BEYOND FIRING LESS: ACTIVITY OBJECTIVES SHAPE POPULATION ORGANIZATION IN A SPIKE-DRIVEN TRANSFORMER

Anonymous authors
Paper under double-blind review

ABSTRACT

An activity objective specifies what training rewards. The statistic measured after training can move opposite to the objective's nominal direction, because training can reorganize which responses remain active. We study this mismatch in a Spike-Driven Transformer using per-input population firing-rate coefficient-of-variation regularization (PopFR-CV), together with mean-rate, lifetime-sparsity, and entropy-style objectives. PopFR-CV explicitly rewards greater firing-rate variation across activation coordinates for each input. Training produces the reverse measured response: evaluation CV falls by more than an order of magnitude, the equal-tensor silent-coordinate fraction increases by over 30 percentage points, mean firing decreases only modestly, and endpoint accuracy remains comparable. Controlled continuations from identical mature states show immediate trajectory separation followed by accuracy recovery. Paired CIFAR-100 endpoints localize the reorganization to strong activity withdrawal from monitored attention pathways; fully retained non-attention responses show little change in mean CV. The low-CV, enlarged-silence endpoint recurs across three static-image datasets within the studied Transformer. Other spiking systems and event data show distinct responses. These findings show that population participation is a distinct outcome of activity regularization that is not reliably predicted by mean firing or by the nominal direction of the optimized statistic.

1 INTRODUCTION

Recent methods make SNNs more efficient by suppressing events or dropping tokens, thereby converting sparse activity into fewer operations (Bu et al., 2025; Zhuge et al., 2024). We ask how an activity objective reshapes both total firing and the location of activity across a trained network. This distinction matters because two networks with similar mean firing can rely on different coordinates and pathways.

We study this question without pruning, token selection, or changes to the dense execution graph. Each objective is evaluated through four complementary outcomes: task accuracy, mean firing rate (MFR), per-input coefficient of variation (CV), and the fraction of activation coordinates that remain silent over the evaluation set. An activation coordinate is a channel–position or token–feature index. Evaluation silence measures dataset-level participation. It carries no biological-death interpretation and applies no pruning mask.

Our primary system is a four-timestep Spike-Driven Transformer (SDT) (Yao et al., 2023). We compare direct mean-rate suppression, lifetime Hoyer sparsity, a Bernoulli-entropy objective, per-input population firing-rate CV (POPFR-CV), and weight-CV controls. These objectives summarize activity along different axes and reach different joint outcomes (Fig. 1). Direct rate suppression lowers firing most. Entropy-style regularization raises firing and reduces the silent-coordinate fraction. POPFR-CV modestly lowers firing and greatly expands that fraction. Firing, dispersion, and participation therefore describe distinct properties.

The most striking case is POPFR-CV. Its auxiliary loss explicitly rewards *greater* variation across activation coordinates within each input. Training nevertheless produces the opposite measured response: across five matched seeds on CIFAR-10 and CIFAR-100, evaluation CV falls 11–12-fold

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and the equal-tensor silent-coordinate fraction rises by 31–33 percentage points. We call this the *objective–outcome inversion*.

A common-state branch test separates intervention-induced redirection from an association between independently trained endpoints. Restoring the same mature model, optimizer, scheduler, data order, and random state into paired CIFAR-100 branches produces immediate CV separation; accuracy first falls and later recovers. Aligned endpoint readouts then locate the aggregate change: participation withdraws strongly from monitored attention-path tensors, while all non-attention responses remain active in both branches and their group-mean CV changes little. The main endpoint difference lies in response participation; total firing changes little.

Finally, we use Hoyer sparsity as a related-objective control. Ideal CV and Hoyer sparsity have identical ordering on the same nonzero vector. Aggregation assigns response-dependent weights during task-coupled training. Source-matched per-input and lifetime Hoyer continuations reproduce the qualitative low-CV, enlarged-silence response and reach different quantitative endpoints. Validation-selected and larger-dataset confirmation establish recurrence within static-image SDT; several other systems delimit the scope. An objective's nominal direction and aggregation-axis label are insufficient predictors of the population organization ultimately learned.

Our contributions are:

- **Population participation emerges as a distinct design outcome.** Mean-rate, lifetime-sparsity, entropy-style, and per-input objectives produce distinct combinations of firing, dispersion, and dataset-level participation. Similar CV reductions can therefore encode opposite ways of organizing network activity.
- **The optimized statistic can move opposite to its rewarded direction.** POPFR-CV rewards higher per-input CV, yet static-image SDT learns much lower hard-spike CV and a larger silent-coordinate set. Common-state continuations establish training redirection and accuracy recovery, and aligned readouts locate the aggregate change in selective withdrawal from monitored attention pathways.
- **The results motivate a joint evaluation principle for activity objectives.** Related-objective, component, dataset, and system controls show that loss geometry alone cannot forecast the learned endpoint. Evaluating task performance, total activity, within-response dispersion, and cross-dataset participation together reveals what an objective actually changes and where that response holds.

2 RELATED WORK

Operational sparsification in SNNs. Activity Pruning suppresses spike events and reports lower firing and modeled synaptic operations, including a full-ImageNet experiment combined with weight pruning (Bu et al., 2025); STATA drops tokens, yielding GPU training speed and memory gains plus operation-based inference-energy estimates (Zhuge et al., 2024). They establish an operational route from exploitable sparsity to less work. We retain the dense execution graph and ask how objectives reshape hard-spike dispersion and participation. Silent-coordinate fraction is an organizational readout and makes no savings claim.

Activity objectives in SNNs. Mean-firing penalties directly reduce spiking activity (Pellegrini et al., 2021; Narduzzi et al., 2022). IM-Loss maximizes a Bernoulli-information surrogate (Guo et al., 2022). We use surrogate-gradient training with TET classification loss (Neftci et al., 2019; Cramer et al., 2022; Deng et al., 2022). These objectives constrain different activity summaries.

Sparsity and CV. Hoyer's normalized ℓ_1/ℓ_2 sparsity (Hoyer, 2004; Yang et al., 2020) is a strict monotone of ideal CV on the same nonzero vector. We apply this geometry per input during task-coupled training and characterize the resulting organization. Parameter pruning and dynamic sparse training alter connectivity (Han et al., 2015; Mocanu et al., 2018; Evci et al., 2020). Our procedure preserves every parameter and uses no mask or separate fine-tuning.

Representation objectives. DeCov, orthogonality constraints, and spectral regularization target activation or parameter redundancy (Cogswell et al., 2016; Bansal et al., 2018; Yoshida & Miyato,

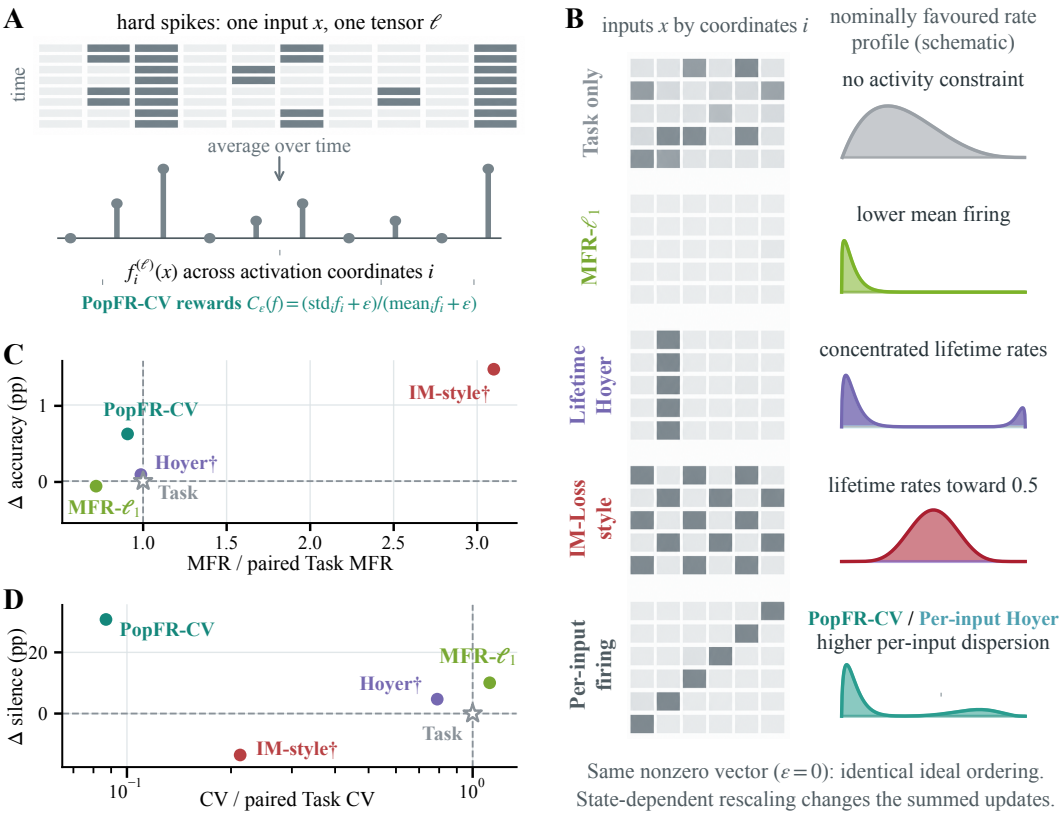

Figure 1: **From activity objectives to learned endpoints. (A)** For each input and monitored tensor, PopFR-CV time-averages hard spikes, forms a firing-rate vector across activation coordinates, computes its CV, and then averages over inputs and tensors ($\epsilon = 10^{-8}$). **(B)** The matrices show the aggregation axis used by each objective; the curves depict schematic rate profiles favored by the nominal objective. Measured endpoint distributions appear in the subsequent figures. MFR-ℓ_1 suppresses the mean, lifetime Hoyer favors concentrated lifetime rates, and the entropy-style objective favors intermediate lifetime rates. PopFR-CV and per-input Hoyer act on the same per-input firing vector and have identical ideal same-vector ordering. Their aggregated task-coupled updates can diverge. **(C–D)** CIFAR-10 five-seed means preview the measured accuracy–firing and dispersion–participation endpoints (star: Task). Figure 2 shows both datasets, individual seeds, and uncertainty. Daggers denote exploratory coefficients.

2017). Our comparison tracks the joint accuracy, firing, dispersion, and participation outcomes of activity objectives.

3 What is optimized, measured, and controlled

3.1 Objectives: the same activity, summarized along different axes

Let $s_{t,\ell,i}(x) \in \{0, 1\}$ be the spike of coordinate i in monitored tensor ℓ at time t for input x, and let

$$f_{\ell,i}(x) = \frac{1}{T} \sum_{t=1}^{T} s_{t,\ell,i}(x). \tag{1}$$

Here i flattens every non-batch, non-time dimension: it indexes a channel–spatial-position pair in a feature map or a token–feature pair in a token representation. The term *activation coordinate* avoids a

biological-neuron or channel interpretation. For each input and tensor, the implementation computes

$$C_\epsilon(f_\ell(x)) = \frac{\mathrm{std}_i\, f_{\ell,i}(x) + \epsilon}{\mathrm{mean}_i\, f_{\ell,i}(x) + \epsilon}, \qquad \epsilon = 10^{-8}, \tag{2}$$

and adds

$$\mathcal{L}_{\mathrm{PopCV}} = -\lambda_{\mathrm{pop}} \mathbb{E}_{x,\ell}[C_\epsilon(f_\ell(x))] \tag{3}$$

to the task loss. The minus sign makes the nominal pressure unambiguous: larger per-input CV is rewarded. For a minibatch $x_1, \ldots, x_B$, define the lifetime firing vector by $p_{\ell i} = B^{-1} \sum_b f_{\ell i}(x_b)$. A *population* statistic such as PopFR-CV compares coordinates separately within each input before averaging inputs; the lifetime comparators act on p_ℓ after averaging each coordinate over samples and time. Figure 1 visualizes this distinction.

PopFR-CV denotes Eq. (3). WeightCV applies an analogous loss to absolute weights, and Weight+PopCV combines the weight and population terms. We compare these CV variants with MFR-ℓ_1, which directly penalizes mean firing; lifetime Hoyer sparsity, which favors concentration in coordinate-wise lifetime rates; and an IM-Loss-style Bernoulli-entropy objective, which favors lifetime firing probabilities near one half. Figure 1 separates these nominal preferences from the endpoints that training actually reaches. Exact definitions and coefficients are in Appendix A.

3.2 Measurements: task performance, firing, dispersion, and participation

The three activity readouts answer different questions. MFR asks how much the network fires on average. Per-input CV asks how unevenly that activity is distributed across coordinates within a response. The silent-coordinate fraction asks which coordinates never participate over the evaluation set.

For $\overline{C}_\epsilon = \mathbb{E}_{x,\ell}[C_\epsilon(f_\ell(x))]$, we take CV across coordinates per input, then average across inputs and tensors; MFR uses the same aggregation order. For silence, we accumulate each coordinate over the evaluation set, calculate each tensor's silent fraction, and average tensors equally. Evaluation-silent means no observed spike, operationally a mean rate below 10^{-6}. This dataset-level quantity differs from a whole input–tensor response containing no spikes (Sec. 5). Appendix Table A1 gives exact weighting and coordinate-weighted alternatives.

3.3 Controls: endpoints, common-state interventions, and related objectives

The primary model is a two-block, 256-dimensional **Spike-Driven Transformer (SDT)** (Yao et al., 2023) with multi-step LIF units, four timesteps, TET classification loss (Deng et al., 2022), and 300-epoch AdamW training. Table 1 separates the roles of the principal protocols. The matched fixed-coefficient comparison uses five seed indices per condition on CIFAR-10/100; auxiliary terms begin after 20 task-only epochs. Lifetime-Hoyer and IM-Loss-style coefficients were chosen in preliminary sweeps and provide descriptive operating points under distinct tuning histories. All outcomes are retained, including one collapsed CIFAR-100 Hoyer run. Appendix A gives the full objectives, coefficients, optimizer settings, tensor inventory, and evaluation procedure; Appendix Fig. A1 maps the systems and readout locations.

To test whether PopFR-CV changes training from a common state, task-only SDT is trained through epoch 100 on a fixed CIFAR-100 train/validation split. Epoch 100 provides a mature task-trained source state while retaining 200 epochs for paired adaptation under the original schedule. Each seed's complete model, optimizer, scheduler, and random state is then restored into matched Task and Task+PopFR-CV branches, which receive the same sample order and stochastic choices through epoch 300. The protocol evaluates redirection from this source state, and the official test set is not used. The same nominal intervention is run in a native SEW-ResNet-18 (Fang et al., 2021). This complete-system comparison bounds transfer across different activation inventories and effective objective scales; it does not isolate architecture.

