Interpreting Compression-Order Effects in Spiking Neural Networks

Anonymous authors
Paper under double-blind review

Abstract

Compression-order comparisons depend on how models are constructed and calibrated. We study channel pruning, unfolded weight-only quantization and threshold scaling in spiking neural networks (SNNs), without intervening training. Matched-state controls distinguish batch-normalization (BN) history from model construction: aggressive clipping without rounding reproduces large history-dependent gaps, including in ReLU networks, and common recalibration yields identical logits. Separately, a prespecified twelve-run CIFAR-100 ResNet-SNN comparison at two time steps gives a mean absolute accuracy gap of 0.17 percentage points (pp) on reused test images, despite 6.0% prediction disagreement under the specified construction. Both orders satisfy the prespecified operational 5-pp budget on evaluated validation and test sets, with 10.08% fewer convolution multiply–accumulate operations. Descriptive extensions include matched T=4 retraining, additional image and event-frame datasets, and Transformer MLPs. Calibration improves both-order budget passage in the complete ImageNet-100 comparison; full reconstruction changes one finite-split qualification decision that final-only recalibration preserves. In a descriptive comparison of twelve T=2 ResNets at α=1.02, the mean between-order relative synaptic-operation range contracts from 12.48% to 0.031% after common recalibration, with all orders remaining budget-feasible. Retrospective candidate policies on reused validation splits quantify compression choices and test-budget violations. These observations distinguish normalization history, model behavior, budget feasibility and activity-dependent work in compression-order evaluation.

1 Introduction

Pruning and quantization combine structural removal with lower precision (Han et al., 2016; Frantar et al., 2022), including in spiking neural networks (SNNs) (Deng et al., 2023; Wei et al., 2025). Their interactions make order a design choice (Harma et al., 2025; Kim et al., 2026), but accuracy gaps alone do not identify their source. Batch normalization (BN) uses running statistics at inference (Ioffe & Szegedy, 2015); stale statistics can distort pruned-candidate evaluation (Li et al., 2020). In our diagnostic pipeline, pruning (P) includes BN recalibration, whereas quantization (Q) leaves the BN buffers unchanged. Before common final recalibration, QP uses final-model statistics; PQ retains pre-quantization statistics. Current-network scores and tie-breaking can select different channels across orders; physical removal can change quantization ranges.

We ask **what order-dependent accuracy gaps measure and which compression decisions they justify.** We prune without retained-weight compensation and quantize unfolded weights without intervening learning. First, matched-state SNN and ReLU controls distinguish BN history from construction; their state condition defines the control's validity rather than predicting unequal-state accuracy. Second, the prespecified T=2 compact comparison measures mean absolute gap and validation-selected regret, complemented by descriptive signed contrasts, prediction disagreement and matched T=4 retraining. Third, calibration can improve worse-order accuracy while preserving better-order accuracy, change qualification upon full reconstruction despite final-only stability, and reduce work differences

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.

while all orders remain accuracy-feasible. Retrospective candidate policies describe finite-split decisions without establishing selector superiority.

2 Related work

SNN compression. QP-SNN trains with quantized weights and fine-tunes after pruning (Wei et al., 2025). We reuse its singular value of spike activity (SVS) and ReScaW components without training; our controls do not reanalyse its reported ordering advantage. Other work covers activity-based pruning (Li et al., 2024), pruning/quantization (Rathi et al., 2019; Chowdhury et al., 2021), and post-training compression (Shi et al., 2025). Q-SNNs quantizes weights and membrane potentials (Wei et al., 2024); PTQ4SNN calibrates weight and recurrent-membrane quantization (Xie et al., 2026). Our Q is weight-only. Joint compression training (Deng et al., 2023) and BASC's temporal-scale and channel-selection corrections (Chen et al., 2026) adapt compression.

Ordering and adaptation. Beyond SNNs (Shen et al., 2024; Harma et al., 2025), Kim et al. (2026) optimize order, proposing weaker perturbations first. Kim and Harma include no-fine-tuning comparisons; we control order-specific BN history without reanalysing their pipelines. AMC/HAQ allocate compression within operator families (He et al., 2018; Wang et al., 2019); LitE-SNN jointly searches SNN architecture, pruning, precision and timesteps with resource objectives and retraining (Liu et al., 2024). Our retrospective policies use fixed candidates (C.5).

Normalization and post-training quantization. EagleEye recalibrates BN before evaluating pruned candidates (Li et al., 2020); universally slimmable networks estimate subnetwork statistics after training (Yu & Huang, 2019). Para-Normalization recollects post-quantization BN moments (Hubara et al., 2021); Shomron & Weiser (2020) recalibrate after hardware precision perturbations. Population estimation and training/inference discrepancies differ (Wu & Johnson, 2021) (A.3). DFQ addresses range imbalance and quantization bias (Nagel et al., 2019); ACIQ studies clipping versus rounding (Banner et al., 2019). AdaRound optimizes rounding (Nagel et al., 2020) and BRECQ reconstructs blocks (Li et al., 2021) without end-to-end retraining; our fixed quantizer performs neither optimization.

Compression evaluation. Pruning needs consistent evaluation (Blalock et al., 2020); similar accuracy can conceal example-level changes (Hooker et al., 2019) and prediction disagreement (Marx et al., 2020). Compress and Compare tracks compression provenance, efficiency and behavior (Boggust et al., 2025). Our matched-state controls distinguish BN history, construction and order-specific feasibility.

3 Model state and controlled comparisons

3.1 Operators and orderings

State (z, r) separates BN running buffers r—means, variances and counters—from the remaining model state z: weights (including BN affine parameters), masks, thresholds and execution semantics. Configuration (p, b, α) specifies pruning ratio, weight bit width and threshold multiplier. PQA applies P, then Q, then A. Without intervening training, we use SVS and ReScaW with the implementation choices specified in appendix A; fixed reference bounds hold ranges constant across orders.

P masks the lowest-SVS output channels and recalibrates BN. SVS averages spatial spike-matrix ranks over calibration examples from the current network. Mask-only P leaves BN affine outputs active; post-BN-clamped P zeros masked-channel BN outputs and recalibrates again. Thus P bundles pruning and normalization. Masking preserves graph size; §4.1 tests removal.

Q in the original studies uses a per-output-channel, symmetric zero-free midrise grid, with $n = (2^b - 1)/2$:

$$Q(w) = c\left[\frac{\text{round}((\text{clip}(w/c, -1, 1) + 1)n)}{n} - 1\right].$$

Here range denotes the symmetric magnitude bound c of $[-c, c]$. The current channel supplies either maximum range $c = \max|W|$ or mean range $c = (2^b - 1)2^{1-b}\,\text{mean}\,|W|$. With current bounds, all-zero channels stay zero and stored masks are reapplied. All-layer coverage includes the stem and classifier; blocks-only coverage leaves them full precision. BN is not folded into the quantized weights; §4.2 separately tests a zero-containing grid. Direct clipping uses identical ranges and coverage without rounding or the quantizer's division and rescaling by c; max-range clipping is identity.

A overwrites thresholds with α times their stored base values and changes no weights, masks or calibration state. It is identity at $\alpha = 1$. Neuron, surrogate-gradient and BN conventions are specified in Appendices A.1, A.3 and E.

3.2 Validity and limits of the fixed-state control

Proposition 1 (common recalibration at fixed model state z). Suppose different orders construct identical z. If deterministic recalibration on stream C replaces r by B(z;C) independently of initial buffers, recalibrated inference functions are identical.

Proof. Recalibration maps each model to R(z,r;C) = (z,B(z;C)). Complete functional states and inference rules coincide. We use reset followed by cumulative moving averaging (reset-CMA; A.3) to ensure initial-buffer independence. Logit comparisons on evaluated inputs check implementation; agreement there alone does not prove all-input equivalence.

Corollary (repeated recalibration). R is idempotent. Original post-pruning and final BN recalibration share the map and stream. At $\alpha = 1$, QP already ends with this calibration, so repeating it leaves the branch unchanged, irrespective of whether QP has highest accuracy.

A fixed whole-output-channel mask leaves each filter and its range unchanged or zeroed, with zeros preserved by mask reapplication. Final states z agree despite different BN histories. Reselection breaks this premise through activity-dependent scores, which also depend on pre-scoring recalibration. Physical removal can change downstream ranges and weights even with fixed channels. For example, 4-bit max-range Q maps the first coordinate of (1,2) to 14/15; removing the second coordinate before Q maps it to 1. Unequal states require measurement; §5 checks accuracy.

Threshold-independent Q and A commute, reducing permutations to at most four order classes (A.2); activation- or membrane-calibrated Q and intervening learning require rechecking.

3.3 Measurements and interventions

For each model and configuration, a(s) is order-class accuracy on the named split without intervening training. The oracle is max a(s); spread S = max a(s) − min a(s); regret = max a(s) − a(chosen). For PQ/QP, S is the within-run absolute accuracy difference. B_0 is original uncompressed accuracy; B_1 and B_{aug} are uncompressed accuracies after the original unaugmented and augmented final BN procedures, respectively (A.3, D.1). The matched reference uses the endpoint's calibration procedure to measure incremental compression loss. Otherwise, losses use B_0 and include calibration-induced loss. The fixed order-spread margin is $\delta = 1.5$ pp; the 5.0 pp budget concerns compression loss from the baseline. The optimistic test oracle is not a selection policy. Feasibility means meeting a stated loss budget on the named split; qualification combines BN preservation and the both-order loss budget on the qualification split (validation for local cohorts); spread is an outcome, not an exclusion rule. These thresholds are operational, not universal (sensitivities: C.1/C.3).

Fixed masks use original models in masking studies and augmented-recalibrated full-precision models in removal controls; reselection scores immediately before P. Final unaugmented recalibration fixes z, resets BN buffers and averages the same training subset with only BN in training mode; evaluation uses running statistics (A.2–A.3).

Spread-reduction contrasts compare Q, identity and matched clipping across range choices (B.2). They are not accuracy gains or additive causal decompositions; extremizing orders can change.

3.4 Calibration procedures and validation

In the T=2 crossed ResNets, unaugmented recalibration lowers every model's accuracy: 76.614%→73.773% on average (change -2.841 pp). We compared augmented CMA with unaugmented sequential estimation on validation data, using upstream BN in inference mode for the latter (Wu & Johnson, 2021). Only augmented CMA met the BN-preservation tolerance for every model: mean/maximum validation loss 0.452 pp/0.903 pp. The selected procedure yielded a mean test accuracy of 76.848% versus 76.614% originally, maximum loss 0.280 pp over 12 models. Appendix D.1 reports the failed procedure and opposite compressed-accuracy changes under mean/max ranges: baseline preservation does not ensure compressed accuracy. Spikingformer's input/procedure comparison (E.3) likewise identifies neither an individual transform nor a universal estimator.

3.5 Evidence and statistical scope

Table 1: Questions, evidence and interpretation scope.

Question	**Evidence**	**Interpretation scope**
BN history or model construction?	Matched-state SNN/ReLU controls (B)	Original T=2 families; descriptive T=4; no external-pipeline correction
Same accuracy, same behavior?	Primary T=2 compact ResNet; T=4 and image/event/MLP extensions (C, E)	Distinct predictions; trained-pipeline comparisons; reused cohorts
Same accuracy, same choices/work?	Candidate policies and active thresholds (C.5, F)	Finite-split decisions; modeled SOP, not energy

For the local compact cohorts, validation is reserved before training; epoch and order selection precede test evaluation (A.1, A.4). Matched T=4 retraining selects epochs separately, comparing trained pipelines rather than fixed-weight inference; its analyses are descriptive. Within each run we average 4 VGG clipping or 8 crossed-ResNet configurations per setting.

Records retain every run/configuration. Test-selected IN-100/DVS/recovery and public SEW masking informed development but are excluded; secondary recovery/range/frontier/regression analyses are archived. Reused-model calibration follow-ups retain all policies/failures/fallbacks (C.5), streams and the original Transformer failure (E).

The original T=2 VGG two-contrast and crossed ResNet four-contrast t families use 97.5% and 98.75% confidence, respectively (B.1–B.2). This nominal family control excludes earlier data-informed family selection. Private records preceding new T=2 compression outcomes fix spread and validation-selected regret as primary (C.1), each with a nominal one-sided 97.5% t bound (nominal family confidence 95%). T=4 compact ResNet repeats these bounds descriptively, without a new confirmatory family. Run-level summaries condition on reused images/configurations and include paired training/calibration randomizations for the primary cohort. Image resampling is a separate descriptive sensitivity (C.1).

4 Ordering opportunity and construction controls

4.1 Different compact states, small accuracy differences

Compact removal preserves residual dimensions while slicing internal channels, BN and downstream inputs. Fixed masks/ranges restore equal final logits (A.2/C.1; Figure 1).

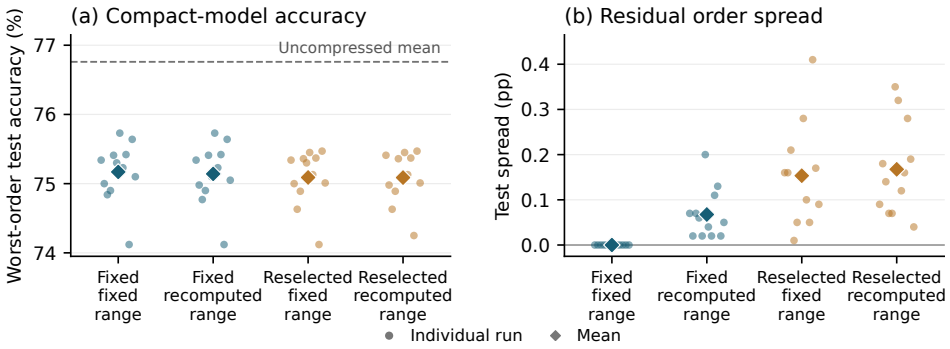

Figure 1: Compact ResNet models at T=2: 12 training/calibration randomizations, conditional on reused test images; 10% internal-channel removal, 4-bit max-range block-only quantization, no threshold scaling. Fixed/reselected masks cross fixed/recomputed ranges. Dots are runs; diamonds are run means. (a) Worst-order test accuracy; dashed: uncompressed mean. (b) Observed absolute accuracy gap. Reselected masks/recomputed ranges define the prespecified primary; other cells are descriptive. Both orders meet 5-pp validation/test loss budgets in every shown run. The text gives mean bounds; C.1 compares T=4.

