How Sparsity Allocation Shapes Label-Free Post-Pruning Recoverability

Anonymous authors
Paper under double-blind review

Abstract

Post-training pruning involves two coupled design choices: how sparsity is allocated across layers and how the resulting sparse model is repaired. We study whether these choices can be treated independently when recovery must be label-free, using only a small unlabeled calibration set. Across ResNet-18/34/50 on CIFAR-10, CIFAR-100, and Imagenette, we combine four unstructured sparsity-allocation strategies with the same suite of gradient-free repair methods. We find that allocation changes both the attainable repaired accuracy and the repair method that works best, with the strongest interaction appearing in a repair-sensitive transition regime between stable recovery and collapse. Across 27 main ResNet settings, the best repair method changes across allocations in 21 settings, and allocation changes the recoverability regime in 13. We further introduce LDMS-σ, a calibration-only diagnostic based on dense–pruned activation statistics, which selects the best repaired allocation in 25 of 27 settings and remains within one percentage point in 26. These results show that post-pruning recoverability is a property of the allocation–repair pair rather than either component alone.

1 Introduction

Modern neural networks are often pruned after training to reduce memory, computation, and deployment cost (Han et al., 2015; Frankle & Carbin, 2019; Blalock et al., 2020). This paper studies *unstructured* post-training pruning: individual convolutional weights are removed while the network topology is kept fixed. For a target global sparsity, an allocation strategy decides how many weights each layer may keep. When labeled data or further training is unavailable, the pruned model is often repaired using only an unlabeled calibration set. Such label-free repair recomputes BatchNorm statistics or corrects activation statistics distorted by pruning (Lazarevich et al., 2021; Saikumar & Varghese, 2025; Wang et al., 2024). The deployed model therefore depends on both where sparsity is placed and how the sparse model is repaired.

Existing work studies these two choices largely in isolation, with allocation methods typically evaluated under a fixed recovery pipeline and repair methods typically evaluated on a fixed sparse model (Evci et al., 2020; Lee et al., 2021; Frantar & Alistarh, 2023; Sun et al., 2024; Wang et al., 2024). Sparsity allocation methods determine how a fixed sparsity budget is distributed across layers and are commonly evaluated before or after labeled fine-tuning. Label-free repair methods instead take the sparse model as fixed and recover its behavior through BatchNorm recalibration or activation-statistic correction. These studies can identify a strong allocation under one recovery procedure or a strong repair method for one sparse model, but they do not answer the system-level question considered here: *how does sparsity allocation shape what label-free repair can recover, and can the two choices be made independently?*

We study this question by treating allocation and repair as two axes of the same post-training compression pipeline. We pair four allocation strategies, Erdős–Rényi-Kernel (ERK), Layer-Adaptive Magnitude-based Pruning (LAMP) (Evci et al., 2020; Lee et al., 2021), uniform layerwise allocation, and unconstrained global magnitude pruning, with a shared suite of label-free repair methods. The main ResNet evaluation contains 27 architecture–dataset– sparsity settings: three ResNet depths, three image datasets, and three high-sparsity targets. Additional experiments evaluate ImageNet-100 and held-out convolutional architectures. This design lets us compare repair methods under the

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.

same allocation, allocations under the same repair, and the best attainable repaired accuracy at fixed architecture, dataset, and global sparsity.

The results organize into three recoverability regimes. In the stable regime, BN recalibration is already close to the best attainable label-free repair. In the collapse regime, all evaluated repairs remain ineffective because too little usable activation signal survives pruning. Between them lies a repair-sensitive transition regime, where BN recalibration is insufficient but stronger activation-statistic repair can still recover substantial accuracy. Allocation matters most in this regime: at the same global sparsity, different allocations can produce different repair ceilings, occupy different recoverability regimes, and favor different repair methods.