A source-matched component-removal continuation compares full $C_\epsilon = (\sigma + \epsilon)/(\mu + \epsilon)$ with $D_\epsilon = \sigma/(\mu + \epsilon)$ at the same coefficient and training stream. Every endpoint is evaluated with the same C_ϵ readout; D_ϵ changes only the training objective. This tests whether the removed numerator term is necessary for the measured response (Appendix C.9).

We use a separate, pre-specified Hoyer extension to compare CV and Hoyer on the per-input axis and Hoyer on the lifetime axis. Five epoch-100 CIFAR-100 SDT states (seeds 301–305) are restored into four branches. Source-training batches select each Hoyer coefficient to match the *magnitude* of POPFR-CV's initial three-update, task-relative AdamW deviation; update direction and later evolution are deliberately left unconstrained. Appendix E.2 gives the calibration equation, zero-boundary implementation, errors, and provenance.

We separately assess generalization after validation-only coefficient selection on CIFAR-10. Three tuning seeds compare four fixed coefficients; a rule fixed before official-test evaluation selects the smallest coefficient meeting an accuracy tolerance and activity threshold. The selected value is evaluated once on the official test split with three fresh paired seeds (Appendix C).

Table 1: Principal protocols and their distinct evidential roles. Full settings and outcomes are in the appendix.

Protocol	Question	Key control or boundary
Fixed objectives	What joint endpoints arise?	Same SDT, duration, and seed indices; descriptive coefficients; all failures retained
Common state	Does POPFR-CV redirect training?	Complete epoch-100 state and stochastic stream; validation only
Paired endpoint readout	Which responses account for lower CV?	Fixed endpoints and aligned inputs/tensors; post hoc common-nonzero conditioning; no checkpoint or run selection
Component removal	Is numerator stabilization required?	Same source, stream, coefficient, and endpoint readout
Matched Hoyer	Do related objectives end alike?	Same source; initial displacement magnitude calibrated without endpoint access
Held-out confirmation	Does validation selection confirm?	One official-test evaluation on fresh paired seeds

Seeds are the independent units, and each experimental protocol is analyzed separately. For paired differences, $\pm$ is the sample standard deviation of same-seed differences; for marginal rows, it is the sample standard deviation across seeds. The matched-axis comparison reports paired seed-level means and pointwise 95% Student-t intervals; its protocol-specified contrasts compare per-input Hoyer with POPFR-CV and lifetime with per-input Hoyer. Epochs, batches, tensors, and coordinates are not replicates. The modest accuracy changes are not treated as a universal improvement or a formal non-inferiority claim.

4 DIFFERENT OBJECTIVES PRODUCE DIFFERENT ACTIVITY ORGANIZATIONS

4.1 FIRING, DISPERSION, AND PARTICIPATION FORM DISTINCT OUTCOME AXES

Figure 2 and Table 2 answer the first question: what combination of accuracy, firing, dispersion, and participation does each objective produce? Direct rate suppression behaves as expected: MFR-ℓ_1 gives the lowest mean firing on both datasets while keeping mean accuracy close to Task. The IM-Loss-style configuration raises firing substantially and reaches the highest mean CIFAR-10 accuracy among these endpoints. Lifetime Hoyer remains near Task on CIFAR-10; one of five CIFAR-100 runs collapses, so the appendix shows every outcome and we report no stable-only aggregate.

POPFR-CV reaches a different operating point. It modestly lowers firing, retains comparable or higher mean accuracy in the matched runs, and produces by far the largest change in coordinate participation. The comparison separates three outcome axes: mean firing, per-input CV, and the silent-coordinate fraction describe different properties and cannot substitute for one another.

4.2 SIMILAR CV REDUCTIONS CAN CONCEAL OPPOSITE PARTICIPATION CHANGES

On CIFAR-10, IM-Loss style and POPFR-CV both lower CV. IM-Loss style raises MFR to $3.10\times$ Task and lowers silence by 13.54 points; POPFR-CV lowers MFR to $0.91\times$ and raises silence by 30.76 points. CIFAR-100 repeats the contrast. Thus lower CV can accompany opposite changes in firing and dataset-level coordinate participation.

The weight-axis controls answer a separate design question. WEIGHTCV stays close to Task, whereas WEIGHT+POPCV stays close to POPFR-CV; adding weight CV provides no consistent

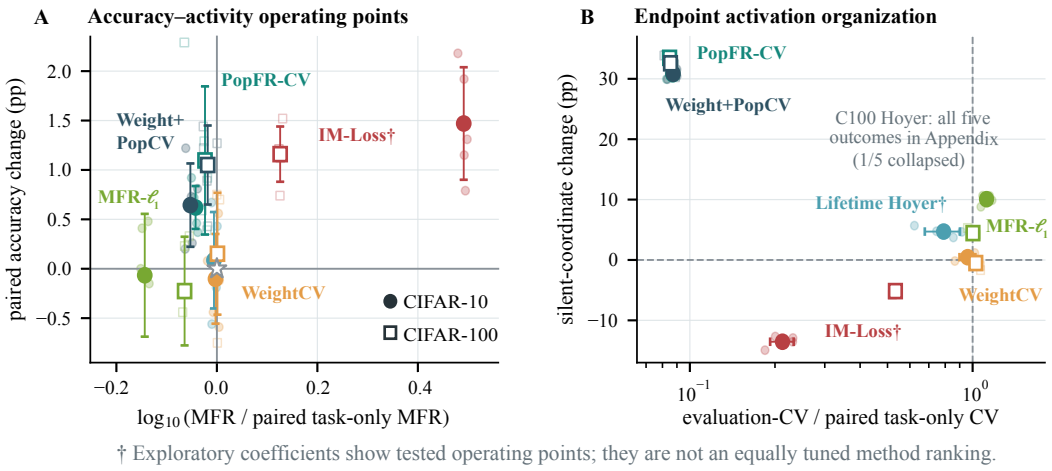

Figure 2: **The objectives separate firing from population participation.** **(A)** Paired accuracy change versus the log MFR ratio to Task. **(B)** Paired evaluation-CV ratio versus the change in equal-tensor silent-coordinate fraction. Filled circles denote CIFAR-10 and open squares CIFAR-100; light markers are seeds, and large markers with bars are means ± sample SD ($n = 5$). IM-Loss style and PopFR-CV both lower CV. Their firing and participation outcomes move in opposite directions. Population-CV configurations show an 11–12× CV decrease together with a 31–33 percentage-point silence increase. Daggered Hoyer and IM-Loss-style coefficients are exploratory operating points produced under separate tuning histories. One CIFAR-100 lifetime-Hoyer run collapsed; all five outcomes are retained in the appendix and no stable-only aggregate is plotted.

Table 2: **Observed SDT operating points at the reported coefficients.** All conditions use $T = 4$, 300 epochs, onset after 20 task-only epochs, and five seed indices. Task, MFR-ℓ_1, and PopFR-CV form the matched fixed-coefficient comparison. †Hoyer and IM-Loss-style coefficients were chosen in preliminary sweeps and show additional trade-offs under distinct tuning histories. One CIFAR-100 Hoyer run collapsed; no aggregate is reported. Values are mean ± sample SD.

Data	Condition	Accuracy (%)	MFR	Eval. pop. CV	Silent (%)
C10	Task	87.78±.39	.0585±.0006	19.15±.71	22.05±.56
	MFR-ℓ_1	87.71±.56	.0421±.0008	21.43±.44	32.12±.44
	PopFR-CV	88.40±.19	.0531±.0007	1.664±.012	52.81±.14
	Lifetime Hoyer†	87.87±.17	.05775±.00031	15.12±2.16	26.76±.66
	IM-Loss style†	89.25±.31	.18140±.00219	4.060±.254	8.51±.48
C100	Task	65.79±.31	.0734±.0008	18.26±.55	19.07±.58
	MFR-ℓ_1	65.56±.27	.0633±.0005	18.30±.60	23.52±.73
	PopFR-CV	66.88±.59	.0695±.0009	1.547±.008	52.54±.04
	Lifetime Hoyer†	1/5 chance-level collapse; all outcomes in Appendix Table A4			
	IM-Loss style†	66.95±.13	.09802±.00046	9.736±.380	13.91±.27

endpoint advantage. Population-only CV is therefore the parsimonious configuration examined below; Appendix Table A2 gives the complete weight ablation.

The PopFR-CV endpoint is especially informative because a large increase in the silent-coordinate fraction accompanies only a modest reduction in mean firing, while evaluation CV moves opposite to the reward in Eq. (3). This objective–outcome inversion motivates the controlled test in Sec. 5: does adding the objective redirect training from the same mature state?

4.3 THE POPULATION-CV ENDPOINT SURVIVES VALIDATION SELECTION AND DATASET SCALE

The validation rule selects $\lambda_{\text{pop}} = 2 \times 10^{-4}$. On three fresh CIFAR-10 official-test pairs, every seed has lower MFR and CV, a larger silent-coordinate fraction, and higher accuracy: CV falls by $91.08 \pm 0.28\%$, silence rises by 32.783 ± 0.392 points, and accuracy changes by $+1.057 \pm 0.179$ points (Appendix Table A6). On Tiny-ImageNet, five matched fixed-coefficient pairs similarly move CV from 29.12 to 1.70 and silence from 30.72% to 53.41%, with $+0.580 \pm 0.372$ accuracy points. These protocols extend the endpoint within SDT; cross-system responses are separate.

5 POPULATION-CV REDIRECTS A COMMON TRAINING STATE

5.1 THE TRAJECTORIES DIVERGE IMMEDIATELY, WHILE ACCURACY LATER RECOVERS

Every Task and Task+POPFR-CV pair in the controlled continuation begins from the same epoch-100 SDT model and complete training state within a seed. One epoch after the split, the geometric Task+POPFR-CV/Task validation-CV ratio is already 0.230. It reaches 0.0911 at epoch 300, with lower endpoint CV in all five seeds. Accuracy reacts on a different timescale: the mean paired difference initially falls by 20.288 points, then recovers to $+0.276 \pm 0.412$ points at the endpoint. This paired design attributes the subsequent trajectory divergence to adding the objective after a common training state.

5.2 THE AGGREGATE CV FALL LOCALIZES TO MONITORED ATTENTION PATHWAYS

An algebraic endpoint decomposition separates whole-zero occupancy from each branch's conditional-nonzero CV and firing (Appendix Table A9). Those conditional sets can contain different events, so we pair the same validation input and tensor across endpoints. Events nonzero in both branches cover $48.36 \pm 1.20\%$ of the 23-tensor inventory. Their geometric POPFR-CV/Task CV ratio is 1.157 $[0.784, 1.708]$: the point estimate is above one, the interval includes both a moderate decrease and increase, and the paired subset does not reproduce the order-of-magnitude aggregate fall.

All 11 stem, shortcut, MLP, and head tensors remain nonzero for every input in both branches and have CV ratio 1.0046 $[1.0015, 1.0077]$, a numerically small change; only 1.032% of attention-path events are nonzero in both branches. Together with the attention tensors' large increases in whole-zero responses (Appendix Fig. A3), these results localize the aggregate CV drop to attention-path withdrawal. Uniform within-response equalization receives no support from this readout. Accuracy recovery rules out whole-network collapse. Functional dispensability of any attention path remains untested. Dataset-level silence rises by 31.94 ± 0.46 equal-tensor points (14.89 ± 0.28 coordinate-weighted), while MFR is only $2.09 \pm 1.08\%$ lower.

5.3 REMOVING THE NUMERATOR STABILIZER MAINLY WEAKENS THE EARLY TRANSITION

To test a second explanation, write $C_\epsilon = D_\epsilon + B_\epsilon$, with $D_\epsilon = \sigma/(\mu + \epsilon)$ and $B_\epsilon = \epsilon/(\mu + \epsilon)$. In five matched continuations we delete B_ϵ, retain the coefficient and training stream, and evaluate every endpoint with C_ϵ. Deletion weakens the epoch-101 response (0.761 versus 0.212). Strong endpoints persist: D_ϵ and full C_ϵ reach ratios 0.1250 and 0.0877, about 32 additional silence points, and recovered accuracy. Their paired endpoint-CV ratio is 1.425 ($[1.196, 1.697]$; higher in 5/5 seeds), while the frozen retention summary is $\rho = 0.959$. B_ϵ shapes the early trajectory; the strong long-run response persists without it. The intervention leaves the mechanism non-unique (Appendix C.9).

6 WHERE THE RESPONSE RECURS ACROSS TESTED SETTINGS

The complete endpoint appears across CIFAR-10, CIFAR-100, and Tiny-ImageNet in four-timestep SDT. Eight-timestep SDT, lower-coefficient Spikformer, native SEW, a spiking MLP, and CIFAR10-DVS show partial, weak, or absent responses. These complete-system comparisons define empirical boundaries; the source of cross-system variation remains unresolved (Appendix Tables A12 and A14).

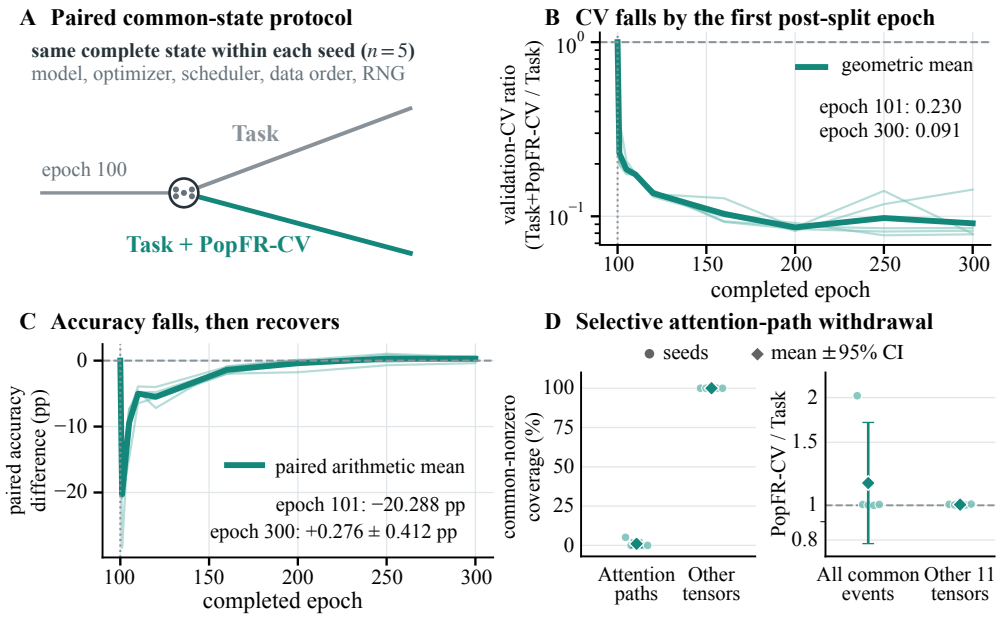

Figure 3: **Population-CV redirects training and withdraws activity selectively from monitored attention pathways. (A)** Within each seed ($n = 5$), the complete epoch-100 model, optimizer, scheduler, data order, and random state are restored into matched Task and Task+PopFR-CV continuations. **(B)** Validation CV separates by the first post-split epoch and ends at a geometric ratio of .091. **(C)** Accuracy initially falls by 20.288 points and recovers to $+0.276 \pm 0.412$ points relative to Task. **(D)** For aligned input–tensor events, common-nonzero coverage is nearly absent in 12 attention-path tensors and complete in the other 11; the fully retained non-attention group has a CV ratio near one. Points are optimization seeds. Coverage diamonds show arithmetic means $\pm$ sample SD; CV diamonds show geometric means with pointwise 95% intervals. This localizes observed withdrawal. Causal attribution and functional dispensability remain unresolved. No official test data are used.