The prespecified T=2 comparison (Figure 1) and matched T=4 retraining each use twelve ResNet-SNNs: reselected masks/recomputed ranges, p=.10, 4-bit max-range block Q, no threshold scaling. Augmented recalibration precedes SVS scoring and follows P and final construction (C.1).

At T=2, masks differ in all twelve pairs, with mean Hamming disagreement 1.699% of scored channels; mean worst-order test accuracy is 75.09% versus 76.760% uncompressed. Both orders meet the 5-pp validation/test budgets in every model at both T values. Physical removal saves 10.08% dense convolution MACs against the same-T baseline; T=4 doubles the dense count for a fixed graph, not measured energy or SOPs.

Different predictions, small accuracy differences. S is a finite-dataset absolute test gap, not a noise-corrected population effect. Mean S/regret receive nominal one-sided 97.5% run-level t limits conditional on reused images; only the T=2 pair is primary. Signed PQ−QP intervals are descriptive 95% run-level t intervals (A.4/C.1). Prediction disagreement averages within-run fractions of examples with differing predictions.

Table 2: Compact ResNet results across T.

T	Mean S / upper (pp)	Signed PQ−QP [95% CI] (pp)	Prediction disagreement (%)	Mean regret / upper (pp)
2	0.1675 / 0.232	0.0175 [-0.1106, 0.1456]	5.971	0.062 / 0.107
4	0.1533 / 0.205	-0.1283 [-0.2045, -0.0521]	6.355	0.0092 / 0.029

At T=2, correctness differs on 3.7025% of images. With u_i, v_i images correct only under PQ/QP in run i, correctness disagreement is $100(u_i + v_i)/N_i$, but $S_i = 100|u_i - v_i|/N_i$. Within-run gains and losses nearly balance, accounting for the gap without explaining why. We average the within-run absolute gaps. At p=.25, no T=4 pair meets either split's budget. C.1 reports passage at tighter budgets.

T=2 image/model-bootstrap 97.5th percentiles are 0.3542 pp/0.2025 pp for S/regret, a separate descriptive resampling target (C.1). The equal-correctness conditional-binomial reference gives an expected absolute T=2 gap of 0.1489 pp; it is neither a correction nor a lower bound (C.1). QP wins eleven of twelve T=4 runs, descriptively.

Implementation sensitivity (T=2). Tie-breaking and pre-scoring variants retain both-order budget passage in every model (C.2). Same-order prediction disagreement averages 6.278% (PQ) and 6.902% (QP) when pre-scoring recalibration changes, comparable to cross-order disagreement.

4.2 Transfer across architectures and input domains

T=4 CIFAR-10 VGG uses hidden-channel removal, reselected masks and recomputed max ranges and a zero-containing 4-bit quantization grid, with unfolded BN and full-precision stem/classifier. Architecture, data and quantizer change jointly relative to ResNet (C.3).

At p=.40, convolution MACs fall 40.04%; both-order 5-pp passage is 11/12 on validation and 12/12 on test, which cannot reverse qualification. Mean worst accuracy is 90.598% versus 94.081% uncompressed, with spread 0.2425 pp. T=2 passes both splits throughout at this ratio; T=4 does so only at p∈{.10,.20,.30}, and none qualifies at p=.50. At the tighter 3-pp budget, p=.40 passes 1/12 validation and 3/12 test models (C.3).

Additional datasets and architectures. Epoch and order selection use internal validation; qualification requires baseline BN loss≤1.5 pp and each order's loss≤5 pp. Denominators count reused models. All rows reselect masks/recompute ranges; spreads follow final recalibration.

Table 3: Transfer results and calibration qualification.

Setting (models)	BN criterion (val.)	Budget (val. / test)	Qualification (val.)	Mean test S (pp)
IN-100, p=.05, 4-bit (6)	6/6	6/6 / 6/6	6/6	0.357
IN-100, p=.05, 8-bit (6)	6/6	6/6 / 6/6	6/6	0.147
DVS, p=.1/.2, 4/8-bit (8)	8/8	8/8 / 8/8	8/8	0.2875–0.5125
Spikingformer MLP, crop, 4-bit (8)	0/8	8/8 / 8/8	0/8	0.188
Spikingformer MLP, informed, 4-bit (8)	8/8	8/8 / 8/8	8/8	0.210

Scale and event inputs. ImageNet-100 saves 4.507% convolution MACs at p=.05; p=.10 fails both bit widths. At 4/8 bits, test predictions differ on 14.670%/12.830%. Additional streams yield small terminal gaps, yet full reconstruction reduces validation qualification from 6/6 to 5/6 for one stream; final-only recalibration preserves it (E.1). DVS uses T=10 event-count frames, saving 9.960%/19.971% convolution MACs at p=.1/.2 with maximum test spread 1.10 pp (E.2). Asynchronous processing and threshold–work transfer are untested.

Calibration-dependent Transformer transfer. On CIFAR-100, crop inputs fail BN preservation for all eight models under CMA and sequential estimation; training-informed inputs pass under both (E.3). We selected training-informed CMA using uncompressed-model validation accuracy after BN recalibration; both rebuilt orders qualify. This isolates input-recipe dependence, not individual transforms. Compression covers MLP hidden channels/weights only, excluding attention, tokenizer and head.

Public ImageNet-1k model. A single public checkpoint of Spikingformer crosses MLP removal p∈{0.05,0.10}, 4/8-bit Q, fixed/reselected masks and frozen/recomputed ranges (E.4). All 16 configurations meet the 5-pp budget for both orders on qualification and evaluation partitions of reused official validation. With reselected masks and recomputed ranges, terminal spread is 0.009–0.216 pp. Fixed masks/ranges yield identical final logits.

4.3 Matched clipping controls

The matched-state T=4 VGG comparison uses post-BN clamping and 8-bit mean-range bounds (B.1), averaging p∈{0.01,0.10}, α∈{1,1.05} within each model. Identity, clipping and

Q share masks, coverage and models; the table reports fixed-mask mean spreads before/after final recalibration.

Table 4: VGG spreads under matched clipping controls.

Q-position operation	**CIFAR-100 spread, before → after**	**CIFAR-10 spread, before → after**
Identity	0.185 pp → 0.000 pp	0.174 pp → 0.000 pp
Direct clipping	65.059 pp → 0.000 pp	77.816 pp → 0.000 pp
Quantization	65.117 pp → 0.000 pp	77.786 pp → 0.000 pp

Clipping reproduces these gaps; common recalibration yields identical full test logits for fixed masks. B.1 reports descriptive run-level clipping-versus-identity and Q8-minus-clipping contrasts. In the separate T=2 SNN/ReLU $\alpha = 1$ fixed-mask mean-range clipping comparison, orders ending with P have higher accuracy before common recalibration in every ANN and SNN instance: their statistics were estimated after clipping (B.3). We evaluate the released ReScaW mean-range formula without its upstream quantization-aware training (code mapping: A.2).

With mask reselection, T=4 clipping yields mean spreads of 0.807 pp and 0.190 pp after common recalibration; different channels break the equality premise. In the separate T=2 CIFAR-10 comparison, 5/48 dependent model–configuration instances exceed δ, reaching 5.230 pp; all exceed the loss budget (B.1). Small means do not bound individual instances.

4.4 Crossing range, coverage and pruning–normalization protocol

The original T=2 ResNet study crosses max/mean range, all-layer/block coverage and mask-only/clamped pruning–normalization protocol at p ∈ {0.01, 0.1}, b ∈ {4, 8}, and $\alpha \in \{1.0, 1.1\}$. Appendix B.2 reports the four fixed-mask mean-minus-max R_Q contrasts.

The four nominal within-family intervals exclude zero at fixed coverage/P protocol, including clamped P's extra recalibration. Zero fixed-mask final spreads and matched identity responses reduce mean-minus-max R_Q to pre-recalibration Q spread differences (B.2). Max-range clipping is identity; Q and post-P normalization can still yield nonzero pre-recalibration spread.

In the fixed-mask max-range, all-layer, mask-only ResNet comparison, quantization minus clipping is 2.421 pp at 4 bits versus 0.252 pp at 8 bits. The bit-width-specific controls in B.2 show that the rounding contrast is not uniformly small. With the identity control at the Q position, fixed-mask max-range, all-layer, mask-only spread is 0.000 pp without threshold scaling and 13.427 pp with threshold scaling ($\alpha = 1.1$). The interaction between threshold scaling and post-pruning normalization therefore produces a gap even without weight quantization.

4.5 ReLU control

Twelve ReLU ResNet counterparts use the same training recipe and splits (A.1, B.3), with P and Q but no analogue of threshold scaling A. With fixed masks, mask-only P and all-layer mean-range 8-bit clipping, spread is 59.0 pp in the ANN and 62.1 pp in the historical T=2 SNN cohort at $\alpha = 1$; common recalibration yields identical logits across orders in each network. Four prespecified SNN-minus-ANN contrasts are 3.3–4.2 pp (nominal Welch/Bonferroni 95% family intervals exclude zero; B.3); earlier SNN-informed design choices lie outside that adjustment. Near-chance worst-order accuracy makes these contrasts largely track post-clipping endpoint accuracy. Descriptive max-range comparisons reverse direction with bit width: ANN gaps are larger at 4 bits, SNN gaps at 8 bits.

Under the chosen support-rank rule, pruning alone costs 6.2 pp on average; compact mean spread is 0.41 pp but no ANN pair qualifies at the 5-pp validation budget. A subsequent descriptive shared-L1 comparison with T=2 SNNs qualifies 12/12 ANN and 12/12 SNN

pairs, with mean spreads 0.127 and 0.162 pp (C.4). Compact feasibility in the ANN therefore depended on the pruning rule; under the shared L1 score both networks qualify with small spreads.

4.6 Threshold scaling and synaptic work

With active threshold scaling, the original T=2 compact ResNet cohort uses reselected masks and recomputed ranges (F). Modeled synaptic operations (SOPs) count binary input events over convolution connections, excluding neuron updates, memory and hardware energy. At α = 1.02, all four order classes pass the 5-pp validation/test budgets in every run before and after final recalibration.

Table 5: Accuracy spread and SOP range before and after recalibration.

Final recalibration	**Test spread: mean / max (pp)**	**Relative SOP range (%)**
Before	0.845 / 1.26	12.48
After	0.373 / 0.64	0.031

Relative SOP range averages each run's order range divided by its order mean. Common recalibration contracts this range despite distinct masks/ranges and small accuracy gaps; this unequal-state result does not follow from Proposition 1. Compared with $\alpha = 1$, final SOP falls about 8.2%, at an additional 0.48–0.65 pp mean accuracy loss across corresponding orders. Appendix F reports validation spreads above δ and failures at $\alpha = 1.05$, driven mainly by lower best-order accuracy. F.2 reports stage/image checks and sampled-channel replay.

5 Absolute accuracy and budget feasibility

5.1 Spread reduction and absolute accuracy

Spread can shrink through higher worst-order accuracy, lower best-order accuracy, or both. Defining $\Delta S = S_{\text{pre}} - S_{\text{post}}$ gives $\Delta\text{S} = \Delta\text{worst} - \Delta\text{best}$, with accuracy changes measured after minus before final recalibration. Extremizing orders can change; we track the previously best order too.

In the complete six-model IN-100 p=.05/4-bit comparison, both-order 5-pp passage rises from 2 to 6 on validation and 4 to 6 on test; mean worst test accuracy improves by 1.000 pp while the accuracy of each previously best order is unchanged (D.2). Conversely, the paired DVS example reduces spread while the accuracy of its best order before recalibration changes by -1.00 pp.

All 6 IN-100 p=.10/4-bit fixed-mask/range pairs fail validation qualification and the both-order test budget despite identical final logits and zero spread (E.1). Equality need not imply useful accuracy. Reselected T=4 compact ResNet gaps are already small before final BN (maximum 0.98 pp), following pre-scoring/post-P recalibration; large-gap clipping controls are separate.

5.2 Calibration does not uniformly improve compressed models

In the T=2 crossed grid (D.1), augmented final recalibration increases spread under original P; under recalibrated P it increases mean-range but decreases max-range spread. With scoring inputs/kernels fixed and a common augmented endpoint, recalibrated P lowers setting-mean oracle accuracy and spread throughout. It changes both pre-scoring recalibration and post-P inputs; C.2 isolates the former. Fixed-mask gaps remain zero despite opposite accuracy changes under mean/max ranges. Smaller gaps need not mean better masks or accuracy.

5.3 Budget-constrained order choice

At T=4 VGG p=.40 and the 5.0 pp budget, validation supports at least one order in 12/12 models and both in 11/12. Seed 508's PQ/QP losses are 5.2734/4.7363 pp. Across candidates, fixed PQ/QP select p=.30/.40 (30.03%/40.04% MAC saving). Both p=.40 orders pass the test budget, so this validation boundary is not reproduced on test; we keep the validation-selected choices. Cohort mean savings for fixed PQ/QP/search are 39.206%/40.041%/40.041%, with no test violations. This individual contrast yields no mean gain over fixed QP. Elsewhere, additional search saving occurs with or without test violations (C.5). Validation also selected epochs: these are finite-split policy choices, not independently certified budget boundaries.

6 Discussion

Practical implications. In the evaluated pipelines, BN statistics estimated from training inputs and validation checks of baseline preservation provide controls for interpreting compression-order gaps. A small accuracy gap or all-order budget passage implies neither agreement in predictions or modeled work nor qualification on new data. Candidate selection requires an explicit loss budget, eligibility rule and full-precision fallback. Calibration sensitivity also requires distinguishing final-only recalibration from full reconstruction (§4.2). These comparisons exclude intervening learning.