We further ask whether the most recoverable allocation can be identified without labeled evaluation or exhaustive repair sweeps. We introduce LDMS-σ, a calibration-only diagnostic computed from dense and pruned activations on unlabeled data. It combines layerwise dispersion in preserved activation scale with within-layer channel heterogeneity. In the main ResNet evaluation, the strongest repair method changes across allocations in 21 of 27 settings by exact three-seed mean argmax. LDMS-σ selects the best repaired allocation in 25 of 27 settings using only unlabeled calibration activations. The diagnostic has no learned weights or fitted hyperparameters; the same formula is applied to every setting, and component ablations test whether its two terms are both needed. The anonymous release includes code, table and figure generation scripts, canonical machine-readable result files, verification scripts, and instructions for running the experimental pipeline with dense model weights.

Our contributions are threefold:

1. We formulate label-free post-pruning recoverability as a coupled allocation–repair problem and show that identical global sparsity can lead to substantially different repair ceilings.
2. We provide a controlled evaluation across four allocation strategies, multiple label-free repair methods, three ResNet architectures, three datasets, and high-sparsity settings, showing that both recoverability regime and preferred repair method depend on the allocation.
3. We introduce LDMS-σ, a calibration-only diagnostic based on layerwise scale dispersion and within-layer channel heterogeneity, and show that it can identify the most recoverable allocation using only unlabeled activations.

2 Related Work

Post-training pruning combines a pruning rule, an allocation rule, and a recovery step. Prior work usually optimizes one component while holding the others fixed. We do not propose a new pruning rule or study labeled sparse fine-tuning; instead, we examine how allocation changes what label-free recovery can recover.

Sparsity allocation. Sparsity allocation determines how a fixed global parameter budget is distributed across layers. ERK uses layer shape to assign nonuniform density (Evci et al., 2020), while LAMP constructs layer-adaptive magnitude scores that account for the cumulative distortion induced by pruning (Lee et al., 2021). More recent methods optimize allocation directly. FCPTS learns layerwise sparsity under an explicit global constraint through a differentiable post-training objective (Gong et al., 2024), whereas DSA searches over functions that map per-layer importance statistics to sparsity rates for large language models (Li et al., 2024). Related LLM work also allocates sparsity according to activation outliers or layerwise reconstruction behavior (Yin et al., 2024; Huang et al., 2025). These methods primarily seek an allocation that preserves task performance under a fixed downstream recovery pipeline. Our focus is complementary: we study how the allocation itself changes the effectiveness and ranking of label-free repair methods.

Post-training pruning and recovery. The standard route to recovering sparse-model accuracy is masked fine-tuning with labeled data and gradient updates (Han et al., 2015; Blalock et al., 2020). Post-training pruning methods target settings where such retraining is unavailable or too costly. SparseGPT solves a layerwise reconstruction problem for one-shot LLM pruning (Frantar & Alistarh, 2023), while Wanda prunes using weight magnitudes modulated by input activation norms (Sun et al., 2024). RIA and related plug-and-play methods further use relative importance

and activation information for efficient post-training pruning (Zhang et al., 2024). These works show that calibration signals can support pruning without labeled retraining. In parallel, prior work has studied lightweight recovery methods such as BatchNorm recalibration, bias correction, channel equalization, affine calibration, and layerwise reconstruction to compensate for distributional shifts after compression (Shomron & Weiser, 2020; Nagel et al., 2019; 2020; Lazarevich et al., 2021). Benchmarking work on post-training sparsity commonly represents pruning as an allocation stage followed by a reconstruction or calibration stage (Wang et al., 2024). These studies establish the value of post-training recovery, but usually evaluate it for a fixed pruning configuration. We instead vary allocation and repair jointly under the same global sparsity.

Activation collapse and statistical repair. Several lines of work connect model degradation to disrupted activation statistics. REPAIR rescales activations to mitigate variance collapse in interpolated neural networks (Jordan et al., 2023). In one-shot pruning, signal-collapse analysis attributes severe failure to the loss of representation-level activation variance and proposes restoring signal without updating the retained weights (Saikumar & Varghese, 2025). Layerwise calibration and related post-training corrections likewise restore activation or output statistics after pruning (Lazarevich et al., 2021). The ASR-q50 and clipped-ASR repairs evaluated in this paper perform channel-wise variance matching between dense and sparse vision models and stabilize the correction through reliability-based shrinkage (Zhan et al., 2026); Section 4 gives the exact equations used in our experiments. Concurrent work on post-training LLM pruning applies analytic energy compensation to correct statistical distortions after weight removal (Yu et al., 2026). Together, these results motivate activation statistics as a description of post-pruning recoverability.

