms per forward pass, GH200).

S	5	10	20	40
Vanilla SA	3.5	3.2	3.3	3.3
Evidence depletion	13.2	26.1	50.5	101.1

J FUSS benchmark details

The Free Universal Sound Separation (FUSS) benchmark [Wisdom et al., 2021] contains real-world audio mixtures of 1–4 everyday sounds (speech, music, environmental noise, etc.) summed additively at 16 kHz in 10-second clips (released on Zenodo under CC BY 4.0). We crop to 4 seconds (64,000 samples), tokenize with 32 ChunkFFT chunks (FFT, project to d=128), and extract 3 ground-truth parameters per source: dominant frequency, RMS amplitude, and spectral centroid (all normalized to $[0, 1]$). Table 13 reports per-seed results.

Table 13: FUSS results: slot collapse rate per seed (200 epochs, 5,000 train / 1,000 val, 32 tokens, d=128).

Method	Seed 0	Seed 1	Seed 2	Seed 3	Seed 4	Mean $\pm$ std
Vanilla SA	0.24	0.33	0.37	0.23	0.32	0.30 ± 0.05
Sequential SA	0.96	0.92	0.96	0.92	0.94	0.94 ± 0.02
Evidence linear	0.04	0.04	0.04	0.06	0.02	$\mathbf{0.04 \pm 0.01}$
Evidence quadratic	0.30	0.25	0.26	0.28	0.19	0.26 ± 0.04

Quadratic depletion $(1-\alpha^2)$ improves only marginally over vanilla on FUSS (0.26 vs. 0.30) – it is nearly inert on spectrally complex audio. Linear depletion $(1-\alpha)$ achieves a $\approx 8\times$ reduction in peak overlap ($0.30 \to 0.04$); cosine overlap is 0.79/0.97/0.50 for vanilla/sequential/linear (Table 8). These are the runs reproduced by our released code; an earlier run of the same configuration with a previous version of our FUSS script, reported in the submitted version, gave 0.29/0.93/0.05/0.31 (vanilla/sequential/linear/quadratic).

Separation diagnostics. To probe downstream utility, we converted each slot's attention map into a time–frequency mask without additional parameters and measured SI-SDR improvement over the mixture. No variant separates meaningfully (SI-SDR improvement between -2.1 and $+1.1$ dB across all variants), whereas an oracle mask from the same readout family gains $+7.7$ dB; the readout family is therefore capable, but attention maps trained for parameter inference rather than reconstruction are not separation-optimal. Per-source estimates of the three spectral parameters likewise did not rise meaningfully above simple baselines for any variant. We therefore use FUSS only to test whether slot collapse and its mitigation transfer to mixtures of real recorded audio, and make no waveform-separation claim.

Practical guidance on depletion form. Across all synthetic benchmarks and FUSS, linear depletion matches or outperforms quadratic. On LISA, both prevent collapse: linear gives lower overlap (0.030 vs. 0.054), quadratic slightly better NLL (-5.94 vs. -5.49; Sec. 3.4). We recommend linear $(1-\alpha)$ as the starting point; quadratic $(1-\alpha^2)$ may be worth trying when the downstream task uses density estimation (e.g., normalizing flows), given its slightly better NLL on LISA (3 seeds). Developing a principled selection criterion – e.g., based on the spectral entropy of the input or the number of sources – is an open direction.

K Visualizations

Additive superposition in the audio domain. Figure 3 contrasts the partitionable token structure of object-centric images with additive superposition in audio, using synthetic signals and our FUSS tokenization.

Evidence depletion on a simulated LISA input. Figure 4 illustrates the allocation argument of Sec. 2.2 on a simulated LISA input: a K=3 sample with three sources at nearly identical frequencies ($\Delta f < 1\,\mu$Hz), all contributing to a single unresolved spectral feature. The attention heatmaps (5

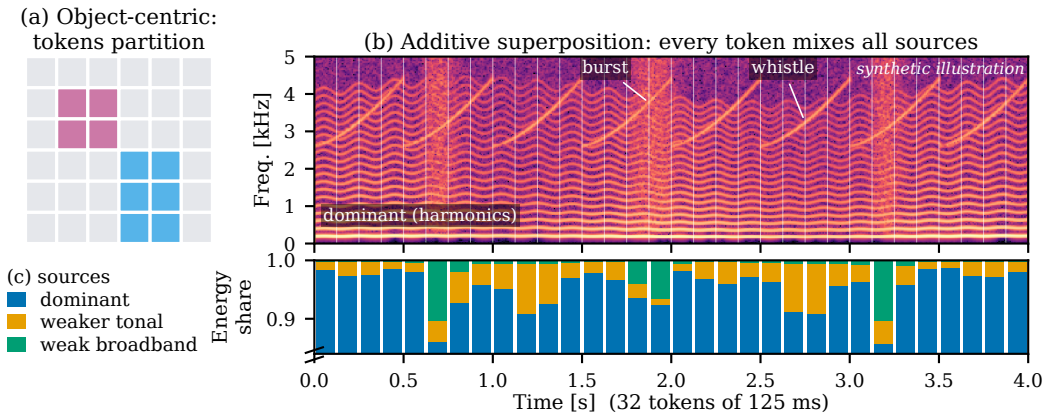

Figure 3: **Partition versus additive superposition** (synthetic illustration). (a) In object-centric images each patch token belongs to one object, so assigning tokens to slots separates objects. (b) In a 4 s, 16 kHz audio mixture split into 32 tokens of 125 ms (our FUSS tokenization), a dominant harmonic source, a $6\times$ weaker whistle, and an $8\times$ weaker broadband source overlap in every token. (c) Per-token energy shares (y-axis truncated at 0.84): the dominant source supplies 86–99% of every token's energy, so choosing *different tokens* mostly reads the *same source*.