Limitations and scope. Local qualification reuses epoch-selection images. Loss differences and final-epoch sensitivity do not isolate selection bias (Cawley & Talbot, 2010) (C.2); independent qualification requires a reserved split. Exact-image exclusions preserve primary budgets at fixed predictions; epoch-selection effects remain unmeasured (C.1).

ReLU history effects are shared, but L1 does not isolate neuronal effects and A has no ANN comparison (B.3/C.4). ImageNet/DVS reuse benchmarks; public IN-1k selection is uncontrolled. CIFAR Spikingformer's original calibration fails; its selected alternative passes (E.3). Mean-range Q uses full-precision-trained weights. MACs/SOPs establish neither energy/latency nor packed-storage gains. Same-order CIFAR calibration-stream variation and its interaction with threshold-work differences remain unmeasured.

T=2/T=4 retraining does not isolate neuronal mechanisms; alternative surrogates, resets and time-specific BN such as BNTT (Kim & Panda, 2021) remain untested. We evaluate unfolded weight quantization with floating-point neurons; folding BN before quantization, retained-weight compensation, and activation or membrane quantization remain unevaluated. Common deterministic conversion preserves identical complete-state equality (A.3).

7 Conclusion

Construction and calibration jointly shape compression-order effects. Across the controlled SNN settings studied here, small aggregate accuracy gaps can coexist with different predictions, budget decisions, and modeled synaptic work. Matched-state controls show that common recalibration can remove history-driven differences, while compact-model experiments show that distinct states may still yield similar accuracy despite different behavior. Calibration also changes absolute feasibility: it can improve both-order budget passage, contract work differences even when compact states remain distinct, and alter qualification under full reconstruction. These effects persist across convolutional SNNs, event-frame inputs, and Transformer MLPs, although their magnitude depends on the construction and calibration procedure. Compression-order evaluation should therefore report not only terminal accuracy gaps, but also model construction, calibration, prediction behavior, budget feasibility, and relevant work measures.

Reproducibility Statement

Appendix A specifies the architectures, operator equations, calibration procedure, data splits and training settings. Appendices B–F provide matched controls, compact grids, calibration

responses, paired accuracy accounting, image/event/Transformer comparisons and synaptic-work definitions. Our code and saved experiment records contain per-order outputs and their links to the reported analyses. The paper and appendices report the results supporting each claim.

Ethics Statement

The study uses public image-classification benchmarks and pretrained models. It involves no human subjects or personally identifying data. Its scope is empirical model compression and evaluation.

Generative AI Use Statement

In this work, we used generative AI tools for literature discovery, brainstorming and refinement of research ideas, assistance with experimental planning and software development, drafting and editing portions of the manuscript, and assistance with mathematical exposition. We reviewed all AI-assisted work. The authors determined the final research questions, experimental protocols, mathematical claims, analyses, interpretations, and manuscript content. AI-assisted code and mathematical arguments included in the paper were checked by the authors, and reported numerical results were checked against the underlying experimental records. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

References

Ron Banner, Yury Nahshan, and Daniel Soudry. Post training 4-bit quantization of convolutional networks for rapid-deployment. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/c0a62e133894cdce435bcb4a5df1db2d-Abstract.html.

Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John Guttag. What is the state of neural network pruning? *Proceedings of machine learning and systems*, 2:129–146, 2020.

Angie Boggust, Venkatesh Sivaraman, Yannick Assogba, Donghao Ren, Dominik Moritz, and Fred Hohman. Compress and Compare: Interactively evaluating efficiency and behavior across ML model compression experiments. *IEEE Transactions on Visualization and Computer Graphics*, 2025. doi: 10.1109/TVCG.2024.3456371. URL https://doi.org/10.1109/TVCG.2024.3456371.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Naz Sepah, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Dmitriy Serdyuk, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems*, volume 3, pp. 747–769, 2021. URL https://proceedings.mlsys.org/paper_files/paper/2021/file/0184b0cd3cfb185989f858a1d9f5c1eb-Paper.pdf.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(70):2079–2107, 2010. URL https://www.jmlr.org/papers/v11/cawley10a.html.

Linliang Chen, Yan Zhong, Xin Liu, Sai Li, and Wang Kang. BASC: Behavior-aligned quantization and pruning for low-bit spiking neural networks. *arXiv preprint arXiv:2608.19239*, 2026. URL https://arxiv.org/abs/2608.19239.

Sayeed Shafayet Chowdhury, Isha Garg, and Kaushik Roy. Spatio-temporal pruning and quantization for low-latency spiking neural networks. In *International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9. IEEE, 2021. doi: 10.1109/IJCNN52387.2021.9534111.

Ekin D. Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. AutoAugment: Learning augmentation strategies from data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 113–123, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Cubuk_AutoAugment_Learning_Augmentation_Strategies_From_Data_CVPR_2019_paper.html.

Lei Deng, Yujie Wu, Yifan Hu, Ling Liang, Guoqi Li, Xing Hu, Yufei Ding, Peng Li, and Yuan Xie. Comprehensive SNN compression using ADMM optimization and activity regularization. *IEEE Transactions on Neural Networks and Learning Systems*, 34(6): 2791–2805, 2023. doi: 10.1109/TNNLS.2021.3109064.

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*, 2022. URL https://openreview.net/forum?id=_XNtisL32jv.

Terrance DeVries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. URL https://arxiv.org/abs/1708.04552.

Wei Fang, Zhaofei Yu, Yanqi Chen, Tiejun Huang, Timothée Masquelier, and Yonghong Tian. Deep residual learning in spiking neural networks. *Advances in neural information processing systems*, 34:21056–21069, 2021.

Elias Frantar, Sidak Pal Singh, and Dan Alistarh. Optimal brain compression: A framework for accurate post-training quantization and pruning. *Advances in Neural Information Processing Systems*, 35:4475–4488, 2022.

Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In *International Conference on Learning Representations*, 2016. arXiv:1510.00149.

Simla Burcu Harma, Ayan Chakraborty, Elizaveta Kostenok, Danila Mishin, Dongho Ha, Babak Falsafi, Martin Jaggi, Ming Liu, Yunho Oh, Suvinay Subramanian, and Amir Yazdanbakhsh. Effective interplay between sparsity and quantization: From theory to practice. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=wJv4AIt4sK.

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)*, pp. 770–778, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Yihui He, Ji Lin, Zhijian Liu, Hanrui Wang, Li-Jia Li, and Song Han. AMC: Automl for model compression and acceleration on mobile devices. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 784–800, 2018.

Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*, 2019. URL https://arxiv.org/abs/1911.05248.

Itay Hubara, Yury Nahshan, Yair Hanani, Ron Banner, and Daniel Soudry. Accurate post training quantization with small calibration sets. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4466–4475. PMLR, 2021. URL https://proceedings.mlr.press/v139/hubara21a.html.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 448–456. PMLR, 2015. URL https://proceedings.mlr.press/v37/ioffe15.html.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2704–2713, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Jacob_Quantization_and_Training_CVPR_2018_paper.html.

Minjun Kim, Jaehyeon Choi, Hyunwoo Yang, Jongjin Kim, Jinho Song, and U Kang. Prune-then-quantize or quantize-then-prune? understanding the impact of compression order in joint model compression. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=KWtOTMMvKU.

Youngeun Kim and Priyadarshini Panda. Revisiting batch normalization for training low-latency deep spiking neural networks from scratch. *Frontiers in Neuroscience*, 15:773954, 2021. doi: 10.3389/fnins.2021.773954. URL https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2021.773954/full.

Raghuraman Krishnamoorthi. Quantizing deep convolutional networks for efficient inference: A whitepaper. *arXiv preprint arXiv:1806.08342*, 2018. URL https://arxiv.org/abs/1806.08342.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Duong H. Le and Binh-Son Hua. Network pruning that matters: A case study on retraining variants. In *International Conference on Learning Representations*, 2021.

Bailin Li, Bowen Wu, Jiang Su, and Guangrun Wang. EagleEye: Fast sub-net evaluation for efficient neural network pruning. In *European Conference on Computer Vision*, pp. 639–654, 2020. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123470630.pdf.

Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient ConvNets. In *International Conference on Learning Representations*, 2017a. URL https://arxiv.org/abs/1608.08710.

Hongmin Li, Hanchao Liu, Xiangyang Ji, Guoqi Li, and Luping Shi. Cifar10-dvs: an event-stream dataset for object classification. *Frontiers in neuroscience*, 11:309, 2017b. doi: 10.3389/fnins.2017.00309.

Yaxin Li, Qi Xu, Jiangrong Shen, Hongming Xu, Long Chen, and Gang Pan. Towards efficient deep spiking neural networks construction with spiking activity based pruning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 29063–29073. PMLR, 2024. URL https://proceedings.mlr.press/v235/li24bz.html.

Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. BRECQ: Pushing the limit of post-training quantization by block reconstruction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=POWv6hDd9XH.

Qianhui Liu, Jiaqi Yan, Malu Zhang, Gang Pan, and Haizhou Li. LitE-SNN: Designing lightweight and efficient spiking neural network through spatial-temporal compressive network search and joint optimization. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 3097–3105, 2024. doi: 10.24963/ijcai.2024/343. URL https://www.ijcai.org/proceedings/2024/343.

Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJlnB3C5Ym.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6765–6774. PMLR, 2020. URL https://proceedings.mlr.press/v119/marx20a.html.

Markus Nagel, Mart van Baalen, Tijmen Blankevoort, and Max Welling. Data-free quantization through weight equalization and bias correction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1325–1334, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/papers/Nagel_Data-Free_Quantization_Through_Weight_Equalization_and_Bias_Correction_ICCV_2019_paper.pdf.

Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization. In *International conference on machine learning*, pp. 7197–7206. PMLR, 2020.

Nitin Rathi, Priyadarshini Panda, and Kaushik Roy. STDP-based pruning of connections and weight quantization in spiking neural networks for energy-efficient recognition. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 38(4):668–677, 2019. doi: 10.1109/TCAD.2018.2819366.

Alex Renda, Jonathan Frankle, and Michael Carbin. Comparing rewinding and fine-tuning in neural network pruning. In *International Conference on Learning Representations*, 2020.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y. URL https://link.springer.com/article/10.1007/s11263-015-0816-y.

Yingtao Shen, Minqing Sun, Jianzhe Lin, Jie Zhao, and An Zou. Order of compression: A systematic and optimal sequence to combinationally compress CNN. *arXiv preprint arXiv:2403.17447*, 2024.

Lianfeng Shi, Ao Li, and Benjamin Ward-Cherrier. Spiking brain compression: Exploring one-shot post-training pruning and quantization for spiking neural networks. In *OPT 2025: Optimization for Machine Learning*, 2025. URL https://openreview.net/forum?id=ElKIRZT2Pb.

Gil Shomron and Uri Weiser. Post-training BatchNorm recalibration. *arXiv preprint arXiv:2010.05625*, 2020. URL https://arxiv.org/abs/2010.05625v1.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1409.1556.

Martino Sorbaro, Qian Liu, Massimo Bortone, and Sadique Sheik. Optimizing the energy consumption of spiking neural networks for neuromorphic applications. *Frontiers in Neuroscience*, 14:662, 2020. doi: 10.3389/fnins.2020.00662. URL https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2020.00662/full.

Kuan Wang, Zhijian Liu, Yujun Lin, Ji Lin, and Song Han. HAQ: Hardware-aware automated quantization with mixed precision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8612–8620, 2019.

Wenjie Wei, Yu Liang, Ammar Belatreche, Yichen Xiao, Honglin Cao, Zhenbang Ren, Guoqing Wang, Malu Zhang, and Yang Yang. Q-SNNs: Quantized spiking neural networks. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 8441–8450, 2024.

Wenjie Wei, Malu Zhang, Zijian Zhou, Ammar Belatreche, Yimeng Shan, Yu Liang, Honglin Cao, Jieyuan Zhang, and Yang Yang. QP-SNN: Quantized and pruned spiking neural networks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=MiPyle6Jef.

Yuxin Wu and Justin Johnson. Rethinking "batch" in BatchNorm. *arXiv preprint arXiv:2105.07576*, 2021. URL https://arxiv.org/abs/2105.07576.

Hui Xie, Tong Shi, Haotong Qin, Aishan Liu, Xiaode Liu, and Jinyang Guo. PTQ4SNN: Membrane-aware post-training quantization for spiking neural networks. *arXiv preprint arXiv:2608.07066*, 2026. doi: 10.48550/arXiv.2608.07066. URL https://arxiv.org/abs/2608.07066. Version 2, September 22, 2026.

Jiahui Yu and Thomas S. Huang. Universally slimmable networks and improved training techniques. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Yu_Universally_Slimmable_Networks_and_Improved_Training_Techniques_ICCV_2019_paper.html.

Chenlin Zhou, Liutao Yu, Zhaokun Zhou, Han Zhang, Jiaqi Wang, Huihui Zhou, Zhengyu Ma, and Yonghong Tian. Spikingformer: A key foundation model for spiking neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2236–2244, 2026. doi: 10.1609/aaai.v40i3.37207. URL https://ojs.aaai.org/index.php/AAAI/article/view/37207.

Max Zimmer, Christoph Spiegel, and Sebastian Pokutta. How I learned to stop worrying and love retraining. In *International Conference on Learning Representations*, 2023.

A. Shared experimental protocol

A.1 Cohorts, training and neurons

The table counts independent training runs, with dependent within-run configurations/streams. L: local CIFAR recipe; E: transfer recipes/exceptions.

Table 6: Cohorts, training recipes and selection.