Allocation-dependent recoverability. Allocation research typically compares pruning rules under one recovery procedure (Gong et al., 2024; Li et al., 2024; Wang et al., 2024), whereas repair research starts from a fixed sparse model (Lazarevich et al., 2021; Saikumar & Varghese, 2025). As a result, existing evaluations do not characterize whether two allocations at the same global sparsity have different repair ceilings, whether the preferred repair method changes with allocation, or whether those differences can be diagnosed from unlabeled dense–pruned activation statistics. We address this missing joint view through a controlled allocation–repair matrix and a calibration-only diagnostic for ranking candidate allocations by recoverability.

3 Problem Setting

Figure 1 summarizes the coupled allocation–repair pipeline, the LDMS-σ diagnostic, and the three empirical recoverability regimes studied below.

Let f_θ denote a dense pretrained network and let $f_{\hat{\theta}}$ be the sparse model obtained by applying an allocation strategy $\mathcal{A}$ followed by unstructured magnitude pruning to a target global sparsity s. Global sparsity is the fixed fraction of convolutional weights removed from the whole model; $\mathcal{A}$ decides how that budget is distributed across layers. This is distinct from the "Global" baseline, which applies one magnitude threshold across all repairable convolutional weights. Let $\mathcal{C} = \{x_1, \ldots, x_n\}$ be an unlabeled calibration set, with $n = 128$ unless otherwise stated. A label-free repair method $\mathcal{R}$ maps the sparse model to a repaired model using only $\mathcal{C}$ and, when required, the dense reference f_θ. No labels or gradient updates are used.

For each architecture, dataset, and sparsity target, we instantiate four candidate allocations: ERK, LAMP, Uniform, and Global. The core empirical question is whether the repaired accuracy $\mathrm{Acc}(\mathcal{R}(\mathcal{A}(f_\theta, s), \mathcal{C}))$ separates into an allocation-independent repair effect and a repair-independent allocation effect, or whether the allocation–repair pair must be evaluated jointly. The setting is label-free but not reference-free: LDMS-σ and the activation-statistic repairs use dense-model activations on $\mathcal{C}$.

4 Label-Free Repair Methods

We evaluate a shared suite of label-free, gradient-free repair methods as a controlled downstream axis for measuring allocation-induced recoverability. All methods use the same unlabeled cali-