slots $\times$ 128 tokens, organized as 4 temporal segments $\times$ 32 frequency patches) show how each variant handles this shared-input regime.

Vanilla SA (overlap 0.98) exhibits exactly the failure mode predicted by the memoryless-attention analysis: all slots distribute attention uniformly because the dominant component attracts every slot equally, and there is no mechanism to redirect subsequent slots. Sequential SA (overlap 0.48) provides some gradient diversity through ordering but no state over explained evidence – the slots still largely attend to the same tokens. With evidence depletion (overlap 0.07), the picture changes qualitatively: once the bias suppresses already-attended frequency bins, remaining slots discover complementary structure along LISA's orbital modulation axis – the sources are unresolved in frequency but separable in time via direction-dependent antenna patterns. This illustrates that evidence depletion does not merely reduce overlap numerically but redirects attention to complementary features, consistent with Proposition 1: dissimilar readouts require moving to tokens not dominated by the same component.

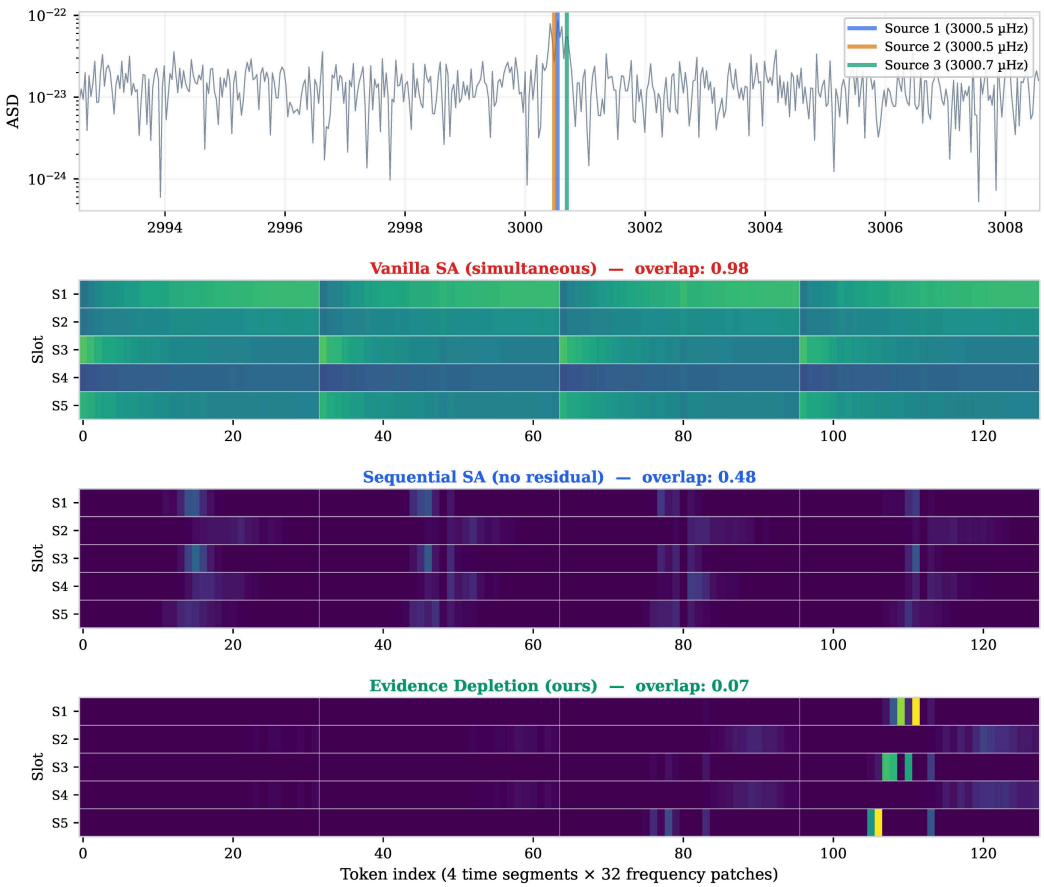

Figure 4: **Evidence depletion produces distinct attention on a simulated LISA input.** *Top:* Amplitude spectral density of a $K{=}3$ input with three nearly co-frequency sources ($\Delta f < 1\,\mu$Hz), forming a single unresolved feature. *Bottom:* Attention heatmaps (5 slots $\times$ 128 tokens) for three variants. **Vanilla SA:** All slots attend nearly uniformly (overlap 0.98). Because every slot observes the same input, gradients are dominated by the same component and all slots settle on the same allocation (collapse). **Sequential SA:** Processing slots one at a time introduces ordering but not state; the slots still largely attend to the same tokens (overlap 0.48). **Evidence depletion:** After each slot, attended tokens are downweighted via a residual evidence variable, changing the effective input seen by subsequent slots, which move to complementary tokens (overlap 0.07), here exploiting temporal modulation induced by LISA's orbital motion rather than frequency separation. Token indices: 4 temporal segments $\times$ 32 frequency patches; segment boundaries shown as white lines.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and contributions (Sec. 1) identify slot collapse as a structural limitation, propose evidence depletion, and validate via controlled ablations. All claims are supported by experiments in Sec. 3.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Sec. 5 lists the limitations: proxy nature of the collapse metrics, no identifiability guarantee, the equal-amplitude regime, linear cost in S and the cardinality readout under slot over-provisioning, L=64 instability, empirically chosen depletion form and hyperparameters, and the specialized, simulated LISA setting with its seed counts and prior-limited parameters. Details appear where they arise: encoder compression (App. C), data-space vs. evidence-space (Sec. 3.1), γ/τ sensitivity (App. E), stress tests (App. I), FUSS separation diagnostics (App. J).

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Prop. 1 (readout alignment under uniform dominance) states its assumptions (common unit direction, per-token dominance ratio $\kappa < 1$, nonnegative weights) in the main text, with a complete proof, tightness example, and scope in App. D. Prop. 2 (monotone depletion) states its assumptions (depletion form $f(\alpha) \in [0, 1]$, floor ϵ, $\sum_j f(\alpha_{j\ell}) = \infty$ for the limit statement) and gives a complete proof. Neither is an identifiability result, which we state explicitly.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Sec. 3 specifies the synthetic, FUSS, and LISA benchmark parameters (signal types, N, K, seeds, training budget); App. H gives the synthetic and FUSS optimizer settings and noise level. Appendices A–C describe the LISA architecture, training, and losses; the LISA ablation's conclusions depend only on the attention module, which is fully specified (Sec. 2.3, Algorithm 1), while some LISA feature-extraction details (e.g., the exact hand-crafted features) are given only at the level of the architecture description. FUSS is publicly available [Wisdom et al., 2021]. Code for the synthetic, baseline, and FUSS experiments is released (see Q5).