Dataset / model / role	**Models**	**T**	**Recipe / selection**
CIFAR-100 VGG, clipping	12	2	L; internal validation
CIFAR-10 VGG, clipping/compact	12	2	L; internal validation
CIFAR-100 ResNet, crossed	12	2	L; internal validation
CIFAR-100 ResNet, compact/threshold	12	2	L; internal validation
CIFAR-100 ResNet/VGG; CIFAR-10 VGG	12 each	4	L; matched to corresponding T=2 runs, with separate epoch selection
CIFAR-100 ReLU ResNet	12	1	L; training streams matched to compact SNNs
ImageNet-100 SEW-18	6	2	E; internal validation
CIFAR10-DVS VGG	8	10	E; internal validation
CIFAR-100 Spikingformer	8	4	E; internal validation
ImageNet-1k Spikingformer	1 public checkpoint	4	E; external selection uncontrolled

T=4 retraining matches the corresponding local SNN cohorts' T=2 initial weights, training indices, minibatch/augmentation streams, optimizer, loss and 200-epoch schedule. Calibration inputs match across T; temporal BN activations differ. T changes during training and epoch selection, not only during fixed-weight inference. The crossed SNN cohort, the SNN arms used in the ReLU/L1 comparisons, the boundary/tie/epoch analyses, and the ResNet threshold/work follow-ups remain T=2; the ReLU models use T=1.

Wide ResNet-20 uses stem width 64, three three-block stages of widths 128, 256, 512, two convolutions/block, subsampling/channel-padding shortcuts and average-pool/linear readout (He et al., 2016). RGB VGG: eight Conv–BN–LIF layers, widths 64, 128, 256, 256, 512, 512, 512, 512, average pools after layers 2, 4, 6, 8, 2,048-feature readout (Simonyan & Zisserman, 2015). CIFAR (Krizhevsky, 2009) reserves training indices [1,536:5,632] for 4,096 validation images, leaving 45,904 learning images. Calibration [0:1,536] uses retained training; test has 10,000 images. L: Nesterov SGD (momentum 0.9, initial learning rate 0.1, weight decay 0.0005), batch 128, drop-last, 200 cosine-decay epochs, TET-form loss (Deng et al., 2022):

$$L = (1-\lambda)T^{-1}\sum_t \mathrm{CE}(h_t, y) + \lambda \mathrm{MSE}(h, \mu), \quad \mu = 1,\ \lambda = 0.05.$$

CE averages examples; MSE averages batch/time/classes. ReLU uses T = 1. Recipe L training and its T=4 compression use FP32 tensors/deterministic algorithms; TF32 is off for matrix multiplication, on for cuDNN convolutions. Training: crop (padding 4), flip, CIFAR AutoAugment (Cubuk et al., 2019), tensor conversion, normalization, Cutout (side 16) (DeVries & Taylor, 2017). CIFAR-100 mean/std: (0.5071, 0.4867, 0.4408) / (0.2675, 0.2565, 0.2761); CIFAR-10: (0.4914, 0.4822, 0.4465) / (0.2470, 0.2435, 0.2616). Evaluation omits augmentation.

Local CNNs (including SEW/DVS): base threshold 1, decay 0.5, fresh membranes each forward, temporal-average logits; RGB repeats over time:

$$u_t = \tau v_{t-1} + x_t, \quad s_t = \mathbf{1}[u_t > \theta], \quad v_t = (1-s_t)u_t, \quad v_0 = 0.$$

The triangular surrogate is $g_{\gamma_s}(a) = \max(\gamma_s - |a|, 0)/\gamma_s^2$, $a = u_t - \theta$, $\gamma_s = 1$; reset is not detached: $\partial v_t/\partial u_t = (1-s_t) - u_t g_{\gamma_s}(a)$. Native Spikingformer uses $u_t = v_{t-1} + (x_t - v_{t-1})/2$, $s_t = \mathbf{1}[u_t \geq \theta]$, $v_t = (1 - \mathrm{sg}(s_t))u_t$, detached reset sg, zero initial state, sigmoid derivative $4\sigma(4a)[1-\sigma(4a)]$. Threshold: 1 except post-attention LIF, 0.5; membranes reset before/after each minibatch. ReLU has no surrogate/reset; compression/calibration require no backward pass.

A.2 Operators and compact construction

Remark (order classes for threshold controls). Q changes weights and reads masks; A changes only thresholds. Thus threshold-independent Q and A commute, giving at most 4 classes: {APQ}, {QPA}, {QAP, AQP}, {PQA, PAQ}; functions may coincide further. Activation/membrane-calibrated Q or intervening learning requires rechecking.

SVS averages per-example/channel spatial ranks of time-averaged spikes (library relative tolerance), removing round(pC) lowest-scoring channels. Original mask controls and CPU float32 compact ResNet scores use library sorting, without guaranteed index ties; VGG/DVS/Transformer use channel-index ties (C). ANN support-rank averages binary ReLU-support spatial ranks, not magnitudes; shared L1 sums absolute current-prefix weights in float64 with index ties (C). Mask-only and clamped P follow §3.1; clamping excludes shortcut inputs and its repeated recalibration keeps output-clamping hooks active. Fixed-mask references and reselection follow §3.3. Modified P adds pre-scoring recalibration and augmented post-P inputs; original P omits the former and uses unaugmented post-P inputs. Scoring images/kernel are unchanged (D). Removing filters/downstream inputs (Li et al., 2017a) can change our quantizer's downstream ranges. Physical removal slices first-convolution outputs, bias/BN entries and next-convolution inputs, preserving residual/pair output widths. Internal-channel fraction p differs from uniform output masking; Transformer removal is MLP-only. Parameters follow tensor shapes; convolution MACs count T once, excluding classifiers (ResNet sizes: C). Quantized tensors remain floating point, unlike integer-only inference (Jacob et al., 2018); counts are not hardware measurements.

Mean-range ReScaW follows ReScaWConv at commit 01c031bc: `scaling_factor` is c, `clip_val` is 2. Max bounds/zero-row guards are local; linear layers use rows. Masks are reapplied after rounding; Q and clipping follow §3.1 (projection/MLP coverage: B–E). Identity leaves weights unchanged. Frozen bounds slice original FP ranges to retained output rows; recomputed bounds use current support. Q leaves BN affine parameters/buffers unchanged.

All-zero surviving rows map nonzero under positive reference bounds, unlike zero-current-bound rows. The symmetric zero-containing alternative (Krishnamoorthi, 2018) is

$$Q_0(w) = (c/q)\,\mathrm{round}(\mathrm{clip}(qw/c, -q, q)), \quad q = 2^{b-1} - 1.$$

c is per-output max magnitude; ties round to even; zero rows stay zero (scale 1). Biases/BN/activations/thresholds/membranes remain floating point. A multiplies stored base thresholds; ReLU uses no analogue of A. SVS/ReScaW use released components (Wei et al., 2025), not the complete upstream recipe. Retained-weight compensation (Frantar et al., 2022; Shi et al., 2025) and activity-dependent quantizer ranges need separate state-dependence analysis.

A.3 Calibration and state equality

Spatial BN maps (B,T,C,H,W)→(BT,C,H,W), sharing affine parameters and running statistics over time; native BN1d/2d flatten T×B. Epsilon is 10^{-5} and training momentum is 0.1. Normalization uses biased batch variance; running updates use unbiased variance once per minibatch.

Reset-CMA resets BN running means, variances and counters, then processes cached inputs without labels/gradients. Only BN enters training mode (`momentum=None`); other modules remain in evaluation mode, with membranes reset each forward. Evaluation restores BN evaluation mode/original momentum. Weight 1/k at minibatch k averages minibatch means/unbiased variances, without pooled moments or between-batch correction. Upstream calibration batch statistics differ from inference running statistics. Original CIFAR uses 1,536 unaugmented retained-training images, batch 256. For augmented CMA, we cache calibration inputs after applying the model's full training transform once (seed 915000 plus model seed), then reuse them across orders and calls. Original post-P/terminal streams/hooks coincide (§3.2). Compact ResNet augments pre-scoring/post-P/finally; SVS inputs remain unaugmented (C). Transfer inputs/call sites differ (E).

Sequential estimation writes BN input means/unbiased variances over batch/time/spatial axes in execution order, using finalized upstream BN in evaluation mode (Wu & Johnson, 2021). ResNet overwrites without global reset. CIFAR Spikingformer resets before graph discovery; public ImageNet discovers with reference BN, then resets before estimation. For Spikingformer, N counts pooled entries/channel and M2 their centered sum of squares; means and M2/(N−1) are computed in FP64, stored in FP32. Upstream execution and estimation change jointly at fixed z. ResNet requires every BN-only validation loss $\leq$1.5 pp, selects maximum mean correct count (ties: augmented CMA), and tests only that candidate; E's rules differ. Deterministic EMA from common resets also ensures initial-buffer independence; we evaluate reset-CMA. Orders are constructed without terminal-state copying; numerical states, byte hashes and full logits are checked separately. Signed zeros are normalized only in temporary verification copies. Earlier VGG development sometimes scored one representative; matched controls score orders independently. Common deterministic folding preserves equality; folding distinct BN buffers before Q can change z. Logit checks establish equality only on evaluated inputs.

A.4 Selection, statistics and verification scope

Validation is held out before local training and selects the earliest maximum-correct-count epoch. Compact and calibration choices precede test. Historical T=2 VGG clipping lacks compression validation; crossed T=2 tests precede corresponding terminal validation. An archived leave-one-model-out analysis selected one order per construction across configurations using other models' validation counts; C.5 selects within-model candidates. Earlier benchmark use remains. Terminal validation selects PQ/QP (ties: PQ). Baseline references follow §3.3. Qualification follows §3.3 and §4.2; any-order feasibility is weaker. Test does not alter validation qualification.

The primary Student-t bounds (§3.5) use 11 df and Bonferroni adjustment, including all 12 runs. With multiple variability sources (Bouthillier et al., 2021), configurations are averaged

within models and inference conditions on reused images. We use a normal working approximation for run-level independently trained/calibrated outcomes; bounded spreads/zero-inflated regrets need not be normal, so exact small-sample coverage is not guaranteed. C.1 gives SDs; B separates clipping/crossed/ANN families. C.1 records the T=2/T=4 chronology and descriptive extensions. Validation serves epoch selection/qualification; benchmarks/hypotheses are reused (Cawley & Talbot, 2010). CIFAR-100 has 10 exact train–test pairs and one train–validation pair (fixed-prediction exclusions/limits: C.1). E covers uncontrolled public-model selection and reused official-validation partitions.

We verify labels, predictions, counts and retained logits. Primary/ANN/L1 reconstructions use separate slicing, ranges and scores or saved choices; ANN history checks layers without replaying every forward. Initial transfer replay covers all evaluation states and CIFAR Spikingformer validation, using shared forwards; DVS/Transformer reuse masks and BN. IN-100 independently checks SVD scores; rebuilt streams recompute BN with shared primitives. F replays channels on a test-image subset. Manifests record per-array coverage, hashes and shared code; agreement does not establish independent training, calibration or statistical confirmation.

B. Matched-state and ReLU history controls

B.1 VGG clipping

The 8-bit VGG slice crosses p∈{0.01, 0.1}, α∈{1.0, 1.05}, clamped P and FP stem/classifier. Both CIFAR datasets retain T=2 identity/clipping/Q and two clipping-minus-identity spread-reduction contrasts. At each T, four configurations are averaged within each of 12 models. Original T=2 model-level contrasts use nominal 97.5% Student-t intervals; reused models do not test clipping–Q equivalence. Q-minus-clipping: 0.030 pp [-0.078 pp, 0.137 pp] (CIFAR-100), -0.014 pp [-0.123 pp, 0.094 pp] (CIFAR-10), descriptive 95% t intervals (11 df). The table gives mean pre-common-BN spreads (pp); fixed-mask post-BN spreads are zero. Contrast intervals: nominal 97.5% at T=2; descriptive 95% at T=4.

Table 7: VGG clipping controls across datasets and T.

Dataset / T	**Identity (pp)**	**Clipping (pp)**	**Q8 (pp)**	**Clipping−identity (pp) [CI]**
C100 / 2	0.509	66.309	66.339	65.800 [64.925, 66.676]
C100 / 4	0.185	65.059	65.117	64.874 [62.668, 67.080]
C10 / 2	0.806	78.365	78.351	77.559 [76.465, 78.653]
C10 / 4	0.174	77.816	77.786	77.642 [74.936, 80.349]

T=4 Q8-minus-clipping spread-reduction contrasts: 0.0577 [-0.0142, 0.1296] pp (C100), -0.0304 [-0.0738, 0.0130] pp (C10), descriptive 95% intervals. Independently constructed T=4 clipping orders have equal final full states/logits. CPU tensor checks establish numerical-state/execution equality in 96 fixed-mask Q8 groups; raw-byte differences are signed zeros only; evaluation uses the original tensors (verification copies: A.3). Reselected clipping after common BN gives T=4 mean spreads 0.807 pp/0.190 pp (C100/C10), averaging configurations within each model before averaging across models. At T=2, mask reselection/common BN gives clipping spreads 0.460 pp/0.500 pp and original-reference oracle losses 6.960 pp/5.161 pp. CIFAR-10 retains 5/48 instances above δ (maximum 5.230 pp), all over budget. Small means do not bound individual instances.

B.2 Crossed ResNet controls and rounding

With $\Delta S = S_{\text{pre}} - S_{\text{post}}$, define R_Q=ΔS(quantization)−ΔS(identity); replacing Q by direct clipping defines its matched contrast. Mean-minus-max R_Q compares range rules at fixed coverage and pruning–normalization protocol. These spread contrasts are not additive accuracy gains.

The T=2 crossed grid (seeds 700–711; §4.4) averages 8 pruning/bit/threshold combinations/run. Four mean-minus-max R_Q contrasts fix coverage/P protocol. Nominal Student-t intervals (four comparisons, 11 degrees of freedom) condition on reused images/data-informed design, with within-family summaries but no study-wide error control.

Table 8: Range-rule contrasts across ResNet controls.

Coverage	**Pruning/BN**	**Mean−max R_Q, pp [98.75% interval]**
all	mask-only	50.513 [48.727, 52.300]
all	clamped	49.908 [48.093, 51.723]
blocks	mask-only	57.392 [56.009, 58.775]
blocks	clamped	56.777 [55.485, 58.069]

The fixed-mask reduction is given in §4.4; D.1 covers reselected models. Bit-specific C_Q=ΔS(quantization)−ΔS(clipping) has descriptive t intervals outside the four-contrast family.

Table 9: Quantization-minus-clipping contrasts by bit width.