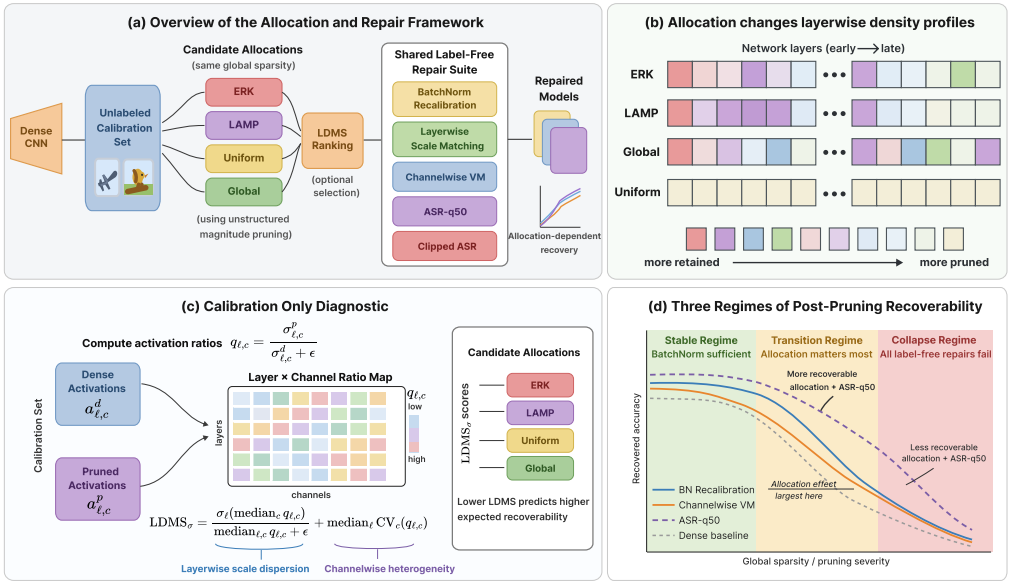


Figure 1: Overview of the allocation–repair framework. (a) Candidate allocations at fixed global sparsity are repaired with a shared label-free suite. (b) Allocation changes layerwise density profiles. (c) LDMS-σ summarizes dense–pruned activation ratios through layerwise scale dispersion and channelwise heterogeneity. (d) Repaired models fall into stable, transition, and collapse regimes. Channelwise VM denotes channelwise variance matching; panels (b) and (d) are schematic.

bration set, and methods that modify activations or weights are followed by the same BatchNorm recalibration procedure.

The main repair set includes six methods. No repair evaluates the pruned model directly. BN recalibration recomputes BatchNorm running statistics (Shomron & Weiser, 2020; Ioffe & Szegedy, 2015) from unlabeled calibration data. Layerwise scale matching applies one scalar correction per repairable convolutional layer, serving as a minimal activation-statistic baseline before more aggressive channelwise correction. Channelwise variance matching applies the corresponding variance correction per output channel without shrinkage. Following prior channelwise repair work (Zhan et al., 2026), ASR-q50 applies reliability-aware shrinkage to channelwise variance matching, so channels with weak post-pruning activation variance do not receive extreme inverse-variance corrections. Clipped ASR further constrains the resulting scale factors to $[0.5, 2.0]$ to reduce unstable corrections in fragile regimes.

For a channel c, the raw channelwise variance-matching factor is

$$\gamma_c^{\mathrm{raw}} = \sqrt{\frac{\sigma_{d,c}^2}{\sigma_{p,c}^2 + \varepsilon}}, \tag{1}$$

where $\sigma_{d,c}^2$ and $\sigma_{p,c}^2$ are dense and pruned activation variances measured on the calibration set. ASR-q50 replaces this raw correction with a shrunk factor

$$\rho_c = \frac{\sigma_{p,c}^2}{\sigma_{p,c}^2 + \mathrm{Med}_j\{\sigma_{p,j}^2\}}, \tag{2}$$

$$\gamma_c^{q50} = \rho_c \gamma_c^{\mathrm{raw}} + (1 - \rho_c), \tag{3}$$

and clipped ASR uses

$$\gamma_c^{\mathrm{clip}} = \mathrm{clip}(\gamma_c^{q50}, 0.5, 2.0). \tag{4}$$

Equations 1–4 are the complete closed-form definition of the ASR-q50 and clipped-ASR variants used in this paper. They use only dense and pruned activations measured on the unlabeled calibration set and do not introduce learned parameters.

4.1 Calibration Diagnostics

In addition to evaluating repaired accuracy, we compute calibration-only diagnostics from dense and pruned activations. These diagnostics are not used to fit the repair methods; they characterize the activation landscape induced by each allocation. For each repairable convolutional layer, let

$$q_{\ell,c} = \frac{\sigma^p_{\ell,c}}{\sigma^d_{\ell,c} + \epsilon} \tag{5}$$

denote the pruned-to-dense activation standard-deviation ratio for channel c in layer ℓ, computed over the calibration set.

Our main diagnostic is

$$\mathrm{LDMS}_\sigma = \frac{\sigma_\ell(\mathrm{median}_c\, q_{\ell,c})}{\mathrm{median}_{\ell,c}\, q_{\ell,c} + \epsilon} + \mathrm{median}_\ell\, \mathrm{CV}_c(q_{\ell,c}). \tag{6}$$

The first term measures layer-to-layer dispersion in preserved activation scale: large values indicate that different layers retain very different amounts of activation signal, making lightweight recalibration unreliable. The second term measures within-layer channel heterogeneity: large values indicate that channels within the same layer are damaged unevenly, making raw channelwise correction less stable. This form can be viewed as a signal-preservation diagnostic: the denominator normalizes by the typical retained activation scale, while the two dispersion terms penalize uneven preservation across layers and channels, consistent with the view that severe pruning failure is tied to loss of activation signal and representation-level variance (Saikumar & Varghese, 2025). We use equal weighting to keep LDMS-σ calibration-only and untuned: both terms are scale-normalized dispersion measures, and the component ablations in Table 3 show that combining them reduces regret relative to either term alone.