7 Related objectives share a response and reach different endpoints

For one nonzero response vector $f \in \mathbb{R}^n_{\geq 0}$, $n > 1$, with positive mean, population variance, and ideal $\epsilon = 0$, normalized Hoyer sparsity and CV satisfy

$$H(f) = \frac{\sqrt{n}}{\sqrt{n}-1}\left(1 - \frac{1}{\sqrt{1+\mathrm{CV}(f)^2}}\right). \tag{4}$$

This strictly increasing mapping gives the same ordering and collinear gradients for one positive, nonuniform vector. Aggregation introduces response-dependent scaling, so the task-coupled training objectives differ. Figure 4 asks how different the learned endpoints remain after matching the magnitude of the initial three-update perturbation.

All three branches lower evaluation CV and add 29.6–32.1 silence points relative to Task. Per-input Hoyer ends at $4.94\times$ PopFR-CV's CV (95% interval $[4.08, 5.98]$) and a lower-MFR operating point; its accuracy difference is unresolved. Lifetime Hoyer changes the aggregation axis and produces a numerically similar silence increase. Same-vector ordering and an axis label are therefore insufficient endpoint predictors. The data leave weighting, aggregation, boundary, clipping, and optimizer-history effects unresolved (Appendix Table A17).

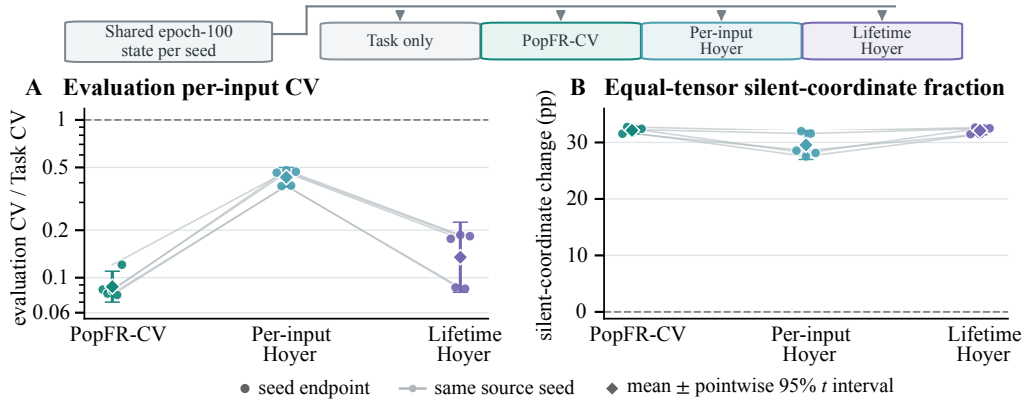

Figure 4: **Related objectives share the qualitative response and reach different endpoints.** Five CIFAR-100 seeds branch from complete epoch-100 states; Hoyer coefficients match only the magnitude of PopFR-CV's initial three-update, task-relative AdamW deviation. **(A)** All regularized branches lower CV; per-input Hoyer ends at $4.94\times$ PopFR-CV's CV. **(B)** All enlarge the equal-tensor silent-coordinate fraction by about 30 points. Gray lines join branches sharing a source; diamonds show paired means with pointwise 95% intervals. Appendix Fig. A4 adds MFR and accuracy.

8 Discussion

The experiments establish population organization as a distinct outcome of activity regularization. Direct rate suppression, entropy-style regularization, and PopFR-CV produce different combinations of accuracy, firing, dispersion, and participation. IM-Loss style lowers CV while recruiting more coordinates; PopFR-CV lowers CV while many more coordinates remain silent. A single firing or dispersion statistic cannot summarize these organizational changes.

This view complements operational sparsification. Activity Pruning and STATA translate selected spike or token sparsity into reduced computation (Bu et al., 2025; Zhuge et al., 2024). Our dense-graph experiments address the preceding learning question: which responses remain active after an objective redirects training? Mean firing quantifies event volume, and participation readouts locate its concentration. Hardware benefit requires an execution mechanism that exploits this structure, so organizational evidence and computational savings remain separate evaluation stages.

The inversion also carries an optimization lesson. PopFR-CV rewards higher per-input CV, yet a common mature state moves rapidly toward lower hard-spike CV. Source-matched Hoyer branches share ideal same-vector ordering and reach different quantitative endpoints. Auxiliary-loss sign and aggregation-axis labels therefore provide incomplete forecasts of learned hard-forward organization. Several contributors remain compatible with the evidence, including state-dependent weighting, clipping, optimizer history, numerical boundaries, and the discrete forward map. Direct trajectory and endpoint measurements provide the reliable test.

The paired endpoint analysis localizes the aggregate CV fall to strong withdrawal from monitored attention-path responses. Every retained non-attention response stays active in both branches, and its group-mean CV remains near Task. This compositional account leaves functional necessity open. A direct intervention that preserves attention-path occupancy during PopFR-CV training offers a falsifiable next test: attenuation of the inversion would support withdrawal as a mediator; relocation would indicate broader network adaptation.

The complete pattern recurs across three static-image datasets in the studied four-timestep SDT. Other systems and event data define the current boundary. The study provides a measurable link between activity objectives and learned participation structure. Its practical conclusion is to report task performance, total firing, within-response dispersion, and dataset-level participation together. This joint view exposes reorganization hidden by mean firing and tests whether training realizes the behavior implied by an objective.

AI USE STATEMENT

We used OpenAI Codex to debug experiment and analysis code; cross-check numerical summaries; check mathematical statements; and edit the manuscript. The authors selected the protocols, executed the reported experiments, and reviewed the code, raw outputs, statistical summaries, derivations, references, and all AI-assisted text. No reported experimental data were generated by a generative AI model. We take responsibility for the final content, including all AI-assisted text and artifacts.

REPRODUCIBILITY STATEMENT

Section 3 specifies the losses, metrics, training setup, statistical reporting, validation-only coefficient selection, and source-training-only calibration fixed before inspection of the new Hoyer endpoints. Appendix B provides the mathematical derivations, while Appendices C–H report outcomes by protocol, cross-system experiments, diagnostics, hardware-cost assumptions, and provenance. The anonymized supplementary release contains training and aggregation code, protocol manifests, calibration records, selected machine-readable run records, and derived summaries. Appendix H and the release README distinguish direct reaggregation from summary-only evidence and experiments that require retraining.

REFERENCES

Nitin Bansal, Xiaohan Chen, and Zhangyang Wang. Can we gain more from orthogonality regularizations in training deep networks? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Tong Bu, Xinyu Shi, and Zhaofei Yu. Activity pruning for efficient spiking neural networks. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5700. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/fa12d67b5939c37ea8a4659c31a08d2c-Abstract-Conference.html.

Michael Cogswell, Faruk Ahmed, Ross Girshick, Larry Zitnick, and Dhruv Batra. Reducing overfitting in deep networks by decorrelating representations. In *International Conference on Learning Representations (ICLR)*, 2016.

Benjamin Cramer, Sebastian Billaudelle, Simeon Kanya, Aron Leibfried, Andreas Grübl, Vitali Karasenko, Christian Pehle, Korbinian Schreiber, Yannik Stradmann, Johannes Weis, Johannes Schemmel, and Friedemann Zenke. Surrogate gradients for analog neuromorphic computing. *Proceedings of the National Academy of Sciences*, 119(4):e2109194119, 2022.

Shikuang Deng, Yuhang Li, Shanghang Zhang, and Shi Gu. Temporal efficient training of spiking neural network via gradient re-weighting. In *International Conference on Learning Representations (ICLR)*, 2022.

Utku Evci, Trevor Gale, Jacob Menick, Pablo Samuel Castro, and Erich Elsen. Rigging the lottery: Making all tickets winners. In *International Conference on Machine Learning (ICML)*, 2020.

Wei Fang, Zhaofei Yu, Yanqi Chen, Tiejun Huang, Timothée Masquelier, and Yonghong Tian. Deep residual learning in spiking neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 21056–21069, 2021.

Yufei Guo, Yuanpei Chen, Liwen Zhang, Xiaode Liu, Yinglei Wang, Xuhui Huang, and Zhe Ma. IM-Loss: Information maximization loss for spiking neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Song Han, Jeff Pool, John Tran, and William J. Dally. Learning both weights and connections for efficient neural network. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Mark Horowitz. Computing's energy problem (and what we can do about it). In *2014 IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 10–14, 2014. doi: 10.1109/ISSCC.2014.6757323.

Patrik O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004.

Decebal Constantin Mocanu, Elena Mocanu, Peter Stone, Phuong H. Nguyen, Madeleine Gibescu, and Antonio Liotta. Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science. *Nature Communications*, 9(1):2383, 2018.

Simon Narduzzi, Siavash A. Bigdeli, Shih-Chii Liu, and L. Andrea Dunbar. Optimizing the consumption of spiking neural networks with activity regularization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022. doi: 10.1109/ICASSP43922.2022.9746375.

Emre O. Neftci, Hesham Mostafa, and Friedemann Zenke. Surrogate gradient learning in spiking neural networks. *IEEE Signal Processing Magazine*, 36(6):51–63, 2019.

Thomas Pellegrini, Romain Zimmer, and Timothée Masquelier. Low-activity supervised convolutional spiking neural networks applied to speech commands recognition. In *IEEE Spoken Language Technology Workshop (SLT)*, pp. 97–103, 2021. doi: 10.1109/SLT48900.2021.9383587.

Alessandro Treves and Edmund T. Rolls. What determines the capacity of autoassociative memories in the brain? *Network: Computation in Neural Systems*, 2(4):371–397, 1991. doi: 10.1088/0954-898X_2_4_004.

William E. Vinje and Jack L. Gallant. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456):1273–1276, 2000. doi: 10.1126/science.287.5456.1273.

Ben Willmore and David J. Tolhurst. Characterizing the sparseness of neural codes. *Network: Computation in Neural Systems*, 12(3):255–270, 2001. doi: 10.1088/0954-898X/12/3/302.

Huanrui Yang, Wei Wen, and Hai Li. DeepHoyer: Learning sparser neural network with differentiable scale-invariant sparsity measures. In *International Conference on Learning Representations (ICLR)*, 2020.

Man Yao, Jiakui Hu, Zhaokun Zhou, Li Yuan, Yonghong Tian, Bo Xu, and Guoqi Li. Spike-driven transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yuichi Yoshida and Takeru Miyato. Spectral norm regularization for improving the generalizability of deep learning. *arXiv preprint arXiv:1705.10941*, 2017.

Zhengyang Zhuge, Peisong Wang, Xingting Yao, and Jian Cheng. Towards efficient spiking transformer: a token sparsification framework for training and inference acceleration. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62768–62778. PMLR, 2024. URL `https://proceedings.mlr.press/v235/zhuge24b.html`.

A IMPLEMENTATION DETAILS AND EXACT DEFINITIONS

Table A1 records the settings that determine the implemented objectives and their readouts. The shared-state continuation includes a complete software and hardware record; the earlier fixed-coefficient endpoints do not contain the same environment capture.

Table A1: Implementation record. Tensor dimensions exclude the time and batch axes and refer to the primary CIFAR SDT setting.

Item	Executed setting
Implemented POPFR-CV	For each hook $[T, B, \ldots]$, cast hard spikes to FP32, average over time, flatten all coordinate axes, and compute per-input $C_\epsilon = (\sigma_N + 10^{-8})/(\mu + 10^{-8})$ with population SD (divisor N). Average inputs within a tensor and then tensors equally; the loss is $-\lambda C_\epsilon$.
Primary SDT neurons	SpikingJelly `MultiStepLIFNode`, $\tau = 2$, numeric reset 0, `decay_input=True`, `detach_reset=True`, torch backend, hard binary forward, Sigmoid surrogate $\alpha = 4$. Threshold is 1 except the two executed talking-head LIF outputs (one per block), which use 0.5. `reset_net` is called after every train/evaluation minibatch.
Monitored SDT tensors	Stem: `MS_SPS_lif` $[32, 32, 32]$, `lif1` $[64, 32, 32]$, `lif2` $[128, 32, 32]$, `lif3` $[256, 8, 8]$. In each block $b \in \{0, 1\}$: `first_lif`, `q_lif`, `k_lif`, `v_lif` $[256, 8, 8]$; `kv_before`$= k \odot v$ $[8, 64, 32]$; talking-head `kv` LIF $[8, 1, 32]$; `x_after_qkv`$= q \odot$ `kv` $[8, 64, 32]$; `fc1_lif` $[256, 8, 8]$; `fc2_lif` $[1024, 8, 8]$. The head LIF is $[256]$. Thus 19 tensors are explicit LIF outputs and four are binary products.
Precision and finite checks	Fixed-comparison SDT, its continuation, and both systems' one-step probes run in FP32; CV reductions explicitly use FP32. Native SEW trains with CUDA AMP and dynamic scaling and casts hooks to FP32 for CV; its evaluation is outside autocast. Training aborts on nonfinite total loss, and continuation validation rejects nonfinite primary metrics. No replacement of nonfinite headline CV is applied.
Input and optimization	CIFAR input/patch $32/4$; Tiny-ImageNet $64/8$; direct image repeat for T steps. Random crop (padding 4 or 8), horizontal flip, RandAugment, Mixup $\alpha = 0.5$ (probability 1), label smoothing 0.1, AdamW learning rate 3×10^{-4}, minimum 10^{-5}, weight decay .06, cosine decay, and global ℓ_2 gradient clip 1.
Onset and warmup	The fixed comparison applies auxiliary losses after 20 completed task-only epochs; the first treated update is zero-based index 20 and produces completed epoch 21. This gate is distinct from the 20-epoch learning-rate warmup. The shared-state continuation branches after completed epoch 100, so the first treated result is epoch 101.
Evaluation averaging	Fixed-coefficient and validation-selected headline MFR and population CV are equal means of per-batch tensor means. Therefore an incomplete final batch (16 of 128 for 10,000 samples; 8 of 128 for 5,000) has one full batch weight. Accuracy and CE are sample weighted; the full-evaluation firing sums used to form coordinate silence are also sample weighted. The headline silent ratio averages within-tensor fractions with rate $< 10^{-6}$ equally, while the coordinate-weighted ratio is separate. The shared-state SDT continuation evaluates 5,000 samples in 50 complete batches of 100, so its MFR and CV are exact uniform sample means. Native SEW accumulates per-sample CV sums.
Zero and conditional cases	Because 10^{-8} is added to numerator and denominator, $C_\epsilon(0) = 1$. Conditioning excludes whole-zero vectors; it preserves zero coordinates within a nonzero vector. The conditional statistic is undefined if every input–tensor vector is zero. Every reported SDT decomposition contains nonzero vectors.
Continuation state	The completed-epoch-100 snapshot stores model, optimizer, scheduler, data-order state, Python/NumPy/Torch/CUDA RNG, and scaler state (none for FP32 SDT; a mapping for native SEW). Restoration checks state hashes; paired branches use matched epoch-derived data generators and preserve RNG across validation.
Matched-axis continuation	Five separate CIFAR-100 SDT sources (seeds 301–305) restore complete epoch-100 model, AdamW, schedule, and RNG state into Task, POPFR-CV, per-input Hoyer, and minibatch-lifetime Hoyer branches. From three deterministic source-training minibatches, coefficients match the norm of each Hoyer branch's three-update deviation from Task, normalized by the Task displacement from source, to the POPFR-CV target within a fixed 15% tolerance. Direction and later evolution remain unmatched. Exact-zero Hoyer vectors share the assigned forward value one with $C_\epsilon(0)$; their boundary derivatives differ. Source and epochwise training-stream hashes are checked within every seed.
Native SEW system	CIFAR-adapted SEW-ResNet-18 with ADD residuals, 3×3 stride-1 stem, no max pool, direct $T = 4$ coding, 20 explicit LIF outputs, and nonbinary ADD residuals excluded. Neuron and optimizer settings match nominal SDT choices; native SEW retains AMP, a different activation inventory, and a different effective objective scale.
Software/source identifiers	The shared-state continuation records Python 3.10.20, PyTorch 2.3.0+cu121, torchvision 0.18.0+cu121, timm 0.6.12, NumPy 2.2.6, CUDA 12.1, cuDNN 8902, SpikingJelly 0.0.0.0.12, and RTX 3090. Versioned protocol manifests and source hashes for the continuation, one-step analysis, and endpoint decomposition accompany the release. Earlier fixed-result JSONs do not contain an equivalent full runtime capture.