Range / coverage / P	**4-bit C_Q (pp) [95%]**	**8-bit C_Q (pp) [95%]**
max / all / mask-only	2.421 [1.934, 2.909]	0.252 [0.044, 0.460]
max / all / clamped	2.413 [1.945, 2.881]	0.362 [0.226, 0.498]
max / blocks / mask-only	0.562 [0.324, 0.801]	0.206 [0.024, 0.388]
max / blocks / clamped	0.627 [0.384, 0.871]	0.318 [0.199, 0.437]
mean / all / mask-only	-0.616 [-1.057, -0.175]	0.015 [-0.150, 0.180]
mean / all / clamped	-0.555 [-0.975, -0.135]	0.049 [-0.100, 0.198]
mean / blocks / mask-only	-0.026 [-0.155, 0.103]	-0.106 [-0.239, 0.027]
mean / blocks / clamped	-0.064 [-0.217, 0.090]	-0.129 [-0.317, 0.059]

B.3 ReLU normalization history

ReLU shares the ResNet graph/training recipe (A). History compares ANN seeds 900–911 with historical T=2 SNN seeds 700–711 at α=1: separate cohorts, no paired neuronal intervention. Fixed masks cross mean/max range, all/block coverage, mask-only/clamped P, 4/8 bits and identity/clipping/Q under unaugmented cumulative BN; no threshold-scaling analogue. The prespecified SNN>ANN mean-range clipping contrast averages two pruning ratios × two bit widths within 12 models/network. Four Welch 98.75% intervals give nominal Bonferroni family confidence 95%; shared images/earlier SNN-informed choices limit inference. Best/worst are pre-terminal accuracies (%).

Table 10: ANN and SNN spreads under matched clipping controls.

Coverage / P	**ANN best / worst**	**SNN best / worst**	**SNN−ANN spread [98.75% CI]**
all / mask-only	59.508 / 1.313	62.857 / 1.255	3.407 [1.526, 5.288]
all / clamped	59.044 / 1.384	62.204 / 1.230	3.314 [1.464, 5.165]
blocks / mask-only	65.096 / 1.927	69.645 / 2.320	4.156 [2.684, 5.628]
blocks / clamped	64.910 / 2.082	69.167 / 2.245	4.094 [2.694, 5.493]

Rows contain 48 dependent instances/network. Here, orders ending in P have strictly higher pre-recalibration accuracy; common recalibration preserves their logits and equalizes logits across orders. The near-chance interpretation follows §4.5; chance (1%) is no mathematical floor. The 8-bit all/mask-only example: ANN best/worst 60.354/1.350%, SNN 63.407/1.310%; worst rises to unchanged best. Uncompressed accuracy: 82.022→80.123%

(ANN), 76.614→73.773% (SNN). Augmented preservation controls cannot validate this unaugmented map.

Data-informed max-range comparisons reverse direction across all four coverage/P settings: ANN mean spread exceeds SNN at 4 bits; SNN exceeds ANN at 8 bits. Minimum worst accuracies: 67.82%/63.62%. Terminal spreads/logit differences are zero; identity at α=1 is zero before/after.

C. Compact compression and feasibility

C.1 ResNet temporal comparison and original primary

Primary chronology. These private records are not public preregistration. Earlier benchmark results and augmented-CMA selection on crossed seeds 700–711 informed primary seeds 900–911, outside nominal family control. All times are KST, 2026.

Table 11: Protocol chronology and fixed decisions.

Cohort	Records and fixed decisions
T=2 primary	Sep 15: protocol 16:30 (training 16:31), statistics plan 16:45, validation gate 21:27. Cohort/configuration below, δ=1.5 pp and 5-pp budget fixed. Mean S/regret receive two nominal one-sided 97.5% t bounds, all runs included; signed/bootstrap descriptive. The plan clarifies initial bound wording before endpoints.
T=4 extension	Sep 24 23:36: mirrored grid/gates specified after T=2 results and T=4 baseline training. Sep 25 00:41: validation lock. Descriptive; no new confirmatory family.

For both cohorts, validation counts, BN-screen results and choices were fixed before test-evaluation jobs started. These private records do not independently certify timestamps or the exact first forward pass. The records are `fresh_compact_20260915/{protocol, statistics_plan, validation_gate}`, `t4_evaluation_20260924/{protocol, validation_lock}`. The 5-pp budget is operational, not universal. Primary T=2 training used four NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs (0–3), one per fit, at most four concurrent primary jobs: 4.80 controller wall-hours versus 19.12 summed model-hours including validation/checkpoint work. Other host activity was not recorded.

Twelve ResNets per T (900–911; matched initial weights) cross p∈{0.10,0.25}, b∈{4,8}, fixed/reselected masks and frozen/current max ranges, with identity A and block-convolution Q. T=4 mirrors the prespecified T=2 primary: p=.10/4-bit, reselected/recomputed. Slicing preserves residual widths (A.2). Fixed masks use augmented-recalibrated FP; reselection uses augmented-recalibrated prefixes and unaugmented SVS. Shared cached augmentation and selection follow A.3–A.4. F/R denote fixed/reselected masks. Spread cells give mean/maximum terminal test gaps in pp; pass cells give both-order (validation, test) counts out of 12 within 5.0 pp of original, for both range rules.

Table 12: Compact ResNet spreads and budget passage across T.

p / bits / mask	T=2 frozen S (pp)	T=2 current S (pp)	T=2 pass	T=4 frozen S (pp)	T=4 current S (pp)	T=4 pass
0.10/4/F	0.000/0.00	0.068/0.20	(12, 12)	0.000/0.00	0.058/0.14	(12, 12)
0.10/4/R	0.153/0.41	0.1675/0.35	(12, 12)	0.137/0.34	0.153/0.32	(12, 12)
0.10/8/F	0.000/0.00	0.026/0.04	(12, 12)	0.000/0.00	0.074/0.22	(12, 12)
0.10/8/R	0.131/0.30	0.137/0.34	(12, 12)	0.111/0.30	0.128/0.31	(12, 12)
0.25/4/F	0.000/0.00	0.228/0.57	(0, 1)	0.000/0.00	0.238/0.81	(0, 0)

p / bits / mask	T=2 frozen S (pp)	T=2 current S (pp)	T=2 pass	T=4 frozen S (pp)	T=4 current S (pp)	T=4 pass
0.25/4/R	0.235/0.68	0.318/0.69	(0, 0)	0.341/0.78	0.295/0.66	(0, 0)
0.25/8/F	0.000/0.00	0.154/0.40	(0, 1)	0.000/0.00	0.188/0.38	(0, 0)
0.25/8/R	0.208/0.46	0.198/0.43	(0, 1)	0.221/0.51	0.187/0.63	(0, 0)

Standalone controls. We report twelve-model means; compressed endpoints use final augmented reset-CMA, p=.10 for P and 4-bit max-range block Q. Losses reference same-model/split original or BN-only accuracy. The T=2 Q-only control comes from the later threshold–work evaluation and matches the primary checkpoints, calibration streams, quantization metadata and calibrated QP-prefix states. This descriptive comparison does not change the prespecified primary family.

Table 13: Matched T=2 standalone controls.

T=2 state	Validation accuracy (%)	Test accuracy (%)	Test loss: original / BN (pp)
Original FP	77.110	76.760	0.000 / -0.055
FP + BN	76.562	76.705	0.055 / 0.000
P only + BN	75.146	75.460	1.300 / 1.245
Q only + BN	76.064	76.487	0.273 / 0.218
PQ + BN	74.615	75.180	1.580 / 1.525
QP + BN	74.713	75.162	1.597 / 1.542

Q-only mean loss is below P-only on both splits, yet T=2 test slightly favors PQ; no prospective weaker-first test. T=4 Q-only was not evaluated. T=4 original/BN-only test means: 79.322%/79.408%; P-only p=.10/.25: 78.148%/72.800%. T=2 p=.25 P-only gives 70.534%. T=4 representative mean worst-order test accuracy: 77.835%. Maximum BN-only losses (validation/test), T=2/T=4: 1.221 pp/0.460 pp and 1.147/0.170 pp. At p=.10, parameters fall 17.10→15.39 million. Every primary/representative pair differs in masks, common-coordinate weights and logits (§4.1). T=4 mean correctness/mask disagreement is 3.997%/1.376%. Fixed-mask/fixed-range final logits coincide at T=2, states/logits at T=4. The MAC scope is defined in §4.1.

Descriptive two-way bootstrap uses 2,000 model-by-image draws, fixed validation choices and image indices shared across both orders/all resampled models, recomputing within-run absolute gaps/regrets. S/regret 97.5th percentiles: 0.3542 pp/0.2025 pp at T=2, 0.3484/0.1150 pp at T=4. These are descriptive resampling summaries with fixed validation choices and differ from the fixed-image run limits.

Sample SDs (S/regret, pp; n=12, ddof=1): (0.1019, 0.0715)/(0.0818, 0.0318) at T=2/4. Validation selects the test-best order in 6/12 and 11/12 runs (zero regret). Always-PQ/QP/uniform regrets: (0.0750 pp, 0.0925 pp, 0.08375 pp) at T=2, (0.1408, 0.0125, 0.0767) pp at T=4. T=2 pre-terminal S: 0.347 pp (maximum 0.93 pp). The table gives descriptive both-order (validation, test) counts out of 12; the 5-pp criterion stays fixed.

Table 14: Compact ResNet budget sensitivity across T.

Budget	1.0 pp	1.5 pp	2.0 pp	2.5 pp	3.0 pp	5.0 pp
T=2	(0, 0)	(0, 3)	(1, 11)	(6, 12)	(9, 12)	(12, 12)
T=4	(0, 1)	(2, 6)	(6, 12)	(9, 12)	(12, 12)	(12, 12)

For T=4 ResNet with reselected masks and recomputed ranges (p=.10, 4-bit), the (both, any, exactly-one) counts at a budget of 2.0 pp are (6, 8, 2) on validation and (12, 12, 0) on

test. At 5.0 pp, the counts are (12, 12, 0) on validation and (12, 12, 0) on test. Counts are out of twelve models, with the original BN criterion unchanged.

Exact-duplicate exclusion. Removing 10 matched test images from fixed T=2 predictions changes mean spread/upper bound 0.1675/0.2322→0.1693/0.2345 pp; both-order passage remains 12/12. Removing the matched validation image preserves joint qualification 12/12. With T=4 exclusions, spread/upper bound: 0.1568/0.2092 pp; test passage 12/12, validation qualification 12/12. Models/validation choices stay fixed; absent per-epoch predictions leave epoch-selection effects unknown, and exclusion does not bound training influence (A.4). Twelve-run mean (u_i, v_i) (§4.1): (186.00, 184.25)/(193.42, 206.25) at T=2/4, N=10,000 each.

Paired correctness counts. Each cohort has twelve models and N=10,000 test images per model; u/v count PQ-only/QP-only correctness. The T=2 primary and descriptive T=4 ResNets use p=.10/4-bit, reselected masks/current ranges; the descriptive T=1 support-rank ANN uses the same compression setting (C.4). Matching seed labels align separate cohorts in the table, not paired neuronal interventions.

Table 15: Paired correctness counts for separate ResNet and ReLU cohorts.

Seed	T=2 correct PQ / QP	T=2 u / v	T=4 u / v	ANN u / v
900	7,541 / 7,550	225 / 234	167/176	273/291
901	7,498 / 7,516	240 / 258	232/246	275/334
902	7,477 / 7,463	97 / 83	200/228	197/204
903	7,496 / 7,489	181 / 174	192/224	207/249
904	7,543 / 7,536	148 / 141	273/279	224/291
905	7,510 / 7,545	191 / 226	202/213	246/282
906	7,577 / 7,545	353 / 321	170/187	176/142
907	7,525 / 7,513	168 / 156	183/203	174/233
908	7,553 / 7,537	72 / 56	177/184	198/253
909	7,425 / 7,453	77 / 105	180/188	157/209
910	7,566 / 7,547	127 / 108	182/167	219/171
911	7,505 / 7,501	353 / 349	163/180	162/181

Under a conditional equal-accuracy null with independent image pairs, $U_i \mid m_i \sim \text{Binomial}(m_i, 1/2)$, where $m_i = u_i + v_i$. Mean exact $E[100|2U_i - m_i|/N]$ across the twelve T=2 observed m values is 0.1489 pp: a sampling reference, not a lower bound, debiasing constant or equivalence result. T=4 reference: 0.1591 pp; QP wins 11/12 versus 4/12 at T=2. Small magnitude can coexist with consistent direction; these descriptive counts do not test cross-T interaction.

C.2 Compression boundary, ties and checkpoint sensitivity

The T=2 six-cell follow-up varies pre-scoring recalibration (No/Yes), retaining 4-bit max-range block Q, reselected masks/current ranges and augmented post-P/final BN. p=.10/Yes exactly replays the primary. S/worst-loss entries give twelve-model mean/maximum terminal test values (pp), losses relative to the uncompressed baseline; passage counts give both-order (validation, test) totals.

Table 16: Compact ResNet pruning and pre-scoring calibration.

p	Pre-scoring recalibration	S (pp)	Worst loss (pp)	Pass
0.10	No	0.223 / 0.66	1.710 / 2.29	(12, 12)
0.10	Yes	0.1675 / 0.35	1.673 / 2.48	(12, 12)
0.15	No	0.219 / 0.74	2.878 / 3.63	(12, 12)
0.15	Yes	0.141 / 0.48	2.821 / 3.65	(12, 12)
0.20	No	0.285 / 0.67	4.367 / 5.49	(2, 10)
0.20	Yes	0.260 / 0.50	4.313 / 5.05	(3, 11)

These reused-cohort comparisons are descriptive. Pre-scoring recalibration lowers each ratio's mean spread, already below δ; p=.15 worst/oracle mean-accuracy changes +0.0575 pp/-0.0208 pp show nonuniform improvement. Test passage cannot reverse p=.20 validation failure. Another follow-up crosses these ratios with library/lexicographic (score, original-index) ties and validation-selected/final-epoch (200) checkpoints; no retraining/primary replacement. Cutoff ties affect 147/216 primary decisions. At p=.20, selected-epoch validation passes 3/12 (library), 4/12 (index); final-epoch passes 9/12 with either rule, mean spreads 0.220/0.300 pp. Tie contributions/epoch-selection bias remain unidentified; independent qualification requires a reserved split (Cawley & Talbot, 2010).