We also report two auxiliary variants. $\mathrm{LDMS}_{\mathrm{var}}$ uses the corresponding variance ratio

$$r_{\ell,c} = \frac{\mathrm{Var}(a^p_{\ell,c})}{\mathrm{Var}(a^d_{\ell,c}) + \epsilon}, \tag{7}$$

and Full LDMS-var adds the median activation mean shift to $\mathrm{LDMS}_{\mathrm{var}}$. These variants test whether the recoverability signal depends on standard-deviation ratios, variance ratios, or additional mean displacement. A more detailed interpretation of these diagnostics is provided in the supplementary material.

5 Experimental Setup

Architectures and datasets. Main experiments use ResNet-18, ResNet-34, and ResNet-50 (He et al., 2016) on CIFAR-10, CIFAR-100 (Krizhevsky, 2009), and Imagenette (Howard, 2019). ResNet-34 provides an intermediate-depth case for testing whether allocation rankings remain stable across depth. DenseNet-121 (Huang et al., 2017) on CIFAR-10, CIFAR-100, and Imagenette, together with VGG16-BN (Simonyan & Zisserman, 2014; Ioffe & Szegedy, 2015), provides held-out architecture checks under different connectivity and activation geometries. ImageNet-100 (Deng et al., 2009) on ResNet-18 provides a data-scale check beyond the primary datasets. A single-seed ResNet-50/ImageNet-1K experiment at 92.5% sparsity is reported in the supplementary material as a large-scale collapse-regime stress test.

Allocation strategies. Four strategies span the design space from structure-aware to unprotected. ERK preserves more weights in layers with fewer parameters, protecting narrow bottlenecks (Evci et al., 2020). LAMP allocates sparsity to minimize a layerwise distortion bound, concentrating pruning in layers whose outputs are less sensitive to weight removal (Lee et al., 2021). Uniform allocation assigns the global sparsity uniformly to every repairable convolutional layer, providing no sensitivity-aware protection. Global magnitude pruning applies a single threshold across all convolutional weights (Blalock et al., 2020), which in practice concentrates sparsity in layers with smaller weight scales. ERK, LAMP, and Uniform determine layerwise budgets before unstructured magnitude pruning, whereas Global applies a single magnitude threshold across all repairable convolutional weights. These strategies are not intended to exhaust all recent allocation algorithms.

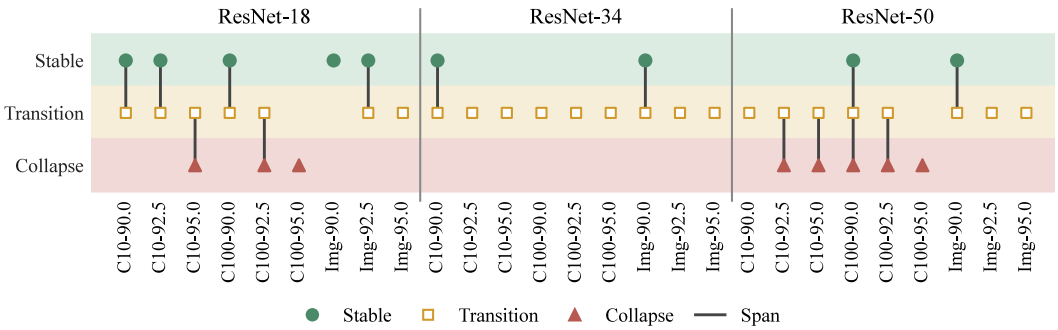


Figure 2: Recoverability regimes across allocations in the 27 main ResNet settings. Each x-axis position corresponds to one architecture, dataset, and sparsity setting. Colored markers show the recoverability regimes reached by at least one allocation under that setting, and black vertical segments indicate settings where different allocations occupy different regimes. In 13 of 27 settings, allocation changes the recoverability regime at the same global sparsity.