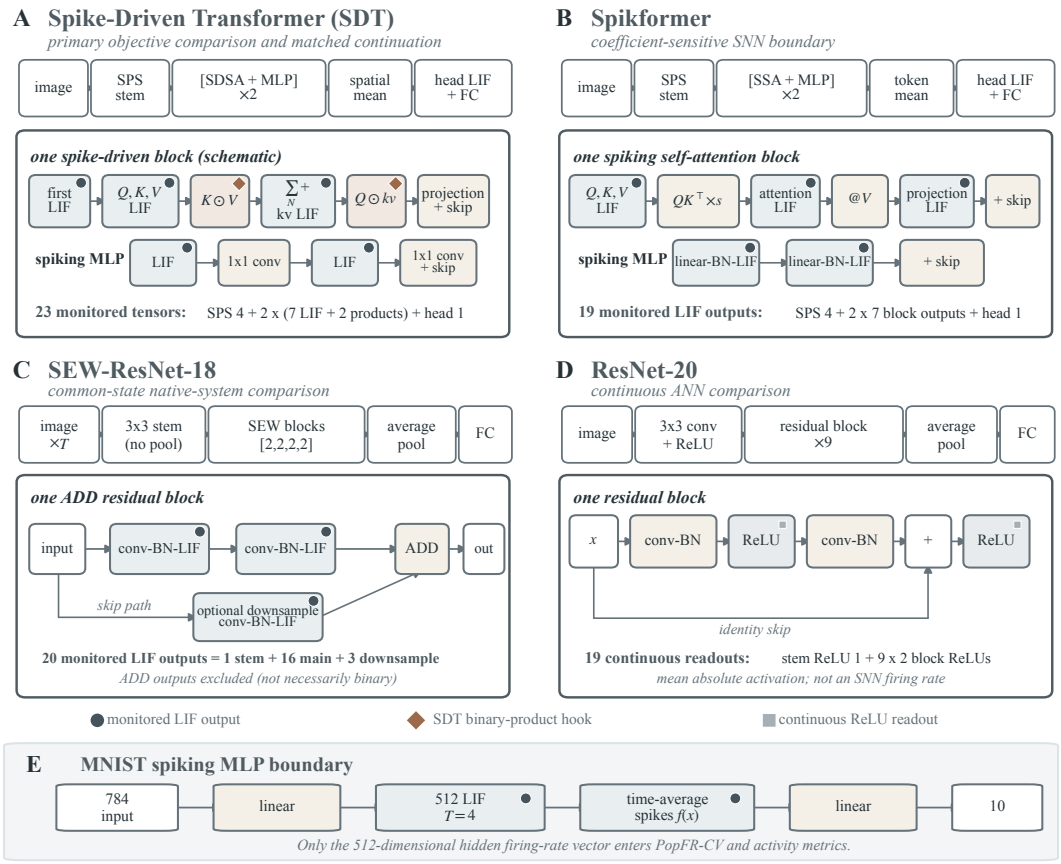

Figure A1: **Architectures and activity readouts used to test scope.** **(A)** The two-block SDT is the primary system. Its 23 monitored tensors comprise 19 explicit LIF outputs and four binary spike products. **(B)** Spikformer provides a second spiking-Transformer setting with 19 monitored spike outputs. **(C)** Native SEW-ResNet-18 provides the common-state residual-SNN comparison; 20 explicit LIF outputs are monitored and nonbinary residual sums are excluded. **(D)** ResNet-20 is an ANN control with 19 post-ReLU readouts; mean absolute activation is continuous and is not called firing rate. The footer shows the separate MNIST spiking MLP and its 512-dimensional hidden LIF readout. Markers illustrate representative locations; labels report the complete inventories. Differences in dynamics, precision, and effective objective scale make these complete-system scope tests. Architecture remains unisolated.

A.1 IMPLEMENTED COMPARATOR OBJECTIVES

All spike-based objectives use the same 23 monitored tensors listed in Table A1 (19 explicit LIF outputs and four binary spike products), flatten every remaining coordinate axis, and average tensors equally. For monitored spike tensor ℓ, let N_ℓ be its number of flattened activation coordinates and define the coordinate-wise minibatch-lifetime firing vector and its mean by

$$p_{\ell i} = \frac{1}{TB}\sum_{t=1}^{T}\sum_{b=1}^{B} s_{t,\ell,i}(x_b), \qquad m_\ell = \frac{1}{N_\ell}\sum_{i=1}^{N_\ell} p_{\ell i}. \tag{5}$$

All reported activity losses average monitored tensors equally. The implemented mean-firing penalty is

$$\mathcal{L}_{\mathrm{MFR}} = \lambda_{\mathrm{MFR}}\frac{1}{L}\sum_{\ell=1}^{L} m_\ell, \qquad \lambda_{\mathrm{MFR}} = 0.5. \tag{6}$$

Because spikes are nonnegative, this is the ℓ_1 penalty on mean activity used for the MFR-ℓ_1 condition.

The minibatch-lifetime Hoyer comparator first forms $p_\ell = (p_{\ell 1}, \ldots, p_{\ell N_\ell})$ and computes

$$H_\ell = \frac{\sqrt{N_\ell} - \|p_\ell\|_1/(\|p_\ell\|_2 + \epsilon_H)}{\sqrt{N_\ell} - 1}, \qquad \mathcal{L}_{\mathrm{Hoyer}} = -\lambda_H \frac{1}{L}\sum_{\ell=1}^{L} H_\ell, \tag{7}$$

with $\epsilon_H = 10^{-8}$ and $\lambda_H = 0.05$. Thus minimization maximizes minibatch-lifetime Hoyer sparsity. No clipping is applied; if $p_\ell = 0$, this stabilized expression is $\sqrt{N_\ell}/(\sqrt{N_\ell} - 1)$. This is not the matched per-input Hoyer diagnostic reported later in Appendix E.

The matched-axis continuation uses standard normalized Hoyer sparsity on nonzero vectors, either on each per-input vector $f_\ell(x)$ before averaging inputs or on the minibatch-lifetime vector p_ℓ. Because firing rates are nonnegative, the executed numerator is `vectors.sum(dim=1)`; no absolute-value or ℓ_1-norm operator is used. Consequently, an exact zero coordinate inside a nonzero vector receives the direct-sum derivative. A sparse-vector automatic-differentiation test verifies collinearity with ideal CV when $\epsilon = 0$. In this experiment only, an entire exact-zero vector is explicitly assigned Hoyer value one through a constant-valued `torch.where` branch, giving zero direct derivative with respect to that rate vector. Hoyer and C_ϵ therefore share the assigned forward value at whole-vector zero; their boundary derivatives differ. The per-source Hoyer coefficients are source-specific calibration controls selected by the procedure in Sec. 3. They are distinct from the exploratory fixed $\lambda_H = 0.05$ cohort.

The entropy comparator clips each lifetime rate as $\widetilde{p}_{\ell i} = \mathrm{clip}(p_{\ell i}, \delta, 1-\delta)$, where $\delta = 10^{-6}$, and uses natural-log Bernoulli entropy

$$h(p) = -p\log p - (1-p)\log(1-p), \qquad \mathcal{L}_{\text{IM-style}} = -\lambda_{\mathrm{IM}}\frac{1}{L}\sum_{\ell=1}^{L}\frac{1}{N_\ell}\sum_{i=1}^{N_\ell} h(\widetilde{p}_{\ell i}), \tag{8}$$

with $\lambda_{\mathrm{IM}} = 0.5$. It therefore favors lifetime firing probabilities near $1/2$. The label *IM-Loss-style* identifies this simplified Bernoulli-entropy term; the published IM-Loss method contains additional components.

For WEIGHTCV, each output filter or output unit r of a `Conv2d` or `Linear` module k is flattened after taking absolute values, $a_{kr} = \mathrm{vec}(|W_{k,r}|)$. The implementation computes

$$C^w_{kr} = \frac{\mathrm{std}_N(a_{kr}) + \epsilon_w}{\mathrm{mean}(a_{kr}) + \epsilon_w}, \qquad \mathcal{L}_{\mathrm{WeightCV}} = -\lambda_w\frac{1}{K}\sum_{k=1}^{K}\frac{1}{R_k}\sum_{r=1}^{R_k} C^w_{kr}, \tag{9}$$

with population standard deviation, $\epsilon_w = 10^{-8}$, and $\lambda_w = 10^{-3}$. Biases, BatchNorm and LIF parameters, and the talking-head `Conv1d` weights are excluded. The combined WEIGHT+POPCV condition adds this term and Eq. (3); the fixed-coefficient POPFR-CV term uses $\lambda_{\mathrm{pop}} = 5\times 10^{-4}$, whereas the separate validation-selected confirmation uses 2×10^{-4}. In the reported 300-epoch fixed-comparison cohort, every auxiliary term first acts on zero-based epoch 20, producing completed epoch 21. The Hoyer and IM-Loss-style coefficients were selected in exploratory sweeps and are reported as tested operating points, not as outcomes of a common validation budget.

B EXACT RELATIONS FOR IDEAL CV AND HOYER

Let $f \in \mathbb{R}^n_{\geq 0}$, $n > 1$, have $\mu = \mathrm{mean}(f) > 0$, population variance $\sigma^2 = n^{-1}\sum_i (f_i - \mu)^2$, and ideal $\mathrm{CV} = \sigma/\mu$ with $\epsilon = 0$. Then

$$1 + \mathrm{CV}^2 = \frac{n\|f\|_2^2}{\|f\|_1^2}, \qquad \frac{\|f\|_1}{\|f\|_2} = \frac{\sqrt{n}}{\sqrt{1+\mathrm{CV}^2}}. \tag{10}$$

Equivalently, defining $N_{\mathrm{eff}} = \|f\|_1^2/\|f\|_2^2$ gives

$$N_{\mathrm{eff}}(f) = \frac{\|f\|_1^2}{\|f\|_2^2}, \qquad \frac{N_{\mathrm{eff}}(f)}{n} = \frac{1}{1+\mathrm{CV}(f)^2}. \tag{11}$$

This is a graded participation measure. It satisfies $N_{\mathrm{eff}} \leq |\,\mathrm{supp}(f)|$, with equality only when the nonzero entries are equal. The Treves–Rolls activity ratio uses the same ideal normalized quantity,

$$a_{\mathrm{TR}}(f) = \frac{\mathrm{mean}(f)^2}{\mathrm{mean}(f^2)} = \frac{N_{\mathrm{eff}}(f)}{n} = \frac{1}{1+\mathrm{CV}(f)^2}, \tag{12}$$

where smaller a_{TR} denotes a sparser response (Treves & Rolls, 1991; Willmore & Tolhurst, 2001). A common high-is-sparse finite-n normalization is

$$S_{\mathrm{TR}}(f) = \frac{1 - a_{\mathrm{TR}}(f)}{1 - 1/n} = \frac{n\,\mathrm{CV}(f)^2}{(n-1)[1+\mathrm{CV}(f)^2]}. \tag{13}$$

This finite-n normalization is widely used for high-is-sparse reporting (Vinje & Gallant, 2000). These are same-vector parameterizations; they do not equate statistics formed after different sample, time, or tensor aggregations. Substitution into normalized Hoyer sparsity, $H = (\sqrt{n} - \|f\|_1/\|f\|_2)/(\sqrt{n}-1)$, gives Eq. (4). Writing its right-hand side as $h(\mathrm{CV})$,

$$h'(c) = \frac{\sqrt{n}}{\sqrt{n}-1}\,\frac{c}{(1+c^2)^{3/2}} > 0 \quad (c > 0), \tag{14}$$

$\nabla_f H = h'(\mathrm{CV})\nabla_f \mathrm{CV}$ at strictly positive, nonuniform interior points. The executed nonnegative-rate implementation extends this relation to nonzero, nonuniform vectors containing zero coordinates by directly computing $\sum_i f_i$. This remains a same-vector result. Aggregation and the state-dependent weights $h'(c_x)$ distinguish $\mathbb{E}_x[H(f(x))]$ from $H(\mathbb{E}_x[f(x)])$ and reweight summed per-input gradients.

For completeness, when $\mu > 0$ and $\sigma > 0$, direct differentiation of $\mathrm{CV} = \sigma/\mu$ yields

$$\frac{\partial \mathrm{CV}}{\partial f_i} = \frac{f_i - \mu}{n\mu\sigma} - \frac{\sigma}{n\mu^2}. \tag{15}$$

Using $\sum_i (f_i - \mu) = 0$ and $\sum_i (f_i - \mu)^2 = n\sigma^2$ gives

$$\sum_i \left(\frac{\partial \mathrm{CV}}{\partial f_i}\right)^2 = \frac{1+\mathrm{CV}^2}{n\mu^2}, \tag{16}$$

and therefore

$$\|\nabla_f \mathrm{CV}\|_2 = \frac{\sqrt{1+\mathrm{CV}^2}}{\sqrt{n}\,\mu}. \tag{17}$$

This identity characterizes local sensitivity for an observed vector. Together with Eq. (4), it also shows that the ideal CV and Hoyer objectives do not differ through an intrinsic same-vector gradient direction.