Behavioral sensitivity. Both panels use the same twelve T=2 p=.10/4-bit checkpoints; every construction passes BN screening and both-order budgets on both splits. Entries are means across models on test images, with sample SDs for disagreements and all panel B metrics. In panel A, pre-scoring is on unless stated; the first three rows reselect masks/recompute ranges, and the others use library ties with the named component fixed. Here d is signed PQ−QP accuracy, S is the within-model absolute gap (pp), and mask disagreement divides by original scored channels. Panel B holds order fixed: the signed accuracy difference is library−index ties or pre-scoring on−off (pp).

Table 17: Behavior under construction and implementation changes.

(a) Across compression orders

Construction	Predictions differing (%)	Correctness differing (%)	d (pp)	S (pp)	Masks differing (%)
Library ties	5.971 (3.019)	3.703 (1.874)	0.0175	0.1675	1.699
Index ties	5.872 (2.966)	3.611 (1.810)	0.0042	0.1608	0.570
Pre-scoring off	5.908 (1.200)	3.721 (0.720)	-0.0058	0.2225	2.908
Fixed mask/range	0.000 (0.000)	0.000 (0.000)	0.0000	0.0000	0.000
Fixed mask, current range	1.265 (0.673)	0.743 (0.419)	0.0108	0.0675	0.000
Reselected mask, frozen range	5.870 (3.086)	3.658 (1.938)	0.0067	0.1533	1.699

(b) Within the same compression order

Same-order comparison	Predictions differing (%)	Correctness differing (%)	Signed accuracy difference (pp)
PQ Library − index	2.092 (3.497)	1.266 (2.138)	0.016 (0.106)
QP Library − index	2.924 (4.081)	1.744 (2.462)	0.003 (0.094)
PQ On − off	6.278 (3.128)	3.867 (1.928)	0.022 (0.201)
QP On − off	6.902 (2.795)	4.312 (1.673)	-0.002 (0.280)

Mean same-order pre-scoring prediction disagreement is comparable to or larger than PQ–QP; tie-change means are smaller. Similar rates need not identify the same images. Mean per-model Jaccards (intersection/union) of PQ–QP disagreement sets: 0.691, 0.198, 0.196 for library/index, library/off, index/off. Per-run intersections/unions and candidate losses are retained; neither pure order contributions nor universal saturation are identified.

Without stored sample IDs, alignment is reconstructed from pinned deterministic loaders/transforms, official CIFAR checksums/partition indices and exact primary state/logit/prediction anchors—not labels alone or independent per-image execution logs. CIFAR calibration-stream sensitivity remains unmeasured.

C.3 VGG with a zero-containing grid

Matched T=2/T=4 CIFAR-10 VGGs slice outputs/BN/next-inputs at pairs (0,1), (3,4), (6,7), (9,10), widths 64/256/512/512, preserving pair-output/classifier-input widths. Per-channel affine fake Q covers seven non-stem convolutions: scale max|W|/7, zero point 0,

bounds $-7 \ldots 7$, zero-row scale 1; activations/biases/BN/classifier stay floating point, without folding or packed/integer execution (Krishnamoorthi, 2018). SVS: initial 1,536 unaugmented retained-training images, channel-index ties. Pre-scoring/post-P/final BN share cached augmentation (seed 915000+model seed); validation selection/qualification precedes test. Each T crosses five ratios with dynamic (reselected/current) and fixed (shared/original) masks/ranges. S: dynamic mean/maximum test spread; loss: mean worst-order versus same-T original. Both-order validation/test passage: 5.0 pp, denominator 12. Same-T MAC savings accompany doubled T=4 dense counts; fixed final logits coincide.

Table 18: Compact VGG compression results across T.

p / T	MAC saved (%)	S (pp)	Loss (pp)	Dyn. both	Dyn. any	Fixed both
0.10 / 2	9.96	0.000 / 0.000	0.125	(12, 12)	(12, 12)	(12, 12)
0.10 / 4	9.96	0.072 / 0.200	0.188	(12, 12)	(12, 12)	(12, 12)
0.20 / 2	19.97	0.057 / 0.160	0.236	(12, 12)	(12, 12)	(12, 12)
0.20 / 4	19.97	0.077 / 0.170	0.558	(12, 12)	(12, 12)	(12, 12)
0.30 / 2	30.03	0.096 / 0.280	0.871	(12, 12)	(12, 12)	(12, 12)
0.30 / 4	30.03	0.129 / 0.270	1.667	(12, 12)	(12, 12)	(12, 12)
0.40 / 2	40.04	0.139 / 0.390	1.773	(12, 12)	(12, 12)	(12, 12)
0.40 / 4	40.04	0.243 / 0.440	3.482	(11, 12)	(12, 12)	(11, 11)
0.50 / 2	50.00	0.158 / 0.420	3.926	(7, 12)	(11, 12)	(11, 12)
0.50 / 4	50.00	0.306 / 0.530	6.543	(0, 1)	(0, 2)	(0, 2)

T=2 original/augmented/Q-only test accuracy: 93.26%/93.27%/93.14%. At p=.40, worst accuracy 91.49%, masks differ in 12/12 pairs, prediction disagreement 3.032%: distinct models, no new confirmatory cohort. T=4 original/augmented means rise to 94.081%/94.054%, but p=.40 worst accuracy falls to 90.598% (disagreement 3.691%). The seed 508 qualification boundary is discussed in §5.3; D.2 gives its paired losses, slacks and SEs. Tighter-budget dynamic (validation, test) both-order counts/12 do not replace the 5-pp screen.

Table 19: Compact VGG budget sensitivity across T.

p / T	1.0 pp	1.5 pp	2.0 pp	2.5 pp	3.0 pp
0.10 / 2	(11, 12)	(12, 12)	(12, 12)	(12, 12)	(12, 12)
0.10 / 4	(12, 12)	(12, 12)	(12, 12)	(12, 12)	(12, 12)
0.20 / 2	(9, 12)	(12, 12)	(12, 12)	(12, 12)	(12, 12)
0.20 / 4	(4, 12)	(11, 12)	(12, 12)	(12, 12)	(12, 12)
0.30 / 2	(1, 8)	(9, 12)	(12, 12)	(12, 12)	(12, 12)
0.30 / 4	(0, 0)	(1, 4)	(5, 10)	(8, 11)	(11, 12)
0.40 / 2	(0, 0)	(0, 3)	(1, 8)	(5, 12)	(12, 12)
0.40 / 4	(0, 0)	(0, 0)	(0, 0)	(0, 1)	(1, 3)
0.50 / 2	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(0, 0)
0.50 / 4	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(0, 0)

For dynamic T=4 VGG at p=.40, the (both, any, exactly-one) counts at a budget of 2.0 pp are (0, 0, 0) on validation and (0, 0, 0) on test. At 5.0 pp, the counts are (11, 12, 1) on validation and (12, 12, 0) on test. Counts are out of twelve models, with the original BN criterion unchanged.

In the dynamic T=4 VGG construction at p=.50 and a budget of 5.0 pp, 1 model passes for both orders on test and 2 pass for at least one order; 0 pass for either order on validation, out of twelve models.

T=4 VGG test losses after BN: Q-only 0.141 pp; P-only p=.40/.50 3.139/6.179 pp; joint worst 3.482/6.543 pp. Substantial pruning coexists with mild standalone Q damage. These original-reference losses share calibration and are not additive causal contributions.

C.4 ReLU and shared-score compact comparisons

Seed-matched ANN/T=2 SNNs cross four mask/range cells at p=.10/4-bit block-only max range, using support-rank/SVS (A.2). Later descriptive L1 arms were specified before evaluation; SNN evaluation did not depend on ANN results. L1 sums first-convolution filters' absolute current weights, including quantized QP weights (FP32 weights accumulated in float64, index ties) (Li et al., 2017a). BN leaves L1 unchanged; pre-scoring BN preserves procedural consistency. Validation precedes test. Retraining changes pruning conclusions (Liu et al., 2019; Renda et al., 2020; Le & Hua, 2021; Zimmer et al., 2023), without establishing eval-only feasibility. Entries: 12-model terminal-test means, S/worst-order original-relative loss in pp. Activity-score SNN passes both splits/every cell; ANN passes validation 0/12, test 1/12. Shared L1 passes both splits 12/12, every cell/network.

Table 20: ANN and SNN compact results by pruning score.

Mask / range	**Activity S: ANN / SNN**	**ANN activity loss**	**L1 S: ANN / SNN**	**L1 loss: ANN / SNN**
Fixed / fixed	0.0000 / 0.0000	6.834	0.000 / 0.000	1.411 / 0.963
Fixed / recomputed	0.0750 / 0.0675	6.847	0.047 / 0.061	1.422 / 0.982
Reselected / fixed	0.4533 / 0.1533	6.897	0.108 / 0.153	1.440 / 1.040
Reselected / recomputed	0.4133 / 0.1675	6.852	0.127 / 0.162	1.437 / 1.033

ANN original accuracy: 82.022%; augmented BN/P-only/Q-only test losses: -0.015/6.151/0.444 pp. All pass BN screening; P-only meets 5.0 pp in 2/12 per split. Losses imply no additive causal shares. Support-rank never ties all channels in a scored layer, but saliency remains unvalidated. Reselected/recomputed ANN signed PQ−QP mean: -0.277 pp [-0.513, -0.040] (descriptive 95% run-level t interval). The paired counts appear in C.1.

Reselected/recomputed L1 ANN/SNN mask Hamming: 2.040%/0.496%; prediction disagreement: 3.973%/3.168%. Score choice changes feasibility, but separate learned weights, time steps and baseline accuracies prevent neuronal causal attribution.

C.5 Accuracy regret versus feasible compression choice

Retrospective 4-bit policies use reselected masks/current ranges and original FP fallback. Each candidate is eligible only if its own order's validation loss is within β and the common BN gate passes; all candidates pass this gate. Policies maximize convolution-MAC saving among eligible candidates. Ties favor higher validation correct count, then PQ. Test evaluation keeps these choices, including fallbacks. ResNet grids use p∈{.10,.15,.20} with pre-scoring BN on/off at T=2, and p∈{.10,.25} with BN on at T=4. Both VGG grids use p∈{.10,.20,.30,.40,.50} and C.3's procedure.

Search takes each model's fixed-PQ/QP candidate union: $G_{i,\mathrm{search}}(\beta) = \max\{G_{i,\mathrm{PQ}}(\beta), G_{i,\mathrm{QP}}(\beta)\}$, G=0 for FP fallback. Mean MAC saving versus uncompressed exceeds/matches the larger observed fixed-policy mean by 0–0.857 percentage points; that reference is retrospective per setting–budget. Test losses/violations are separate; reused grids/images establish no independent qualification. Fixed-PQ/QP means differ by up to 2.51 percentage points. T=4 VGG favors PQ at 2/3-pp budgets, QP at 5.0 pp; this descriptive difference is separate from search saving.

Each entry gives **mean convolution-MAC saving (%) / test-budget violation count / full-precision fallback count**, with twelve models per row. R2-off/R2-on denote T=2

ResNet with pre-scoring off/on; R4-on denotes T=4 ResNet with pre-scoring on. V2/V4 denote T=2/T=4 VGG under C.3. Setting labels repeat on each row.

Table 21: Compression policy savings, violations and fallbacks.

Setting	Budget (pp)	Always PQ	Always QP	Validation search
R2-off	2	0.840 / 0 / 11	1.681 / 0 / 10	2.521 / 0 / 9
R2-off	3	8.365 / 0 / 3	10.046 / 1 / 1	10.447 / 1 / 1
R2-off	5	16.611 / 0 / 0	17.468 / 0 / 0	18.325 / 0 / 0
R2-on	2	0.840 / 0 / 11	2.521 / 0 / 9	2.521 / 0 / 9
R2-on	3	9.607 / 1 / 2	10.008 / 1 / 2	10.848 / 1 / 1
R2-on	5	17.040 / 1 / 0	17.468 / 0 / 0	18.325 / 1 / 0
V2	2	31.697 / 0 / 0	31.697 / 0 / 0	32.531 / 0 / 0
V2	3	40.041 / 0 / 0	40.041 / 0 / 0	40.041 / 0 / 0
V2	5	46.680 / 0 / 0	48.340 / 0 / 0	49.170 / 0 / 0
R4-on	2	5.882 / 0 / 5	5.882 / 0 / 5	6.723 / 0 / 4
R4-on	3	10.084 / 0 / 0	10.084 / 0 / 0	10.084 / 0 / 0
R4-on	5	10.084 / 0 / 0	10.084 / 0 / 0	10.084 / 0 / 0
V4	2	26.676 / 0 / 0	24.162 / 0 / 0	26.676 / 0 / 0
V4	3	31.697 / 0 / 0	30.025 / 0 / 0	31.697 / 0 / 0
V4	5	39.206 / 0 / 0	40.041 / 0 / 0	40.041 / 0 / 0

For same-baseline/split losses $L_{\mathrm{PQ}}, L_{\mathrm{QP}}$, exactly-one feasibility is $1[\min(L_{\mathrm{PQ}}, L_{\mathrm{QP}}) \le \beta < \max(L_{\mathrm{PQ}}, L_{\mathrm{QP}})]$: any-order minus both-order. The width of this budget interval equals the within-run accuracy spread; discrete candidate changes can yield larger resource differences. C.1/C.3 give the four T=4 examples with their split-specific counts. Zero exactly-one counts imply neither good accuracy nor equivalence.

The analysis record retains complete budget curves, exactly-one intervals, model choices, split losses/slacks, fallbacks and violations.

D. Calibration robustness and paired accuracy

D.1 Calibration candidates and construction response

The crossed T=2 ResNets (700–711) use A.3's prespecified screening before test on reused models/images. Candidates share training indices; augmented inputs are cached per model. All twelve fail unaugmented-BN tolerance (§3.4).