They are deliberately chosen controls for the interaction study: ERK and LAMP are widely used structure-aware baselines, Uniform removes sensitivity information, and Global tests whether a single model-wide threshold leaves repairable activation structure. This lets us ask whether the same repair method behaves differently when the allocation mechanism changes.

Repair methods. We use the six repair methods introduced in the preceding section. Within each run at the default budget, all methods share the same fixed 128-image unlabeled calibration subset and use no gradient updates. BN statistics are accumulated over 20 mini-batches from the corresponding calibration loader. In a seed-0 sweep over $n \in \{32, 64, 128, 256\}$, the LDMS-$\sigma$ selection is unchanged in 26 of 27 settings, and the median range of the best repaired accuracy is 0.14 percentage points. Main results report means and standard deviations over three pruning and calibration seeds.

Sparsity levels. Main tables report 90.0%, 92.5%, and 95.0% global sparsity, chosen to sample stable, transition, and near-collapse regimes. We focus on high sparsity because lower sparsity settings are often stable after simple BN recalibration, making allocation–repair interactions harder to observe; the conclusion is therefore about the aggressive post-training pruning regime rather than all possible sparsity levels. For ResNet-50 on CIFAR-100, 95.5% is included because the transition regime extends into higher sparsity on this harder dataset. DenseNet-121 uses architecture-specific sparsity ranges because its transition to collapse occurs earlier than in the main ResNet evaluation. We use stable, transition, and collapse as empirical regime descriptors rather than fixed theoretical classes. Labels are assigned at the allocation level. For a fixed architecture, dataset, sparsity, and allocation, we compare BN recalibration with the best label-free repair under the same allocation. A case is stable when BN recalibration is within 5 percentage points of the best repaired accuracy, collapse when the best repair recovers less than 20% of the dense-to-chance accuracy gap, and transition otherwise. These thresholds are descriptive conventions used to organize the experiments, not universal performance laws. Across stable thresholds of 3, 5, and 7 percentage points and collapse thresholds of 15–30% recovered dense-to-chance gap, allocations span at least two regimes in 9–15 of the 27 main settings. Under the default thresholds, Figure 2 shows that different allocations occupy different recoverability regimes in 13 of the 27 main settings, demonstrating that global sparsity alone does not determine whether a sparse model is stable, repair-sensitive, or collapsed.

6 Results

6.1 Best Repair Method Varies with Allocation

The six repair methods span BN recalibration, layerwise correction, raw channelwise correction, and reliability-aware channelwise correction, thereby probing progressively finer corrections to the activation geometry left by each allocation. Full method-level comparisons are reported in the sup-

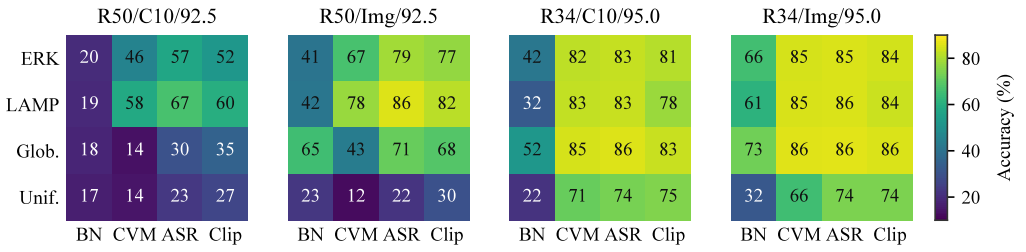


Figure 3: Allocation–repair matrices for representative transition settings. Cells report three-seed mean accuracy (%). Changing the allocation alters both the repair ceiling and, in some cases, the strongest repair method.

Table 1: Allocation–repair matrix for ResNet-50/CIFAR-10 at 92.5% sparsity. Entries are three-seed mean accuracies (%); bold marks the best repair under each allocation.