The implemented loss uses $(\sigma + \epsilon)/(\mu + \epsilon)$ with $\epsilon = 10^{-8}$, and the Hoyer implementation also regularizes its norm denominator. The exact relations above therefore describe the ideal positive-mean definition, while stabilization affects zero and near-zero behavior. In particular, an all-zero input–tensor vector has $C_\epsilon(0) = 1$. If p_0 is the fraction of whole input–tensor vectors that are exactly zero under the same equal-tensor measure used for evaluation, then the identity

$$\mathbb{E}[C_\epsilon] = p_0 + (1 - p_0)\,\mathbb{E}[C_\epsilon \mid \mu(f) > 0] \tag{18}$$

separates the whole-zero contribution from the conditional nonzero contribution. Here p_0 denotes whole-response occupancy; the lifetime silent-coordinate ratio is a separate statistic. Appendix Table A9 reports both terms at the common source and paired endpoints. The identity supplies an exact measurement decomposition and carries no causal attribution.

C ADDITIONAL OUTCOME ANALYSES

C.1 WEIGHT- AND POPULATION-AXIS ABLATION

Table A2 reports the complete fixed-coefficient ablation from the matched SDT cohort. The weight-only objective remains close to Task training on all four endpoint summaries, whereas the combined weight-plus-population objective remains close to the population-only objective. Within this tested design, the population term accounts for the characteristic low-CV, enlarged-silent-set endpoint, and POPFR-CV is therefore the more parsimonious recipe. This ablation does not rank the separate activity-objective families reported below, which were not all selected under a common tuning protocol.

Table A2: Exact weight- and population-axis ablation in SDT at $T = 4$ and 300 epochs (identical seeds 0–4; mean ± sample SD). Auxiliary terms start after 20 completed task-only epochs.

Data	Condition	Accuracy (%)	MFR	Eval. pop. CV	Silent (%)
C10	Task	87.78±.39	.0585±.0006	19.15±.71	22.05±.56
	WEIGHTCV	87.68±.21	.0582±.0013	18.37±.71	22.51±.69
	POPFR-CV	88.40±.19	.0531±.0007	1.664±.012	52.81±.14
	WEIGHT+POPCV	88.42±.27	.0518±.0011	1.670±.018	52.89±.17
C100	Task	65.79±.31	.0734±.0008	18.26±.55	19.07±.58
	WEIGHTCV	65.94±.37	.0735±.0003	18.72±.62	18.56±.24
	POPFR-CV	66.88±.59	.0695±.0009	1.547±.008	52.54±.04
	WEIGHT+POPCV	66.84±.23	.0703±.0012	1.558±.040	51.64±1.53

C.2 PAIRED ACCURACY DIFFERENCES

The matched $n = 5$ experiment uses identical seeds 0–4 across configurations and does not retune coefficients per seed. Relative to the paired baseline, the accuracy differences (mean ± sample SD, percentage points) are:

Configuration	CIFAR-10	CIFAR-100
POPFR-CV	$+0.620 \pm 0.217$	$+1.096 \pm 0.750$
WEIGHT+POPCV	$+0.644 \pm 0.422$	$+1.050 \pm 0.400$

For both configurations, all five paired differences are positive on each dataset. With $n = 5$, the smallest two-sided exact Wilcoxon value for a perfectly consistent sign is 0.0625; we therefore avoid a binary significance claim and report the paired magnitudes and directions.

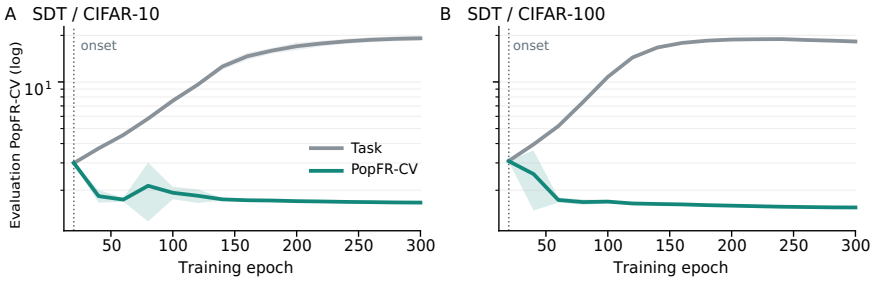

Figure A2: **The evaluation-CV fall repeats across fixed-coefficient SDT runs.** Trajectories are shown for **(A)** CIFAR-10 and **(B)** CIFAR-100. Curves and bands are mean ± sample SD over the same five seeds at each recorded epoch; the vertical line marks the first POPFR-CV-treated epoch after 20 completed task-only epochs. The log-scale curves are evaluation readouts, not the minibatch training objective. Because these runs monitored the evaluation split during training, they are descriptive replications; the independently held-out confirmation is reported in Table A6.

Table A3: Exploratory SDT comparator endpoints at 300 epochs ($n = 5$; mean $\pm$ sample SD). The CIFAR-100 lifetime-Hoyer comparator cohort has one chance-level collapse, so no post-hoc stable-only aggregate or near-zero-MFR ICE is reported.

Data	Objective	Accuracy (%)	MFR	Eval. CV	Silent (%)	ICE
C10	Lifetime Hoyer	$87.87 \pm .17$	$.05775 \pm .00031$	15.12 ± 2.16	$26.76 \pm .66$	1521 ± 6
C10	IM-Loss style	$89.25 \pm .31$	$.18140 \pm .00219$	$4.060 \pm .254$	$8.51 \pm .48$	492 ± 6
C100	IM-Loss style	$66.95 \pm .13$	$.09802 \pm .00046$	$9.736 \pm .380$	$13.91 \pm .27$	683 ± 3
C100	Lifetime Hoyer	1/5 chance-level collapse; aggregate omitted				

C.3 Exploratory lifetime-Hoyer and IM-Loss-style endpoints

Table A3 reports the five-seed comparator endpoints used in Figs. 1 and 2. The coefficients were $\lambda_H = 5 \times 10^{-2}$ and $\lambda_{\mathrm{IM}} = 5 \times 10^{-1}$, both starting at epoch 20. They were chosen in preliminary sweeps and are analyzed separately from the matched fixed-coefficient comparison.

The four stable CIFAR-100 lifetime-Hoyer accuracies range from 65.68 to 66.52%; the fifth run reaches 1.00% accuracy with zero measured MFR. Reporting only the stable runs would conceal the observed instability. This comparator cohort and the onset grid in Appendix E.6 are independent full-training launches. Both start from scratch. Their overlapping nominal condition is discussed there. IM-Loss style lowers evaluation CV on both datasets and decreases the silent-coordinate fraction, separating low CV from the joint low-CV/enlarged-silent-set endpoint reached under PopFR-CV.

Table A4: All five outcomes in the CIFAR-100 lifetime-Hoyer comparator cohort ($\lambda_H = .05$, onset after 20 task-only epochs). Finite metrics are shown for every seed; ICE is undefined for the collapsed zero-MFR run and no stable-only aggregate is computed. Equal-tensor and coordinate-weighted silence are distinct summaries.

Seed	Accuracy (%)	MFR	Eval. CV	Equal-tensor silent (%)	Coord.-weighted silent (%)
0	65.68	.074176	15.3373	20.114	12.411
1	66.52	.074082	16.0782	20.079	12.265
2	1.00	0	1.0000	100.000	100.000
3	66.30	.072477	17.4287	20.224	12.386
4	65.76	.072414	13.9899	18.999	11.682

C.4 Tiny-ImageNet extension

Table A5: SDT/Tiny-ImageNet, $T = 4$, 300 epochs, five matched seeds. The paired changes are $+0.580 \pm 0.372$ accuracy points and $+22.693 \pm 0.808$ silent-ratio points; both are positive in all five pairs. This is a dataset-scale extension within SDT.

Condition	Accuracy (%)	MFR	Eval. pop. CV	Silent (%)
Baseline	45.83 ± 0.35	0.0610	29.12	30.721 ± 0.508
PopFR-CV	46.41 ± 0.36	0.0575	1.70	53.414 ± 0.318

C.5 Validation-only coefficient selection

The validation-selected experiment contains 15 tuning runs (three baseline seeds plus four PopFR-CV coefficients across the same three seeds) and six independent official-test runs. The split, coefficient-selection rule, and test seed list were fixed before official-test evaluation. The rule selects 2×10^{-4}; the manifest distributed with the release records this procedure without machine-specific paths. This experiment is analyzed separately from the fixed-coefficient and Tiny-ImageNet comparisons.

Table A6: Independent CIFAR-10 official-test confirmation after validation-only coefficient selection. Values are mean ± sample SD over fresh seeds 101–103; ICE is reported only as an activity-efficiency proxy.

Condition	Accuracy (%)	MFR	Eval. pop. CV	Silent (%)	ICE
Baseline	$87.453 \pm .159$	$.05745 \pm .00063$	$18.595 \pm .701$	$19.891 \pm .453$	1522.4 ± 17.0
PopFR-CV	$88.510 \pm .261$	$.05441 \pm .00154$	$1.657 \pm .009$	$52.674 \pm .083$	1627.4 ± 44.3

C.6 Silent-ratio aggregation and threshold robustness

Table A7 recomputes each paired silent-set change using both the primary equal-tensor mean and a coordinate-weighted mean. Every paired seed has a positive change under both aggregations. Equal-tensor weighting affects the magnitude and preserves the expansion.

Table A7: Silent-ratio sensitivity (paired mean change ± sample SD, percentage points). The four protocols remain separate.

Protocol	n	Equal-tensor	Coordinate-weighted
CIFAR-10, fixed coefficient	5	$+30.760 \pm 0.575$	$+14.255 \pm 0.426$
CIFAR-100, fixed coefficient	5	$+33.474 \pm 0.587$	$+15.865 \pm 0.410$
Tiny-ImageNet, fixed coefficient	5	$+22.693 \pm 0.808$	$+9.689 \pm 0.649$
CIFAR-10, validation-selected test	3	$+32.783 \pm 0.392$	$+15.389 \pm 0.134$

The 10^{-6} threshold is insensitive within the relevant range because of discrete resolution. At $T = 4$, the smallest nonzero lifetime rate over 5,000 or 10,000 examples is 5×10^{-5} or 2.5×10^{-5}. Hence every threshold in $(0, 2.5 \times 10^{-5})$ gives the same zero-observed-spike mask in these static-image evaluations; no empirical threshold sweep is needed.

C.7 Shared-checkpoint continuation detail

The controlled continuation uses fresh seeds 201–205 and a fixed CIFAR-100 train/validation split. Table A8 shows that CV changes within the first measured epoch after branching, whereas accuracy recovers over the remaining training trajectory. Epochs are repeated measurements, not replicates.

Table A8: SDT shared-checkpoint continuation. CV is the geometric Task+PopFR-CV/Task validation-CV ratio; accuracy is the paired mean difference in percentage points.

Epoch	CV ratio	Δaccuracy (pp)
100	1.000	.000
101	.230	-20.288
105	.186	-9.396
110	.174	-5.004
120	.136	-5.480
160	.103	-1.404
200	.086	$-.396$
250	.098	$+.304$
300	.091	$+.276$

At epoch 300, the SDT log-CV response is -2.395 ± 0.251 (geometric ratio 0.0911; 5/5 lower); SEW gives -0.005 ± 0.016 (ratio 0.9950; 2/5 lower). Their matched native-system response ratio is 0.0916. The SDT equal-tensor and coordinate-weighted silence changes are $+31.944 \pm 0.457$ and $+14.889 \pm 0.281$ points; the SEW equal-tensor change is -0.32 points. In a secondary analysis specified after training, the sample-weighted mean of sequential minibatch CV logged during SDT epoch 300 has a Task+PopFR-CV/Task ratio of 0.1039 and is lower in 5/5 seeds. Because parameters change between minibatches, this quantity is a descriptive training-side check and supplies no static evaluation or additional replicate.

A post hoc comparison computed from stored paired endpoints shows a strongly asymmetric membership difference at the same within-tensor indices. Under equal-tensor aggregation, $32.039 \pm 0.370\%$ of indices are Task-active/Task+POPFR-CV-silent; $0.095 \pm 0.091\%$ show the reverse membership. Coordinate weighting gives $15.063 \pm 0.169\%$ versus $0.174 \pm 0.117\%$, with active-set Jaccard $82.59 \pm 0.06\%$. The four monitored spike-product hooks do not account for the full response: among the 19 explicit LIF outputs, the geometric Task+POPFR-CV/Task CV ratio remains 0.1315. The response varies across tensors. These nested readouts provide no causal-layer identification, and the replicate count remains five seeds.

C.8 WHOLE-INPUT–TENSOR ZERO OCCUPANCY AND CONDITIONAL ENDPOINTS

Let $Z = p_0$ be the equal-tensor fraction of whole sample–tensor firing vectors that are exactly zero, $C_+ = \mathbb{E}[C_\epsilon \mid \mu(f) > 0]$, and $M_+ = \mathbb{E}[\mu(f) \mid \mu(f) > 0]$. This Z measures whole-response occupancy; lifetime silent-coordinate ratio measures a separate dataset-level property. Conditioning excludes whole-zero vectors and preserves zero coordinates within a nonzero vector. The Task and Task+POPFR-CV conditional sets can therefore contain different input–tensor events and do not define a fixed survivor group. Table A9 is computed from stored validation activations of the same 23-tensor, seeds-201–205 SDT continuation; it performs no training or optimizer step and does not access the official test set.

Table A9: Whole-input–tensor occupancy decomposition (mean ± sample SD over five seeds). All means use the same equal-tensor measure; the common state is completed epoch 100 and the two endpoints are epoch 300.