Table 22: ResNet calibration candidate screening.

BN procedure	VAL accuracy	Max. VAL loss	TEST accuracy	Max. TEST loss
Uncompressed model	77.126%	0	76.614%	0
Unaugmented cumulative	73.153%	4.785 pp	73.773%	3.930 pp
Augmented cumulative	76.674%	0.903 pp	76.848%	0.280 pp
Sequential unaugmented	—	4.761 pp	Not evaluated	—

Sequential calibration fails with mean validation loss 4.144 pp and is not tested. Unaugmented test results are existing controls; selected augmented calibration changes mean test accuracy by +0.234 pp. The complete grid averages 8 configurations/model in eight settings within 12 runs, retaining severe degradation. U/A denote unaugmented/augmented final maps; M/C denote mask-only/clamped P. Fixed-mask cells give terminal accuracy (%) U→A, identical across orders; reselected cells give mean oracle accuracy (%)/spread (pp).

Table 23: ResNet accuracy and spread across calibration maps.

Coverage / P	Fixed U→A	Orig. P/U	Orig. P/A	Recal. P/U	Recal. P/A
Max range					
all / M	61.41→63.72	61.74/0.617	64.11/0.775	61.85/0.621	64.00/0.555
all / C	60.82→62.91	61.31/0.797	63.51/0.881	61.38/0.711	63.30/0.580
blocks / M	62.08→64.40	62.37/0.564	64.84/0.746	62.54/0.613	64.73/0.575
blocks / C	61.45→63.70	61.98/0.791	64.28/0.870	62.03/0.727	64.04/0.574
Mean range					
all / M	45.71→43.32	46.35/1.301	44.05/1.431	45.97/0.819	43.55/0.966
all / C	45.20→42.81	46.10/1.568	43.79/1.630	45.45/0.994	43.02/1.221
blocks / M	54.01→53.67	54.52/1.399	54.44/1.587	54.36/0.661	54.01/0.807
blocks / C	53.36→52.82	54.13/1.513	53.85/1.642	53.87/0.874	53.23/0.915

§5.2/A.2 describe the intervention/directions. Fixed-mask max-range accuracy exceeds mean-range under both maps despite zero gaps.

D.2 Paired accuracy accounting

For the complete IN-100 cohort (§5.1), test mean worst/best accuracy changes 71.890/73.230%→72.890/73.247%, and spread changes 1.340→0.357 pp. The original IN-100/DVS/CIFAR Spikingformer grids contain 22 models and 204 dependent model–configuration–split pairs; §5.1's identity holds in integer correct counts throughout. This descriptive summary includes dynamic/recomputed and fixed/frozen constructions, excluding later IN-100 streams, selected Transformer calibration and public IN-1k. Shrinking-gap counts distinguish higher worst (W up) from lower best (B down), nonoverlapping here; pairs are not independent replications.

Table 24: Paired gap changes across transfer cohorts.

Cohort / split	Pairs	Gap smaller / same / larger	Shrinking: W up / B down
IN-100 / VAL	30	25 / 1 / 4	20 / 5
IN-100 / TEST	30	25 / 0 / 5	21 / 4
DVS / VAL	40	27 / 2 / 11	20 / 7
DVS / TEST	40	24 / 3 / 13	12 / 12
Spikingformer, crop-CMA / VAL	32	23 / 1 / 8	11 / 12
Spikingformer, crop-CMA / TEST	32	25 / 0 / 7	17 / 8

Examples pass original-reference budgets at both phases and the validation BN criterion. IN-100: p=0.05, 4 bits, N=5000. DVS: fixed/frozen seed 3001, p=0.10, 4 bits; dynamic seed 3002, p=0.20, 8 bits; N=1000. The last column tracks the previously best order, not re-ranked extrema.

Table 25: Correctness counts and extrema before and after recalibration.

Case	Correct counts (PQ, QP): pre → terminal	ΔW / ΔB (pp)	Change for pre-BN best order (pp)
IN-100 4004 TEST	(3621, 3700) → (3663, 3700)	+0.84 / +0.00	QP +0.00
DVS 3001 TEST	(801, 819) → (819, 819)	+1.80 / +0.00	QP +0.00
DVS 3001 VAL	(827, 819) → (819, 819)	+0.00 / -0.80	PQ -0.80
DVS 3002 TEST	(800, 786) → (790, 786)	+0.00 / -1.00	PQ -1.00

IN-100 pre-BN spread 1.58 exceeds δ=1.5 by 4 correct predictions: an observed boundary, not a population-effect test. DVS validation/dynamic-test rows retain best-order deterioration. Equal correct counts establish accuracy agreement; state/logit equality needs separate checks (A).

Paired losses near the budget boundary. For IN-100 seed 4003 validation, $(N, K_{\mathrm{FP}}) = (5{,}000, 3{,}962)$; for T=4 VGG seed 508, p=.40 validation/test, $(N, K_{\mathrm{FP}}) = (4{,}096, 3{,}892)/(10{,}000, 9{,}431)$. K_{FP} counts original FP correct predictions. If a/b count FP-only/compressed-only correctness, both $= K_{\mathrm{FP}} - a$, neither $= N - K_{\mathrm{FP}} - b$, loss is $100(a-b)/N$ pp and slack is 5−loss. FP accuracy is $100K_{\mathrm{FP}}/N\%$; endpoint accuracy is FP accuracy minus loss. IN ref is stream 0; final1/rebuild1 are stream 1 final-only/full reconstruction.

Table 26: Paired loss and budget slack at selected boundaries.

Endpoint / order	Only FP / compressed	Loss / slack (pp)	SE (pp)
IN ref PQ	392 / 184	4.160 / 0.840	0.476
IN ref QP	394 / 179	4.300 / 0.700	0.475
IN final1 PQ	399 / 168	4.620 / 0.380	0.472
IN final1 QP	411 / 171	4.800 / 0.200	0.478
IN rebuild1 PQ	403 / 159	4.880 / 0.120	0.469
IN rebuild1 QP	434 / 162	5.440 / -0.440	0.482
VGG val PQ	271 / 55	5.273 / -0.273	0.433
VGG val QP	244 / 50	4.736 / 0.264	0.412
VGG test PQ	635 / 139	4.960 / 0.040	0.274
VGG test QP	622 / 155	4.670 / 0.330	0.275

For signed imagewise correctness Z=FP−compressed, l=mean(Z), r=mean(Z^2), SE $= 100\sqrt{(r-l^2)/(N-1)}$ pp is descriptive under a fixed-pair iid-image working approximation. It neither corrects epoch/candidate selection nor certifies population budgets; the SE does not alter the qualification criterion. Seed 4003 validation stream 1 has direct rebuild-minus-final-only loss differences PQ 0.26 pp (SE 0.355), QP 0.64 pp (SE 0.399). IN rebuild QP/VGG validation PQ remain failures despite passing test counterparts. The analysis retains all six IN models, streams and split-specific contingencies.

E. Image, event and Transformer extensions

These descriptive extensions reuse benchmarks under A.4's joint validation criteria; configurations/streams are dependent. Dynamic means reselected masks/recomputed ranges; PQ/QP excludes A/intervening training. Tables give model-mean worst-order original-relative loss and S pre/final means/final maximum. Dense MACs count T once; coverage is cohort-specific.

E.1 ImageNet-100 and calibration streams

CMC (`HobbitLong/CMC`) contains 126,689 native-resolution ILSVRC2012 (Russakovsky et al., 2015) training/5,000 official-validation images. Before training, 50/class form 5,000 internal-validation images; removing nine training files byte-identical to validation/evaluation leaves 121,680 training images. Evaluation reuses earlier official validation. Six SEW-ResNet-18s (Fang et al., 2021) (4000–4005): SEW-ADD, T=2, A.1 neurons/TET; 200 epochs, batch 256/drop-last, cosine Nesterov SGD (rate 0.1, momentum 0.9, weight decay 0); AMP FP16 training, FP32 validation/compression, TF32 off. Training uses random resized crop 224 (scale 0.08–1)/flip; evaluation resize 256/center crop 224; ImageNet mean/std (0.485,0.456,0.406)/(0.229,0.224,0.225).

Internal residual-channel P preserves block-output dimensions. SVS uses 1,536 strided retained-training images/evaluation transforms and CPU float32 default `torch.argsort` (index ties unguaranteed). Zero-free max-range Q covers sixteen main-path convolutions;

stem/projections/classifier stay full precision. Cached training crops/flips (seed 915000 + model seed), batch 256, drive augmented CMA pre-scoring/post-P/finally. Fixed/frozen uses the common augmented FP reference. All pass BN preservation; original validation/test accuracy 79.280/76.563%, after BN 78.507/76.453%.

Table 27: ImageNet-100 compact compression results.

p / bits	Conv. MAC saved (%)	VAL joint / TEST budget	TEST S: pre → final / max (pp)	Worst loss (pp)
0.05 / 4	4.507	6/6 / 6/6	1.340 → 0.357 / 0.740	3.673
0.05 / 8	4.507	6/6 / 6/6	0.247 → 0.147 / 0.340	3.267
0.10 / 4	9.145	0/6 / 0/6	1.500 → 0.367 / 0.740	7.160
0.10 / 8	9.145	0/6 / 0/6	0.467 → 0.367 / 0.800	6.773
0.10 / 4 fixed/frozen	9.145	0/6 / 0/6	1.163 → 0.000 / 0.000	7.187

Feasible compression saves 4.507% all-convolution MACs. At p=0.10, P-only test accuracy is 69.830%; matched-BN references do not restore feasibility. Fixed/frozen states/logits coincide despite failure. At p=0.05, prediction disagreement is 14.670/12.830% (4/8-bit). Seed 4004 exceeds the pre-BN spread margin within budget (D.2).

Stream follow-ups. At p=0.05/4-bit, final-only varies cached calibration inputs with non-BN state fixed; rebuild uses them pre-scoring/post-P/finally to reconstruct masks/quantized weights. SVS inputs stay unaugmented/unchanged. Stream 0 reproduces original states/logits in both arms. Streams 1–2 sample 1,536 retained-training images without replacement: NumPy `default_rng` seed 925000 + 100 × model seed + stream ID, indices sorted by class/path. Augmentation seed is 935000 + 100 × model seed + stream ID, with original transforms/batch/CMA. Cached streams are shared across arms/baseline/orders without selection/exclusion; validation decisions precede test evaluation. Panel B pairs same-order test images from stream 0/new streams over 6 reused models. Values are mean (sample SD); d is reference-minus-new accuracy (pp).

Table 28: ImageNet-100 sensitivity to calibration streams.

(a) Qualification and between-order spread

Arm / stream	VAL joint / TEST budget	TEST S: mean / max (pp)	TEST worst loss (pp)
Both arms / 0	6/6 / 6/6	0.357 / 0.740	3.673
Final-only / 1	6/6 / 6/6	0.300 / 0.760	3.510
Final-only / 2	6/6 / 6/6	0.233 / 0.680	3.623
Rebuild / 1	5/6 / 6/6	0.170 / 0.280	3.713
Rebuild / 2	6/6 / 6/6	0.440 / 0.660	3.697

(b) Same-order prediction and correctness changes

Arm / stream / order	Predictions differing (%)	Correctness differing (%)	d (pp)
Final / 1 / PQ	14.840 (0.400)	8.187 (0.287)	-0.327 (0.266)
Final / 1 / QP	15.103 (0.446)	8.190 (0.441)	0.057 (0.162)
Final / 2 / PQ	15.023 (0.677)	8.233 (0.287)	-0.033 (0.335)
Final / 2 / QP	15.323 (0.468)	8.517 (0.375)	0.057 (0.431)
Rebuild / 1 / PQ	16.617 (0.587)	9.287 (0.393)	-0.087 (0.443)
Rebuild / 1 / QP	16.597 (0.313)	9.153 (0.255)	0.353 (0.170)
Rebuild / 2 / PQ	16.540 (1.077)	9.077 (0.813)	-0.070 (0.406)
Rebuild / 2 / QP	16.797 (1.033)	9.480 (0.674)	0.033 (0.412)

All six pass BN screening; rebuild stream 1, seed 4003, fails qualification despite a validation spread of 0.56 pp (paired boundary details: D.2). Both rebuild streams change every order's mask: mean Hamming 1.545/1.536% of 1,920 original channels. Final-only changes validation winners in 4/6 and 2/6 models despite small gaps/stable budgets. Main results use stream 0. Separate score/slicing/Q full-test replay uses stored BN (final-only) or recomputed BN (rebuild), sharing forward/calibration primitives.

Same-order mean prediction disagreement is comparable to or greater than reference cross-order disagreement (14.670%). Ordered image paths/indices and evaluation-batch hashes establish alignment. These contrasts neither supply a CIFAR stream baseline nor isolate an intrinsic order effect.

E.2 Event-frame CIFAR10-DVS

Eight VGG-SNNs (3000–3007) use CIFAR10-DVS (Li et al., 2017b): ten equal-count, two-polarity event bins resized to 48×48. Eight 3×3 Conv–BN–LIF layers have widths (64,128,256,256,512,512,512,512), four average pools, linear readout and temporal-logit averaging (T=10); neurons/BN follow A.1/A.3. Reserving 100/class from 900 training sequences/class gives 8,000 training/1,000 validation/1,000 unchanged test. FP32 TET: 200 epochs, batch 64, cosine Nesterov SGD (0.1, momentum 0.9, weight decay 0.0005), TF32 off. Training/calibration uses crop 128 (scale 0.7–1, nearest), resize 48/horizontal flip.

P prunes first convolutions in four adjacent pairs, including the stem, preserving pair-output dimensions. SVS averages time-averaged-spike spatial ranks on resize-only inputs, with original-index ties. Zero-containing max-range Q covers seven non-stem convolutions. BN-only/pre-scoring/post-P/final CMA shares 1,536 fixed-permutation retained-training sequences, batch 64. All eight pass BN preservation (mean/max validation loss 0.575/1.500 pp; losses equal to the BN-preservation tolerance pass).