Allocation	None	BN	Layer	VM	ASR-q50	Clip
ERK	9.97	19.75	24.02	45.70	**57.22**	52.35
LAMP	10.61	19.29	53.28	58.40	**66.71**	60.28
Global	10.00	18.20	14.17	13.58	30.19	**34.89**
Uniform	10.00	16.67	11.49	14.48	22.92	**26.57**

plementary material; the main text focuses on how allocation changes the strongest repair and the resulting repair ceiling.

Figure 3 illustrates the central interaction: at fixed architecture, dataset, and sparsity, allocation changes both the recoverable accuracy and, in some cases, the strongest repair method. Table 1 provides a concrete numerical instance: at 92.5% sparsity on ResNet-50/CIFAR-10, ASR-q50 reaches 66.71% under LAMP but only 22.92% under Uniform, while clipped ASR becomes the strongest repair under Global and Uniform.

Across the 27 main ResNet settings, the best repair changes across allocations in 21 settings by exact three-seed mean argmax. Under a one-point near-tie rule, 12 settings still have no repair method that is near-best for every allocation. The allocation-specific repair identities and the full tolerance analysis are reported in the supplementary material. A seed bootstrap gives the same qualitative picture, with median counts of 23 exact repair-non-invariant settings and 12 one-point near-tie settings across 10,000 resamples.

The full ResNet-50 repair matrix in the supplementary material provides a representative example on Imagenette at 92.5% sparsity. Under ERK and LAMP, ASR-q50 is the exact mean-argmax repair, although it is close to layerwise repair. Under Global allocation, layerwise repair is the strongest method (80.67±2.32), while raw channelwise variance matching falls to 43.14±3.35. The relevant point is therefore not that every local argmax gap is statistically separated, but that changing the allocation can alter which repair family remains reliable.

The standard deviations further show that transition settings can exhibit high seed variability under BN recalibration. For example, on ResNet-50/Imagenette at 92.5% sparsity, BN recalibration has large seed variation under both ERK (40.93±12.63) and LAMP (41.52±17.68). In the same settings, activation-statistic repairs achieve substantially higher means with much smaller variation. This suggests that, in the transition regime, stronger calibration-based repair can improve both the average recovered accuracy and the stability of recovery across pruning and calibration seeds.

6.2 Allocation Determines the Repair Ceiling

Table 2 summarizes representative cases from the allocation–repair matrix using three-seed means. ResNet-18 and ResNet-50 often favor LAMP in repair-sensitive transition settings. ResNet-34, however, does not support a uniform Global-dominance claim across all datasets: Global has its clearest and most consistent advantage on CIFAR-100, while on CIFAR-10 and Imagenette its margins over ERK or LAMP are often within the observed seed variation. Uniform allocation never provides the

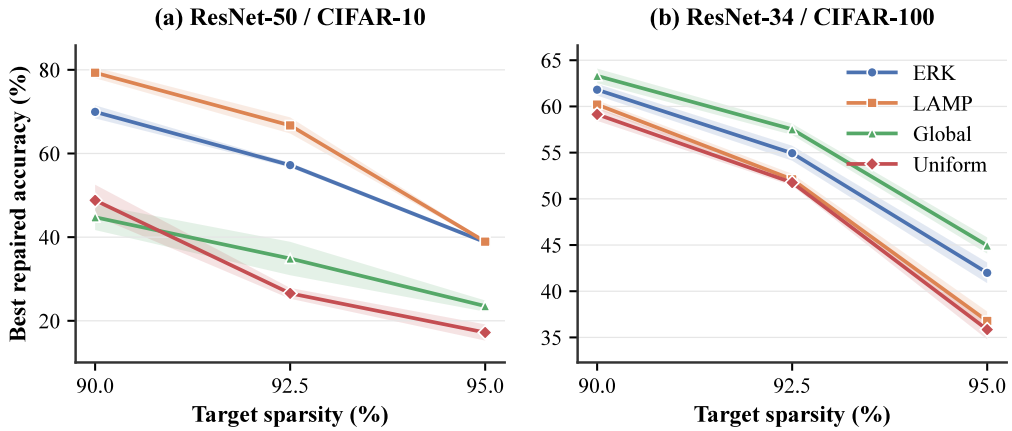


Figure 4: Best repaired accuracy under different sparsity allocations. Curves show the mean test accuracy over three seeds, and shaded bands indicate one standard deviation. The left panel shows a representative setting in which LAMP leaves the most recoverable sparse model, while the right panel shows a representative counterexample in which Global magnitude pruning is stronger.

best recovery in these representative cases and often falls sharply behind at high sparsity. These patterns show that allocation quality is not a universal ordering over pruning rules, but depends on the architecture, dataset, sparsity, and downstream repair method.