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Partially. Code for the synthetic, baseline, and FUSS experiments is released with reproduction instructions at `https://github.com/nhouba/when-attention-collapses` (also installable as `pip install evidence-depletion`). The LISA simulation and training pipeline is not part of the code release; its architecture, training, and losses are described in Appendix A. FUSS is publicly available from Zenodo; LISA data is generated from publicly available tools (JaxGB, SciRDv1).

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Sec. 3 and App. H (synthetic: sample counts, epochs, optimizer, seeds, metrics; FUSS: Sec. 3.2 and App. J; LISA: data model, architecture summary, training budget). Appendix A (full architecture, loss weights, angular encoding, hyperparameters, 10^6 samples, batch 128, Adam, LR schedule, 150 epochs, 4–8$\times$GH200).

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Synthetic, FUSS, and alternative-mechanism experiments report mean $\pm$ std over 5 seeds. The LISA ablation (Table 3) reports mean $\pm$ std over 5 seeds; additional LISA controls use 1–3 seeds (the SA-SH hyperparameter sweep and the 100-epoch convergence extension use one seed), as stated. Sensitivity sweeps (App. E) and stress tests (App. I) use 3–5 seeds, as stated.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: App. A.3: 4–8$\times$NVIDIA GH200 (120 GB), 150 epochs for the full LISA model. Synthetic and FUSS experiments: single GPU, $\sim$2–10 hours per sweep. Loss baselines: single GPU, $\sim$12 hours total. The additional LISA ablations (App. B.1) use the same hardware at $\approx$4.5 hours per 30-epoch run; alternative-mechanism and stress-test sweeps use single GPUs at $\sim$0.5 hours per run.

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The work develops a general inference mechanism validated on synthetic benchmarks and a scientific application. No human subjects, private data, or dual-use concerns.

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Sec. 1 and Sec. 5 discuss broader applicability across audio, spectroscopy, and scientific domains. We do not foresee negative societal impacts from a compositional inference mechanism for scientific data.

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The method is a general attention mechanism for signal decomposition. No risk of misuse beyond standard ML applications.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All software (JaxGB, BSD-3-Clause; lisaorbits, BSD-3-Clause; nflows, MIT; PyTorch, BSD-style) and data specifications (SciRDv1) are cited and used under their

licenses. FUSS dataset [Wisdom et al., 2021] is used under its CC BY 4.0 license. Synthetic data is generated by our code.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Code and experiment scripts are released with documentation at `https://github.com/nhouba/when-attention-collapses`.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No human subjects or crowdsourcing involved.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects involved.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were not used as a component of the core methodology. Any use was limited to writing assistance and does not impact the scientific content.