Table 29: Event-frame DVS compact compression results.

p / bits	**Conv. MAC saved (%)**	**VAL joint / TEST budget**	**TEST S: pre → final / max (pp)**	**Worst loss (pp)**
0.10 / 4	9.960	8/8 / 8/8	0.575 → 0.5125 / 1.100	1.000
0.10 / 8	9.960	8/8 / 8/8	0.338 → 0.3250 / 0.600	0.625
0.20 / 4	19.971	8/8 / 8/8	0.525 → 0.2875 / 0.500	1.775
0.20 / 8	19.971	8/8 / 8/8	0.475 → 0.3625 / 0.800	1.913
0.10 / 4 fixed/frozen	9.960	8/8 / 8/8	0.613 → 0.0000 / 0.000	0.750

Controls give validation/test accuracy (%) without additive causal attribution; Spikingformer uses original crop-CMA (E.3).

Table 30: DVS and CIFAR Spikingformer control accuracies.

Control	**DVS VAL / TEST**	**CIFAR Spikingformer VAL / TEST**
Original	83.375 / 81.263	77.982 / 77.335
BN only	82.800 / 81.275	75.760 / 76.076
P only, p=0.10	81.963 / 80.662	75.055 / 74.981
P only, p=0.20	81.612 / 79.475	—
Q only, 4-bit	82.688 / 81.175	75.845 / 75.890
Q only, 8-bit	82.787 / 81.188	76.013 / 76.061

At p=0.20/4-bit, mean terminal test prediction disagreement is 3.650% despite small spread. D retains budget-feasible pre-BN gaps and previously-best accuracy decreases. All fixed/frozen final states/logits coincide. Full-test slicing/Q replay shares stored masks/BN/forward, without independent scores/calibration. Event-frame P/Q excludes asynchronous hardware, temporal causality and threshold/SOP generalization.

E.3 CIFAR Spikingformer and calibration qualification

Eight CIFAR-100 Spikingformers (Zhou et al., 2026) (2000–2007) use four native blocks, width 384, eight attention heads, MLP ratio 4, T=4 and A.1's input decay/sigmoid surrogate/detached reset. Epoch selection/qualification reuse 4,096 internal-validation images. AdamW uses rate 5×10^{-4}, weight decay 0.06, batch 64, 410 epochs; smoothed cross-entropy on temporal-mean logits, RandAugment, mixup 0.5 through epoch 200, erasing 0.25. Normalization mean/std: (0.4914,0.4822,0.4465)/(0.247,0.2435,0.2616). Training uses AMP; selection/compression use FP32, TF32 off.

At p=.10, P reduces MLP hidden width 1536→1382, preserving residual width and leaving attention/tokenizer/head uncompressed. SVS ranks time-averaged 8×8 hidden spikes with default relative tolerance/original-index ties. Zero-containing Q covers eight MLP convolutional tensors; biases/BN stay FP32. Scoring uses unaugmented training indices [0:1536]. Original calibration caches 32-pixel crops with padding 4 and horizontal flips, batch 64. All 29 BNs use simultaneous reset-CMA pre-scoring/post-P/finally.

All BN candidates. Both procedures (A.3) cross both input recipes (below/E.4); BN-only losses are original-relative pp per split.

Table 31: Spikingformer calibration candidate screening.

Input / procedure	CIFAR VAL mean / max; pass/8	CIFAR TEST	Public qualification / evaluation
Crop / CMA	2.222 / 2.686; 0/8	1.259	-0.100 / 0.304
Training-informed / CMA	0.491 / 1.123; 8/8	-0.164	-0.200 / 0.287
Crop / sequential	2.057 / 2.759; 0/8	1.371	0.060 / 0.391
Training-informed / sequential	0.534 / 1.147; 8/8	-0.144	-0.020 / 0.347

CIFAR training-informed inputs use bicubic resized crop 32 (scale and aspect ratio [1,1]), flip 0.5, `rand-m9-n1-mstd0.4-inc1`, native FP32 normalization, zero erasing 0.25 (one region, area [0.02,1/3], aspect [0.3,1/0.3], ten attempts). Mixing is excluded: all selected epochs exceed 200. Both procedures share indices, cached streams and batches of 64, without labels. Native FP32 normalization subtracts 255μ then divides by 255σ on uint8-valued floats. Sequential estimation jointly changes upstream execution and variance estimation. The comparison does not isolate individual augmentations or explain earlier ResNet losses (§4.2).

Among candidates passing all eight validation screens, maximum total BN-only correct count selects training-informed CMA (table-order ties) before construction/test. For each order, reconstruction recomputes the masks and quantization ranges, with BN recalibration before scoring, after pruning and at the final endpoint. Both orders meet final budgets in all eight models/both splits throughout; BN screening determines joint qualification.

Table 32: CIFAR Spikingformer calibration and compact results.

Recipe / bits / construction	VAL joint/8	TEST S: pre → final / max (pp)	Worst loss (pp)
Crop / 4 / dynamic	0	0.307 → 0.188 / 0.390	2.419
Crop / 8 / dynamic	0	0.176 → 0.163 / 0.290	2.334

Recipe / bits / construction	VAL joint/8	TEST S: pre → final / max (pp)	Worst loss (pp)
Crop / 4 / fixed/frozen	0	0.247 → 0.000 / 0.000	2.406
Crop / 8 / fixed/frozen	0	0.161 → 0.000 / 0.000	2.366
Selected / 4 / dynamic	8	final 0.210 / 0.480	0.725
Selected / 4 / fixed/frozen	8	final 0.000 / 0.000	0.630

MLP-only MAC saving is 10.026%; parameters fall 9360916→8885980 (no full-model MAC/packed-runtime claim). Selected dynamic spread rises 0.188→0.210 pp: selection optimized BN preservation. All eight dynamic pairs differ in masks/weights; mean mask Hamming is 1.404%, lifted MLP-weight mismatch 7.720%, prediction disagreement 9.543%. Mismatch counts exact inequalities over all coefficients in eight original-shape MLP tensors: compact first-convolution rows/second-convolution columns map to original channel IDs, with zeros elsewhere; biases/BN/attention are excluded. Fixed/frozen states/logits coincide. For budgets 1.0 pp/1.5 pp/2.0 pp/5.0 pp, (validation,test) passage out of 8 is (0, 5)/(4, 8)/(4, 8)/(8, 8) for dynamic and (2, 7)/(4, 8)/(7, 8)/(8, 8) for fixed/frozen. Separate SVD/slicing/Q/full-logit replay uses stored BN/shared native forward; sequential moments are checked separately.

E.4 Public ImageNet-1k Spikingformer

One released checkpoint uses eight blocks, width 768, eight heads, MLP ratio 4, T=4 and A.1 Spikingformer neurons (Torch replaces CuPy); external selection is uncontrolled. Bicubic 224-pixel inputs, crop fraction 1.0/ImageNet normalization give 75.806% top-one accuracy on 50,000 official-validation images. Forward: AMP FP16/batch 36; parameters/BN/Q: FP32, TF32 off. Before compression outcomes, five images/class define 5,000 qualification images; remaining 45,000 images are evaluated after locking. Both partitions served baseline checks; neither is fresh test data.

Calibration/SVS use 1,536 shuffled official-training images: one/class plus 536 fixed random classes' second images. Crop calibration: bicubic resized crop (area [1,1], aspect [0.75,4/3])/flip 0.5, batch 32. Training-informed adds `rand-m9-mstd0.5-inc1`, zero erasing 0.25 and mixup/cutmix 0.8/1.0 (application 1, switch 0.5), discarding mixed targets. Cached streams recalibrate all 53 BNs (A.3). Neither alternative recreates historical training RNG/BN buffers. All four E.3 candidates pass; minimum BN-only qualification loss selects training-informed CMA (table-order ties) before compression/evaluation. Evaluation-partition accuracy after selected BN recalibration is 75.660% versus 75.947% originally. The map applies pre-scoring/post-P/finally. SVS averages unaugmented FP32 time-averaged 14×14 hidden-spike ranks, with original-index ties. P slices hidden channels; zero-containing Q covers sixteen MLP convolutional weight tensors; attention, tokenizer and head are uncompressed, and residual widths are unchanged. Parameters: 66,363,208 original, 64,467,160/62,583,424 at MLP removal fractions p=0.05/0.10, not whole-model MAC savings.

All 16 cells pass qualification/both-order evaluation budgets. F/R means fixed/reselected masks before the slash, frozen/recomputed ranges after. Entries: evaluation S/worst original-relative loss (pp); dynamic: pre→final S/final loss; last column: dynamic qualification worst loss. All share one model/45,000 images.

Table 33: Public ImageNet-1k Spikingformer compact results.

p / bits	F/F	F/R	R/F	R/R: pre→final / loss	Qual. loss
0.05 / 4	0.000 / 1.478	0.007 / 1.484	0.002 / 1.478	1.062→0.009 / 1.484	0.820
0.05 / 8	0.000 / 1.036	0.016 / 1.051	0.200 / 1.036	0.249→0.216 / 1.051	0.080
0.10 / 4	0.000 / 3.229	0.136 / 3.364	0.051 / 3.280	1.104→0.084 / 3.364	2.480
0.10 / 8	0.000 / 2.764	0.193 / 2.958	0.002 / 2.764	0.029→0.196 / 2.958	2.300

Fixed/frozen complete states/full logits coincide in every pair/both partitions; reselected masks/predictions differ. Terminal evaluation prediction disagreement spans 10.916–13.384% across four dynamic cells. At dynamic p=0.10/8-bit, PQ correct counts fall 32,946→32,845; QP stays 32,933: the best order before recalibration degrades, widening spread. P-only losses: 1.129/3.009 pp (p=0.05/0.10); Q-only final losses: 0.684/0.400 pp (4/8-bit). Separate full-evaluation slicing/integer-Q replay uses stored BN/shared forward and independent FP/Q4/Q8 scores, without independent training/calibration.

F. Threshold scaling and modeled work

F.1 Protocol, accuracy and cost

This descriptive follow-up adds $\alpha \in \{1, 1.02, 1.05\}$ to C.1's twelve T=2 checkpoints, retaining its recipe/images/streams and A.4 selection. Only PQ/QP differ at α=1; active A gives §3.2's four classes. SOPs count binary input events × valid outgoing convolution edges (Sorbaro et al., 2020), including position/padding/stride/dilation/groups and T=2 once; zero weights are not skipped. RGB-stem/classifier dense MACs are separate. BN, pooling, residual sums, neuron updates and memory/latency are excluded. This models work, not hardware energy; binary-input counting excludes SEW-ADD. Relative SOP range is $100(\max_s C_s - \min_s C_s)/\operatorname{mean}_s C_s$, where C=mean SOPs/image. Panel A gives across-model mean/maximum within-run test spread and mean relative SOP range. All/any counts (out of 12) require every/at least one order within 5.0 pp of original accuracy.

Table 34: Threshold scaling, accuracy and synaptic work.

(a) All threshold and recalibration conditions

α	BN phase	Val. (all, any)	Test (all, any)	Spread: mean / max (pp)	SOP range (%)
1.00	Before	(12, 12)	(12, 12)	0.3467 / 0.93	0.364
1.00	After	(12, 12)	(12, 12)	0.1675 / 0.35	0.009
1.02	Before	(12, 12)	(12, 12)	0.845 / 1.26	12.478
1.02	After	(12, 12)	(12, 12)	0.373 / 0.64	0.031
1.05	Before	(0, 12)	(3, 12)	3.835 / 7.28	25.341
1.05	After	(0, 4)	(3, 5)	0.418 / 0.84	0.068

(b) Mean SOPs at α=1.02 (12 models; millions/image)

Order	Before BN	After BN
APQ	314.420	313.478
QPA	353.713	313.486
QAP	313.447	313.447
PQA	354.845	313.489

At α=1.02, 3/12 pre-BN validation spreads exceed δ despite all-order budget passage (§4.6). Terminal-minus-pre best/worst changes average re-ranked phase-wise extrema: -0.303/+0.169 pp. At α=1.05, changes are -3.410/+0.007 pp, mainly best degradation; no run qualifies all orders on validation, though every model retains a feasible pre-BN order. A-only at α=1.02 changes full-test mean work by +13.206 M SOP/image versus the original before BN; recalibration then changes it by -43.599 M. Higher thresholds need not lower network-wide work. We hold BN affine parameters fixed and measure SOP separately; A-only differs from §4.4's identity-Q ordering interaction.

F.2 Layer, image and sampled-channel diagnosis

At α=1.02, QPA/PQA work falls in every stage mean and all 10,000 test images/model; QAP retains state/outputs/costs. Entries average each run's summed absolute layer-mean

differences and mean absolute per-image total differences (M/image). The small sums rule out cancellation among large layer-mean differences as the sole explanation; per-image work differs and images/layers are dependent. Sampled replay below locates the first firing change; reductions chiefly involve channels active in both phases.

Table 35: Layer, image and channel diagnostics of work changes.

(a) Full-test synaptic operations

Order pair	Layer-mean differences: before	Layer-mean differences: after	Per-image total differences: after
APQ-PQA	40.515	0.177	0.695
QAP-QPA	40.383	0.198	0.637

On the first 256 test images, α=1.02 QPA preserves stem outputs/first-block BN input moments. With affine parameters fixed, new buffers first change firing after this BN; stage 1's second-block output changes at both time steps. Sample-silent channels emit no spikes over these images, both time steps and all positions within an order/model/phase. Matched channels are grouped by pre/final activity; final-minus-pre spikes are summed over channels/19 LIF layers, divided by 256, then averaged over 12 models. Active-in-both channels spike in each phase.

(b) Emitted-spike changes (first 256 images; spikes/image)

Order	Newly silent	Active in both	Newly active	Total
QPA	-0.106	-32040.63	+0.782	-32039.96
PQA	-0.132	-33066.73	+0.801	-33066.07

Entries are sampled spikes/image (rounding affects sums), not a fanout-weighted SOP decomposition. Different masks/shared forward/scoring preclude attributing the full gap to BN, resets or residuals. Sampled state/logit/cost checks, separate fanout arithmetic and unchanged instrumented outputs support consistency, without full-test replay or independent models/calibration (A.4).