A. CV occupancy mixture

State	Z	C_+	Overall C_ϵ
Common source	$.000257 \pm .000080$	$9.0957 \pm .0652$	$9.0936 \pm .0653$
Task endpoint	$.117405 \pm .006388$	$21.3129 \pm .6734$	$18.9286 \pm .6278$
Task+POPFR-CV endpoint	$.516355 \pm .012015$	2.5823 ± 1.0207	$1.7751 \pm .5291$
Paired endpoint change	$+.398950 \pm .016438$	-18.7306 ± 1.0380	$-17.1535 \pm .7491$

B. MFR occupancy mixture

State	M_+	Overall MFR
Common source	$.074869 \pm .000819$	$.074850 \pm .000821$
Task endpoint	$.083817 \pm .000562$	$.073974 \pm .000275$
Task+POPFR-CV endpoint	$.149817 \pm .003797$	$.072425 \pm .000806$
Paired endpoint change	$+.066000 \pm .004284$	$-.001548 \pm .000796$

The decomposition is computed within each seed and only then averaged across the five seeds. Let subscript 0 denote Task and 1 denote Task+POPFR-CV, define $\Delta Z = Z_1 - Z_0$, $\Delta C_+ = C_{+,1} - C_{+,0}$, and $\Delta M_+ = M_{+,1} - M_{+,0}$, and let a bar denote the midpoint of the two endpoints, for example $\overline{Z} = (Z_0 + Z_1)/2$. The identities

$$\Delta C_\epsilon = (1 - \overline{C_+})\,\Delta Z + (1 - \overline{Z})\,\Delta C_+, \tag{19}$$

$$\Delta\mathrm{MFR} = -\overline{M_+}\,\Delta Z + (1 - \overline{Z})\,\Delta M_+ \tag{20}$$

follow from $C_\epsilon = Z + (1 - Z)C_+$ and $\mathrm{MFR} = (1 - Z)M_+$. After averaging the seed-level terms, the rounded values are

$$\Delta C_\epsilon = -17.1535 \approx \underbrace{-4.3605}_{\text{occupancy term}} + \underbrace{-12.7930}_{\text{conditional-}C_+\text{ term}}, \tag{21}$$

$$\Delta\mathrm{MFR} = -0.001548 \approx \underbrace{-0.046626}_{\text{occupancy term}} + \underbrace{+0.045077}_{\text{conditional-}M_+\text{ term}}. \tag{22}$$

Equality holds before rounding; the maximum CV-mixture reconstruction error is 7.81×10^{-8}. These identities show that the two branches have different conditional means and that opposing MFR terms nearly cancel. Because the conditioned event sets differ, they do not establish a change within the same surviving responses, assign causal shares, or identify a training mechanism.

To resolve that ambiguity, we form the paired common-nonzero set $\mathcal{I}_{++} = \{(x, \ell) : \mu_{\mathrm{Task}}(x, \ell) > 0,\ \mu_{\mathrm{PopCV}}(x, \ell) > 0\}$ on the same 5,000 validation inputs and 23 tensors for every seed. This post-specified readout performs no training, optimizer step, checkpoint selection, or official-test evaluation. Within each seed, we pool the CV values over all events in $\mathcal{I}_{++}$ separately for the two branches, form the ratio of those two event means, and then geometrically aggregate the five seed-level ratios. Table A10 reports the result. Across all events, $\mathcal{I}_{++}$ covers $48.36 \pm 1.20\%$, and its geometric CV ratio is 1.157 with pointwise 95% interval $[0.784, 1.708]$. The point estimate is above one and the interval permits both a moderate decrease and increase; the paired subset does not reproduce the order-of-magnitude aggregate fall.

The tensor partition explains the compositional change. Every one of the 275,000 non-attention events (11 tensors $\times$ 5,000 inputs $\times$ 5 seeds) is nonzero in both branches, and their five-seed CV ratio is 1.0046 $[1.0015, 1.0077]$. Only 3,096 of 300,000 attention-path events (1.032%) are nonzero in both branches; 3,091 come from block-0 query responses in seed 201 and five from the same tensor in seed 202, with none in the other attention tensors. Together with the attention tensors' large increases in whole-zero responses, this shows that the branch-specific C_+ decrease accompanies a strong change in attention-path participation. Equalization of a fixed survivor group remains unsupported.

Table A10: Paired common-nonzero endpoint readout. The coverage denominator is all aligned input–tensor events in each row; the numerator contains events nonzero in both Task and Task+PopFR-CV. Within each seed, the reported CV ratio divides the mean PopFR-CV CV by the mean Task CV over that same common-nonzero event set; the five ratios are then geometrically aggregated. Intervals are pointwise Student-t. Attention coverage is pooled descriptively because three seeds contain no common-nonzero attention event.

Event set	Common coverage	PopFR-CV/Task CV ratio	Interpretation
All 23 tensors	$48.36 \pm 1.20\%$	1.157 $[0.784, 1.708]$	no comparable aggregate-scale reduction
12 attention-path tensors	1.032% pooled	not aggregated ($2/5$ seeds defined)	selective withdrawal
Other 11 tensors	100%	1.0046 $[1.0015, 1.0077]$	mean CV changes little

The tensor-resolved view in Fig. A3 shows the same structural division. The 12 query/key/value and spike-product tensors spanning the two attention paths are exactly the 12 of 23 summaries with geometric endpoint CV ratio below .9; every one also has a whole-response zero increase above 25 points. The other 11 stem, shortcut, MLP, and head tensors have mean whole-zero change 0, complete common-nonzero coverage, and a common-event mean-CV ratio close to one. Tensors and coordinates are nested descriptive measurements, so this localizes observed activity withdrawal without identifying a causal layer or establishing functional dispensability.

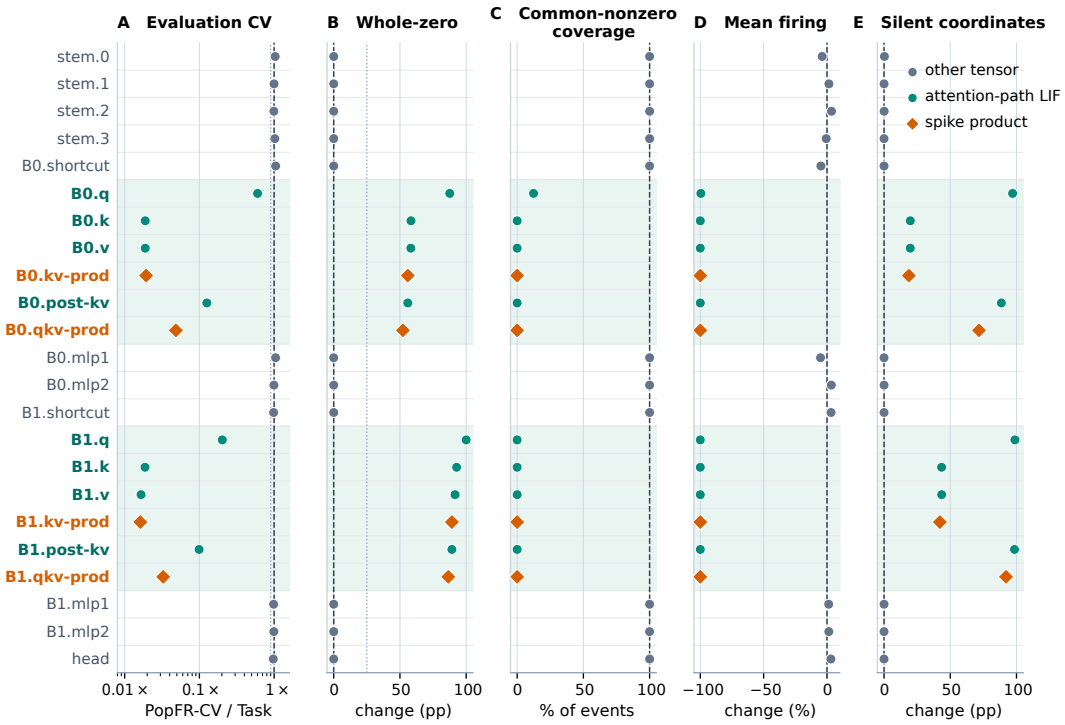

Figure A3: **The endpoint change concentrates in the monitored attention pathways.** Each row is one tensor; shaded rows are the 12 query/key/value or spike-product tensors spanning the two attention paths. Points are descriptive five-seed summaries. **(A)** Geometric Task+PopFR-CV/Task evaluation-CV ratio (log axis; dotted line .9). **(B)** Change in the fraction of whole input–tensor responses that are zero (dotted line 25 points). **(C)** Coverage of aligned events that are nonzero in both branches. **(D–E)** Paired MFR and dataset-level silent-coordinate changes. Teal circles denote attention-path LIF outputs, orange diamonds spike products, and gray circles all other tensors. The figure localizes the observed withdrawal; tensors are nested measurements, not independent replicates, and the analysis does not establish a causal or removable site.

C.9 Objective-component necessity test

The stabilized statistic decomposes exactly as

$$C_\epsilon = \frac{\sigma+\epsilon}{\mu+\epsilon} = \underbrace{\frac{\sigma}{\mu+\epsilon}}_{D_\epsilon} + \underbrace{\frac{\epsilon}{\mu+\epsilon}}_{B_\epsilon}. \tag{23}$$

To test whether the exact-zero-enriching numerator component B_ϵ is necessary for the long-run response, five epoch-100 CIFAR-100 SDT states (seeds 301–305) are continued to epoch 300 under Task, full $-\lambda C_\epsilon$, and dispersion-only $-\lambda D_\epsilon$. The two regularized branches use the same $\lambda = 5 \times 10^{-4}$, 23 monitored tensors, optimizer state, schedule, data order, RNG stream, and validation epochs. All three branches are evaluated with the same canonical validation C_ϵ readout; D_ϵ changes only the training objective. This component-removal comparison leaves initial displacement magnitudes unmatched and does not test B_ϵ alone.

The frozen primary readout is the retained fraction of Task-relative CV reduction,

$$\rho = \frac{1-G_D}{1-G_F}, \tag{24}$$

where G_F and G_D are the geometric endpoint CV ratios to Task. The prospectively fixed "not necessary" criterion requires: integrity-valid records; finite epoch-300 branches with validation accuracy at least 5%; a full response with $G_F \le .90$, lower CV in at least 4/5 seeds, mean accuracy

change at least $-.5$ points, and accuracy change at least -1 point in at least 4/5 seeds; an analogous dispersion response with $G_D \leq .90$; and $\rho \geq .80$. These are operational decision bounds, not null-hypothesis tests. Table A11 gives the endpoint summary. The observed $\rho = 0.9592$ is the pre-specified five-seed sample summary and passes the rule; it is not a percentage of mechanism explained and is not assigned a confidence interval. Dispersion-only remains above full C_ϵ in endpoint CV in 5/5 seeds: their paired geometric ratio is 1.425 with pointwise 95% Student-t interval $[1.196, 1.697]$. Its early transition is also milder: at epoch 101, the geometric Task-relative CV ratios are .212 for full and .761 for dispersion-only. Thus B_ϵ affects the early trajectory. The strong long-run response persists after its removal under this protocol.

Table A11: Matched objective-component continuation ($n = 5$ optimization seeds). Every CV entry uses the same validation C_ϵ readout; "Full" and "Dispersion" identify the training objective. CV entries are geometric endpoint ratios to Task; other entries are paired means $\pm$ sample SD.

Branch	CV ratio	Δaccuracy (pp)	ΔMFR (%)	Δsilent (pp)
Full C_ϵ	.0877	$+.556 \pm .606$	$-1.182 \pm .656$	$+32.146 \pm .495$
Dispersion D_ϵ	.1250	$+.712 \pm .988$	$-1.070 \pm .489$	$+31.999 \pm .477$

The independent unit is the optimization seed; epochs, tensors, coordinates, and validation examples are nested or repeated measurements. The official test set is not constructed or accessed. This result rejects necessity of one implementation-specific component; it does not identify D_ϵ as a unique microscopic mechanism or establish transfer beyond the tested continuation.

D Responses across timesteps, systems, and data

Table A12: Activity responses across the evaluated settings. Recipe identity and measured components are reported separately. Accuracy changes are descriptive; no formal non-inferiority test is performed.

Setting	Recipe; independent n	CV response	Participation	Accuracy	Interpretation
SDT, $T = 4$, static images	PopFR-CV; 5/data set	11.5–17.1× lower	$+22.7$ to $+33.5$ *pp*	$+.58$ to $+1.10$ *pp* observed	Replicated within SDT
Shared C100 state	PopFR-CV; 5/system	ratio .091 SDT; .995 SEW	$+31.94$ *pp* SDT; $-.32$ *pp* SEW	SDT $+.276$ *pp* after -20.288 *pp* transient	Strong SDT response; weak SEW response
SDT, $T = 8$	Weight+PopCV; 3/data set	$10.67 \to 2.21$; $16.28 \to 1.44$	$+41.1/+41.6$ *pp*	$+.50/+.04$ *pp*	Combined-axis response
Spikformer	lower-λ Weight+PopCV; 3 endpoint seeds	lower CV/activity	smaller coordinate reruns only	$-.60/-.14$ *pp*	Partial, coefficient-sensitive
MNIST spiking MLP	PopFR-CV; 5	$1.076 \to 7.547$	$+30.9$ *pp*	$-.51$ *pp*	CV increases; no objective–outcome reversal
CIFAR10-DVS	Weight+PopCV; 2	$1.79 \to 2.30$	0% silence in both	$-.15$ *pp*	Event-data non-replication

D.1 Eight-timestep robustness

The $T = 8$ experiment compares baseline with Weight+PopCV, not population-only CV. Across three paired seeds per dataset, CIFAR-10/CIFAR-100 accuracy changes are $+0.497 \pm 0.390/+0.043 \pm 0.786$ points, evaluation CV changes from 10.67 to 2.21 and from 16.28 to 1.441, and silent-ratio changes are $+41.06 \pm 0.31/+41.63 \pm 0.33$ points. Thus the combined configuration retains the lower-CV, larger-silent-set response at eight timesteps; the accuracy estimates remain descriptive.

D.2 Lower-coefficient Spikformer

At $\lambda_{\text{pop}} = 5 \times 10^{-5}$, the combined-axis Spikformer configuration incurs accuracy costs of $3.53/7.55$ points on CIFAR-10/100 while ICE rises $56.6/79.3\%$. A lower-coefficient experiment uses $\lambda_{\text{pop}} =$

2×10^{-5} on CIFAR-10 and 5×10^{-6} on CIFAR-100. Across three endpoint seeds, the mean accuracy costs fall to $0.60/0.14$ points with ICE gains of $17.6/15.0\%$. Coordinate-level silence is available for smaller, separate runs (CIFAR-10 $n = 2$, CIFAR-100 $n = 1$). These experiments show a coefficient-sensitive accuracy–activity response; they do not provide the matched coordinate analysis used for SDT.

D.3 Non-Transformer spiking MLP

The MNIST experiment uses a $784 \to 512$-LIF$\to 10$ network, $T = 4$, Sigmoid surrogate $\alpha = 4$, and five paired test seeds after validation-only coefficient selection. Table A13 shows a $22\times$ MFR decrease and a large silent-set expansion. Evaluation CV rises and accuracy falls by 0.51 points. The MLP therefore shows a firing-and-participation response. The joint SDT endpoint does not recur.

Table A13: Validation-selected PopFR-CV on a non-Transformer MNIST spiking MLP, five paired test seeds. The activity/silence response appears. The lower-CV/preserved-accuracy joint endpoint does not recur.

Condition	Accuracy (%)	MFR	Eval. pop. CV	Silent (%)
Baseline	97.97 ± 0.07	0.4425	1.076	0.04
PopFR-CV	97.46 ± 0.08	0.0202	7.547	30.9

D.4 Event-based CIFAR10-DVS

The event-data experiment uses SDT, $T = 10$, 200 epochs, and $n = 2$–3 seeds depending on condition. Table A14 reports the key operating points. PopFR-CV and Weight+PopCV reduce activity. Population CV rises and the silent ratio remains zero for every condition. The tested event-data setting therefore produces a different joint response; its origin remains open.

Table A14: SDT/CIFAR10-DVS. The narrower accuracy–activity response remains. The CV-decrease/silent-set endpoint does not reproduce, and the silent ratio is 0% in all rows.

Condition	n	Accuracy (%)	MFR	Eval. pop. CV	ICE
Baseline	2	74.60 ± 1.13	.223	1.79	335
PopFR-CV	2	73.40 ± 1.27	.145	2.42	506
Weight+PopCV	2	$74.45 \pm .64$	.153	2.30	486
Hoyer	3	$73.63 \pm .25$	.121	2.78	610
MFR-ℓ_1	3	73.83 ± 1.56	.056	4.30	1311
IM-Loss style	2	$73.80 \pm .28$	.366	1.02	202

D.5 ANN control

The ResNet-20 control (He et al., 2016) uses ReLU activations and no spike step. Across activation-CV settings, CIFAR-10/100 accuracy remains within 0.5 points and ICE within 5% of baseline. This establishes that the endpoint is not a generic consequence of adding a CV-rewarding term in every network. Because ResNet-20 differs from SDT in architecture, state dynamics, and optimization graph, the comparison cannot isolate the surrogate Jacobian as a unique cause.

E Geometric and local analyses

E.1 Shared-checkpoint one-step probe

Both task and auxiliary gradients use the same implemented backward path,

$$\widetilde{\nabla}_\theta R = \widetilde{J}^\top \nabla_f R, \tag{25}$$

where $\widetilde{J}$ is the surrogate pseudo-Jacobian, not the classical derivative of the hard-spike forward map. Let $g_T = \widetilde{\nabla}_\theta \mathcal{L}_{\text{task}}$ and let $g_C = \widetilde{\nabla}_\theta C$ be the pseudo-gradient of the *positive* CV statistic. With

$P(g) = \mathrm{clip}(g) = \min(1, 1/\|g\|_2)g$, the three separately clipped perturbations are

$$\delta\theta_T = -\eta P(g_T), \qquad \delta\theta_R = -\eta P(-\lambda g_C), \qquad \delta\theta_{TR} = -\eta P(g_T - \lambda g_C). \tag{26}$$

For branch j, the surrogate-linearized prediction is $\widehat{\Delta C}_j = g_C^\top \delta\theta_j$, using the perturbation actually applied, whereas the measured hard-forward response is $\Delta C_j = C_{\mathrm{hard}}(\theta + \delta\theta_j) - C_{\mathrm{hard}}(\theta)$.

The combined-minus-task prediction has a fixed sign under this clipping rule. Euclidean projection onto the unit ℓ_2 ball is firmly nonexpansive, so

$$[P(x) - P(y)]^\top (x - y) \geq \|P(x) - P(y)\|_2^2 \geq 0. \tag{27}$$

Taking $x = g_T$ and $y = g_T - \lambda g_C$ gives

$$\widehat{D} = \widehat{\Delta C}_{TR} - \widehat{\Delta C}_T = \eta g_C^\top [P(g_T) - P(g_T - \lambda g_C)] \geq 0. \tag{28}$$

This statement concerns the contrast, not the absolute combined-branch prediction, and equality is possible. It is a property of the surrogate-linearized perturbation, not of the hard-spike forward map.

Table A15: Measured one-step combined-minus-task hard-forward CV contrast in train mode (mean $\pm$ sample SD over three seeds). The mature-SDT response is negative in all three seeds; early SDT and both SEW values are near zero.

System	Epoch 20 checkpoint	Epoch 100 checkpoint
SDT	$+0.00036 \pm 0.00062$	-0.0906 ± 0.0583
SEW-ResNet	-0.00059 ± 0.00103	$+0.00022 \pm 0.00039$

The probe contains two native systems, two source epochs, three seeds, three fixed clean validation batches of 32, and train/evaluation model modes. The primary analysis uses deterministic replay in train mode; batches 0–2 are averaged arithmetically within seed. Each branch uses $\lambda = 5 \times 10^{-4}$ and the checkpoint learning rate (3×10^{-4} at epoch 20 and 2.4667×10^{-4} at epoch 100). Parameters, buffers, spiking state, and RNG are restored before each branch. The perturbation contains no AdamW moments or weight decay.

Across the 12 primary seed-level system/epoch summaries, surrogate-linearized and hard-forward signs agree in 7 summaries for the regularizer-only branch and 5 for the combined branch; exact zero is treated as a separate tie and none occurs. More directly, Eq. (28) is nonnegative while the mature-SDT measured contrast is negative in all three seeds. The regularizer-only SDT epoch-100 response is mixed (train-mode mean $+0.0708 \pm 0.1374$; evaluation-mode mean -0.0520 ± 0.0833), so the result does not support a general claim that a pure regularizer step lowers hard-forward CV.

The probe shows that a loss-side local reference need not survive the discrete forward map and that the response is state dependent. It does not replay AdamW, predict the long-run endpoint, or isolate surrogate gradients as the cause. The shared-state continuation in Fig. 3 addresses the real-optimizer branch effect over training.

E.2 Pre-specified Hoyer extension from matched full-SDT states

The two Hoyer branches, their calibration rule, and their endpoint analysis were specified before the new Hoyer endpoints were inspected. The five immutable epoch-100 Task sources and completed Task and PopFR-CV anchor branches pre-existed this extension. The added branches apply Hoyer either to each input's response vector or after averaging each coordinate over minibatch samples and time. Endpoint blindness applies to calibration and analysis of the new Hoyer branches; the completed parent experiment predates that protocol. All four branches share the same source hash and epochwise training-stream hashes within each seed; the summary script stops before reporting scientific contrasts if either check fails.

For coefficient calibration, let θ_0 be the restored source, $\theta_T^{(3)}$ the state after three Task updates, and $\theta_{T+R}^{(3)}$ the state after replaying those batches with auxiliary objective R. The source-training-only displacement ratio is

$$q_R(\lambda) = \frac{\|\theta_{T+R}^{(3)} - \theta_T^{(3)}\|_2}{\|\theta_T^{(3)} - \theta_0\|_2}. \tag{29}$$

The POPFR-CV branch at $\lambda_{\text{pop}} = 5 \times 10^{-4}$ defines the target separately within each source state. Calibration replays AdamW moments, weight decay, global gradient clipping, Mixup randomness, and the frozen epoch-100 learning rate. A fixed coarse log grid over $[10^{-6}, 10]$ and a fixed local refinement select the closest Hoyer coefficient; no validation metric or official-test example is constructed or accessed. The scalar q_R matches the magnitude of the initial three-update, task-relative deviation. It does not match the displacement direction, its layerwise allocation, the three-step path, or the relative objective strength later in training. Table A16 shows that all ten selections satisfy the pre-specified 15% gate.

Table A16: Source-training-only calibration for the full-SDT Hoyer extension. Error is the relative deviation from the POPFR-CV anchor's three-update task-relative displacement-norm ratio q.

Seed	Target q	$\lambda_{\text{per-input H}}$	Error (%)	$\lambda_{\text{lifetime H}}$	Error (%)
301	.42365	.72862	.282	.14538	.564
302	.34023	3.65174	1.630	.59566	.144
303	.26780	.39811	1.317	.05309	.947
304	.42773	2.98538	.129	.59566	.038
305	.25359	.39811	1.633	.05309	3.171

Table A17 reports the paired endpoint estimates. CV and MFR ratios are summarized on the log scale and transformed back; additive outcomes use same-seed differences. Relative to POPFR-CV, per-input Hoyer has a 4.937-fold higher endpoint CV (pointwise 95% interval $[4.079, 5.976]$) and 2.586 percentage points less equal-tensor silence ($[-4.778, -.394]$), while its accuracy difference is unresolved ($-.424$ points, $[-1.473, .625]$). Lifetime Hoyer has a 0.311 CV ratio relative to per-input Hoyer ($[.215, .450]$), 2.546 points more silence ($[.432, 4.660]$), and $.764$ points higher validation accuracy ($[.142, 1.386]$). These protocol-specified contrasts show structured branch differences. The intervals are pointwise and provide no simultaneous coverage guarantee. No pre-specified equivalence margin supports a practical-equivalence claim. The Hoyer-trained CV readouts alone cannot establish reversal of the averaged Hoyer objective, because averaging does not commute with its nonlinear mapping from CV.

Table A17: Paired epoch-300 contrasts to Task in the full-SDT matched-axis continuation ($n = 5$ seeds). Brackets are pointwise 95% Student-t intervals. CV and MFR entries are geometric ratios. Coefficients are per-source calibration controls, not a practical hyperparameter ranking.

Branch	CV ratio	Δaccuracy (pp)	Δsilent (pp)	MFR ratio
POPFR-CV	.0877 [.0700, .1099]	$+.556$ [$-.196, 1.308$]	$+32.146$ [$31.531, 32.761$]	.9882 [.9801, .9963]
Per-input Hoyer	.4331 [.3755, .4995]	$+.132$ [$-.335, .599$]	$+29.560$ [$26.994, 32.126$]	.6907 [.4636, 1.0289]
Lifetime Hoyer	.1348 [.0808, .2248]	$+.896$ [$.398, 1.394$]	$+32.106$ [$31.366, 32.847$]	.9639 [.9409, .9876]

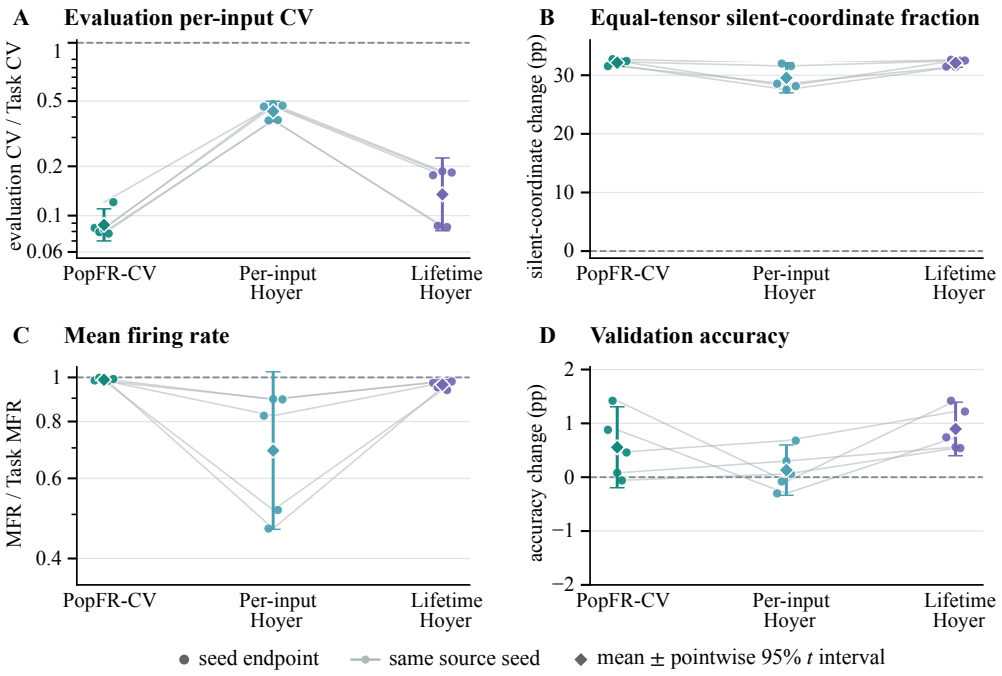

Figure A4: **Full outcomes for the source-matched CV–Hoyer comparison.** Five CIFAR-100 SDT states branch at completed epoch 100 and continue to epoch 300. Per-source Hoyer coefficients match the magnitude of PopFR-CV's initial three-update, task-relative AdamW trajectory deviation. Direction and later evolution remain unmatched. **(A)** Evaluation per-input CV ratio to Task. **(B)** Equal-tensor silent-coordinate-fraction change. **(C)** MFR ratio to Task. **(D)** Validation-accuracy change. Thin lines connect branches sharing a source state and training stream; diamonds show paired means with pointwise 95% Student-t intervals. CV and MFR centres and intervals are computed in log space and displayed on log axes. The branches share the qualitative low-CV, enlarged-silence response and occupy different quantitative firing and dispersion endpoints.

The result rejects the simple prediction that sharing the per-input axis and ideal same-vector ordering is sufficient to match the PopFR-CV endpoint. It also shows that the per-input/lifetime label alone does not order the measured endpoints: lifetime Hoyer is numerically closer to PopFR-CV in silent-fraction increase and endpoint CV than per-input Hoyer is. Several mechanisms remain compatible with these observations; the experiment does not separately identify state-dependent weighting, aggregation, boundary handling, clipping, AdamW history, or the evolving hard-spike map.

E.3 Prospective displacement-magnitude extension

We prospectively extended the per-input PopFR-CV–Hoyer comparison to source-local targets $0.5q$, q, and $2q$, where q is the anchor displacement ratio in Eq. (29). The anchor endpoints pre-dated this extension; the low/high continuations and the across-dose rule were fixed before either new endpoint was inspected. For each of the five immutable sources, low/high coefficients were selected separately for PopFR-CV and per-input Hoyer using the same three source-training minibatches. All 20 selections pass the fixed 15% gate: relative errors range from .09–4.70% at the low target and .53–9.18% at the high target. All 20 continuations complete numerically, and source, calibration, implementation, and epochwise training-stream hashes pass the frozen integrity checks.

At the low target, both objectives remain task-functional. Relative to Task, PopFR-CV has geometric endpoint-CV ratio .1298 $[.0881, .1911]$; per-input Hoyer has ratio .8053 $[.6848, .9470]$; the paired Hoyer/PopFR-CV ratio is 6.206 $[4.435, 8.683]$. Their respective equal-tensor silent-coordinate changes are $+31.857$ $[31.281, 32.434]$ and $+15.483$ $[10.153, 20.812]$ points. At the high target, PopFR-CV remains above the collapse threshold in 5/5 seeds and incurs an accuracy change of -7.848 $[-12.986, -2.710]$ points. Per-input Hoyer reaches chance-level accuracy with zero measured MFR in four seeds; the fifth ends at 52.42% accuracy and MFR .01171. Because the

protocol defined any required collapsed branch as inconclusive, no high-dose Hoyer CV aggregate or stable-only subset is reported.

Table A18: Prospective magnitude extension from the same five full-SDT sources. Brackets are pointwise 95% Student-t intervals over paired seeds. CV entries are geometric ratios. The anchor endpoints pre-date the extension; low/high outcomes and the three-dose decision are prospective. "–" marks an aggregate forbidden by the pre-specified collapse gate.

Target	PopFR-CV/Task CV	Hoyer/Task CV	Hoyer/PopFR-CV CV	Collapses (PopFR-CV/Hoyer)	Frozen decision
$0.5q$	.1298 [.0881, .1911]	.8053 [.6848, .9470]	6.206 [4.435, 8.683]	$0/5$ / $0/5$	stable contrast
q	.0877 [.0700, .1099]	.4331 [.3755, .4995]	4.937 [4.079, 5.976]	$0/5$ / $0/5$	pre-existing anchor
$2q$	.0822 [.0765, .0884]	–	–	$0/5$ / $4/5$	inconclusive

The frozen across-dose outcome is therefore **inconclusive**; dose-robust separation is unsupported. The low target supplies a prospective replication of quantitative endpoint separation. The high target shows that matching an initial three-update displacement norm does not ensure comparable later stability. Universal robustness of PopFR-CV, universal instability of Hoyer, and a unique mechanism remain unsupported; no rescue dose, replacement seed, or post-hoc threshold is introduced.

E.4 Earlier matched-gradient MLP diagnostic

At a shared task-only MNIST spiking-MLP checkpoint, each regularizer coefficient was scaled so that its regularizer-only parameter-gradient norm equaled 1% of the task-gradient norm on a fixed batch. The three branches were per-input CV, per-input Hoyer, and lifetime Hoyer; each used three seeds.

Table A19: Matched-gradient, matched-axis diagnostic ($n = 3$). Similarity between per-input CV and Hoyer varies by endpoint. Aggregation contributes; other factors remain unresolved.

Loss	Accuracy (%)	MFR	Eval. pop. CV	Silent (%)
Per-input CV	$98.06 \pm .12$	$.168 \pm .095$	$2.385 \pm .758$	7.49 ± 7.13
Per-input Hoyer	$98.15 \pm .09$	$.184 \pm .061$	$2.108 \pm .391$	4.49 ± 2.73
Lifetime Hoyer	$98.13 \pm .11$	$.314 \pm .035$	$1.452 \pm .120$	1.63 ± 1.11

E.5 Recovery after removing Population-CV

We test whether induced-silent coordinates can recover after removing PopFR-CV in a 16-unit toy network. All branches start from the same per-seed checkpoint induced with the SDT-compatible Sigmoid surrogate ($\alpha = 4$). This stress intervention deliberately collapses the network and is not an accuracy-preserving operating point. A coordinate is eligible if its validation firing rate is at least 10^{-3} after task-only warmup and in the matched task-only continuation, then falls below 10^{-3} after induction. Across ten seeds, 157 coordinates meet this definition; these pooled coordinates are descriptive, while seeds are the replication units. During recovery, Adam state is reset, minibatch order and hard forward thresholds are fixed, and only the backward surrogate changes. Recovery requires firing rate at least 10^{-3} at two consecutive epoch-end evaluations.

Table A20: Ten-seed toy recovery after removing PopFR-CV. All 157 eligible induced-silent coordinates recover under the smooth Sigmoid and ATan surrogates, whereas none recovers under the rectangular surrogate or while the regularizer remains active. Pooled coordinate counts are descriptive; the seed is the replication unit.

Recovery branch	Recovered / eligible	Median epoch	Final accuracy (%)
Sigmoid $\alpha = 2$, reg. off	$157/157$	4	$86.11 \pm .51$
Sigmoid $\alpha = 4$, reg. off	$157/157$	4	$86.18 \pm .41$
Sigmoid $\alpha = 8$, reg. off	$157/157$	4	$86.22 \pm .53$
ATan $\alpha = 2$, reg. off	$157/157$	4	$86.12 \pm .49$
Rectangular width 1, reg. off	$0/157$	right-censored	$12.70 \pm .00$
Sigmoid $\alpha = 4$, reg. on	$0/157$	right-censored	$12.70 \pm .00$

E.6 Effect of Hoyer onset

The onset study contains 4 start epochs $\times 3$ coefficients $\times 3$ fixed seeds on SDT/CIFAR-100. Earlier starts at epochs 0 or 5 yield 4 collapses among 18 cells-by-seed outcomes. Later starts at epochs 10 or 20 yield none among 18. Collapse means epoch-300 top-1 accuracy below 5%; all four collapsed runs reached 1% accuracy and zero MFR. Reuse of the same three seeds across intervention cells makes these counts descriptive and precludes a binomial comparison. Within this grid, later onset is associated with fewer observed collapses.

The five-seed comparator cohort in Table A4 and this onset grid overlap nominally at $\lambda_H = .05$, onset 20, and seed index 2. They are independent from-scratch cohorts generated by `unified_table_compare.py` and `hoyer_lambda_sweep.py`, respectively. Both drivers use the same nominal SDT, Hoyer formula, optimizer schedule, and onset semantics. Their legacy records do not bind execution to an exact code hash, software/hardware stack, or bitwise-deterministic CUDA setting, so a common integer seed is not a complete replay identifier across cohorts. Seed index 2 collapses in the comparator cohort and remains stable in the onset grid. We retain both observations: the onset grid describes its own $4/18$ versus $0/18$ pattern; the comparator result prevents a delayed-onset safety guarantee. The retained metadata are insufficient to attribute the different outcomes to one source.

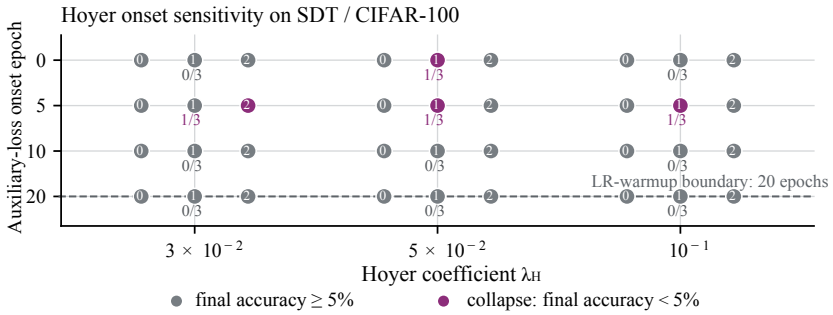

Figure A5: **Later Hoyer onset reduces observed collapse within this grid; an overlapping comparator prevents a safety guarantee.** SDT/CIFAR-100 runs train for 300 epochs. Each cell contains the same three fixed seeds; magenta marks final top-1 accuracy below 5%, and the number below each triplet is the collapse count. Four collapses occur, all for starts at epochs 0 or 5; none occurs for starts at 10 or 20 in this grid. The dashed horizontal line marks the 20-epoch learning-rate-warmup boundary. The $4/18$ versus $0/18$ comparison is descriptive because seeds are reused across cells. A separate comparator cohort contains one onset-20 collapse under an overlapping nominal condition, as discussed in the text.

F Legacy lifetime-distribution diagnostic

This earlier three-seed, weak-coefficient cohort asks whether endpoint means conceal a rare high-firing tail. It is descriptive and protocol-distinct from the five-seed matched-axis continuation above. The full-range binned tail distribution shows broad overlap among conditions, while the lifetime-Hoyer setting retains a small near-saturated tail.

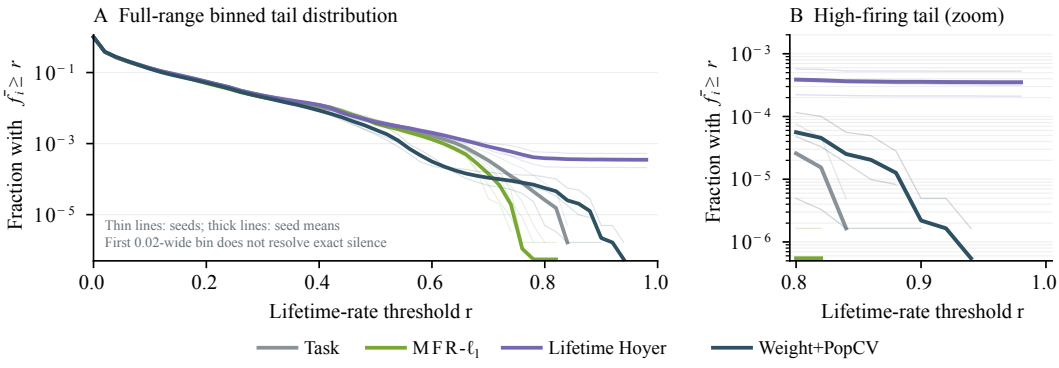

Figure A6: **Legacy lifetime-rate distributions reveal a small high-firing tail.** This exploratory SDT/CIFAR-10 cohort uses $T = 4$, 300 epochs, and seeds 0–2. **(A)** Full-range binned complementary distributions; thin lines are seeds and thick lines their means. **(B)** Enlargement of thresholds 0.8–1.0, where lifetime Hoyer retains a small near-saturated tail. Coefficients are $\lambda_{\mathrm{MFR}} = .01$, $\lambda_H = .005$, and $\lambda_w = .001$, $\lambda_{\mathrm{pop}} = .0005$ for WEIGHT+POPCV; auxiliary terms begin after 20 task-only epochs. The 50 bins have width .02, so the first bin merges zero-observed and low-positive rates and cannot resolve the 10^{-6} silence threshold. This protocol-distinct diagnostic provides context only. The five-seed matched-axis comparison is separate.

G SPIKE-OPERATION MODEL AND ASSUMPTIONS

The SOP calculation assigns each spiking layer the MAC count of its downstream convolution or linear operation, multiplies by its measured neuron-weighted firing rate and time steps, and sums across layers. The 45-nm proxy assigns 0.9 pJ per spike-driven addition, matching the cited estimate for a 32-bit floating-point addition (Horowitz, 2014). Table A21 reports the $T = 4$ architecture accounting. Because it uses neuron-weighted MFR and different seed aggregation, relative SOP values should be compared only within this table.

Table A21: Modeled SDT spike operations at $T = 4$. Values exclude memory, routing, control, leakage, and idle energy and require sparse-execution support. No device energy or speedup is measured.

Data	Configuration	Neuron-wtd. MFR	SOPs (10^6)	Modeled μJ	vs. base
C10	Baseline	.0432	13.4	12.03	1.000
	POPFR-CV	.0415	12.8	11.54	.959
	WEIGHT+POPCV	.0426	13.2	11.84	.985
C100	Baseline	.0546	16.9	15.19	1.000
	POPFR-CV	.0538	16.6	14.98	.986
	WEIGHT+POPCV	.0522	16.2	14.54	.957

H REPRODUCIBILITY AND EXPERIMENTAL PROVENANCE

The anonymized supplement is a lightweight audit release organized by experimental protocol. It contains training and aggregation code, frozen protocols and manifests, selected machine-readable run records, calibration records, derived statistics and tables, plotting sources, and the final figure PDFs. It deliberately excludes datasets, model and optimizer checkpoints, launcher logs, caches, complete server exports, and version-control metadata. Its evidence is delivered at three levels:

- **Direct reaggregation from included records:** the fixed-objective $T = 4$ comparison, validation-selected confirmation, Tiny-ImageNet and $T = 8$ extensions, Hoyer-onset grid, paired common-nonzero readout, and the released matched-axis and magnitude-extension branch records. The magnitude extension includes its frozen manifest, five calibration

records, all 20 branch JSON records with stored training and validation trajectories, every high-dose collapse, and the pre-specified inconclusive decision.

- **Summary and provenance audit:** the shared-state continuation, tensor localization, objective-component continuation, complete four-branch matched-axis comparison, and several cross-system boundaries are supplied as compact seed-level tables, derived statistics, protocol notes, or provenance records. These files support numerical and provenance checks without reconstructing omitted model states.
- **Reproduction by rerunning:** regenerating checkpoints, replaying checkpoint-dependent probes, or reproducing full training trajectories requires downloading the public datasets and rerunning the released code. The README maps every evidence group to its delivery level and primary paths.

The release supports protocol-level inspection, direct reaggregation where the required records are included, and full retraining from the supplied entry points. It does not promise bitwise CUDA determinism. Some earlier fixed-coefficient result files predate complete Python/CUDA/GPU and Git-SHA capture; this is stated in their metadata. The primary SDT runs use single 24-GB GPU workers and require roughly seven GPU-hours per 300-epoch fixed-coefficient run in the reference environment; the small MNIST and toy experiments can also run on CPU. The SDT reference implementation is distributed under Apache-2.0, and public datasets are downloaded through their standard loaders. No human-subject or sensitive data are used.

I LIMITATIONS

- **Central scope.** Full matched evidence is confined to $T = 4$ SDT on three static-image datasets. Tiny-ImageNet extends the class and image scale within the same backbone; architecture and ImageNet-scale generality remain untested.
- **Selection.** Fixed-coefficient $n = 5$, Tiny-ImageNet, $T = 8$, Hoyer-onset, and Spikformer runs monitored their evaluation split. Only the separate CIFAR-10 experiment uses validation-only selection followed by fresh one-time official-test evaluation.
- **Architectures and modalities.** Native SEW-ResNet shows near-zero response under a same-nominal matched continuation; because native activation sets, precision, and effective objective scale differ, this is not architecture isolation. Spikformer and the MNIST MLP show partial responses. CIFAR10-DVS is a non-replication of the joint endpoint, and no harder event benchmark was completed.
- **Hoyer comparison.** The full-SDT matched-axis continuation controls source state and training stream and matches the magnitude of the initial three-update, task-relative AdamW trajectory deviation. Direction and subsequent evolution remain unmatched. Its per-seed Hoyer coefficients are source-specific calibration controls. Outcomes are measured on held-out validation data; no official-test set is accessed. The experiment discriminates simple ordering/axis predictions. Practical equivalence and a ranking of broadly tuned objectives remain untested.
- **Magnitude extension.** The prospective $0.5q/q/2q$ extension reproduces the CV endpoint separation at $0.5q$. Its across-dose decision is formally inconclusive because per-input Hoyer collapses in 4/5 seeds at $2q$. Every outcome is retained; no stable-only aggregate, rescue dose, replacement seed, or relaxed collapse threshold is used.
- **Mechanism.** The paired endpoint readout localizes the aggregate change to selective attention-path withdrawal; within-event CV equalization receives no support. The analysis leaves the path-selection mechanism and functional necessity unresolved. Exact mathematics rules out an intrinsic ideal same-vector CV–Hoyer gradient-direction difference at positive nonuniform interior points. The matched-axis continuation and local probe constrain additional simple explanations without separately identifying state-dependent weighting, aggregation, boundary handling, clipping, AdamW history, numerical stabilization, or system design.
- **Energy.** ICE and SOPs are proxies. Device energy, wall-clock speed, memory traffic, routing, and leakage are unmeasured.

- **Terminology.** In the $T = 4$ static-image evaluations, the 10^{-6} rule denotes only a zero-observed-spike criterion. It carries no claim of biological death, parameter removal, or permanent silence. Equal-tensor weighting changes the magnitude and preserves the direction of the paired silent-set response.