The ResNet-18/Imagenette 95% case remains a useful exception in which Global allocation slightly edges LAMP in mean accuracy (73.59 versus 73.18). Because this margin is small relative to the seed-level variation, we interpret it as evidence that Global is competitive in this setting rather than as a strong ordering claim. ResNet-50 also shows a near-collapse boundary case on CIFAR-100 where ERK becomes competitive at very high sparsity. These cases reinforce the same point: as Figure 4 illustrates, the preferred allocation changes with the recoverability regime rather than following a fixed ranking.

Table 2: Representative allocation effects after label-free repair. Each entry is the highest three-seed mean accuracy (%) over the repair suite for the given allocation; complete results appear in the supplementary material.

Arch.	Data	Sp.	ERK	LAMP	Global	Uniform
R18	C10	90.0	78.26	**79.77**	66.77	57.85
R18	C10	95.0	45.44	**50.59**	27.99	24.06
R18	Img	95.0	64.82	73.18	**73.59**	35.63
R34	C10	95.0	82.96	83.25	**85.64**	74.97
R34	C100	92.5	54.94	52.11	**57.53**	51.76
R34	C100	95.0	41.99	36.78	**44.95**	35.85
R50	C10	92.5	57.22	**66.71**	34.89	26.57
R50	C100	92.5	21.26	**25.31**	3.69	6.61
R50	Img	92.5	78.89	**85.78**	80.67	57.19
R50	Img	95.0	57.65	**60.33**	55.48	30.81

These patterns show why allocation and repair should be evaluated jointly. A repair method cannot recover signal that the allocation has removed. Conversely, an allocation that appears weak under BN recalibration can remain useful if it preserves activation structure that stronger repair can exploit. The preferred pipeline is therefore a property of the allocation–repair pair, not of either component in isolation. The method-level repair comparisons underlying these allocation summaries are reported in the supplementary material.

6.3 Regime Interpretation

Figure 2 shows that allocation changes the recoverability regime in 13 of the 27 main settings, while the allocation–repair matrices show that these effects are largest in repair-sensitive transition settings. In stable settings, many repair methods approach similar accuracy because BN recalibration already restores most of the available signal. In collapse settings, all label-free repairs remain low because too little coherent activation signal survives pruning. The largest allocation-dependent gaps occur between these extremes, where stronger repair can exploit preserved activation signal only when the allocation has left that signal in a coherent form.

Table 3: Calibration-only allocation selection and fixed-repair deployment summary on the 27 main ResNet settings. The upper block evaluates selectors against the best repaired allocation; the lower block evaluates fixed-ASR-q50 pipelines against the oracle allocation–repair pair. Regret is oracle accuracy minus selected accuracy, in percentage points.

Selector / pipeline	Exact	Within 1 pp	Within 1.5 pp	Mean regret	Max regret
Allocation selector evaluated with best repair					
LDMS-σ	25 / 27	26 / 27	26 / 27	0.103	2.684
Layerwise dispersion	20 / 27	21 / 27	23 / 27	1.191	13.003
Channel CV	16 / 27	21 / 27	23 / 27	0.907	8.167
LDMS-var	24 / 27	24 / 27	26 / 27	0.189	2.684
Full LDMS-var	24 / 27	25 / 27	26 / 27	0.191	3.113
Always-LAMP	16 / 27	21 / 27	23 / 27	0.907	8.167
BN-only ranking	16 / 27	18 / 27	19 / 27	2.986	26.537
Complete pipeline with fixed ASR-q50 repair					
LDMS-σ + ASR-q50	12 / 27	22 / 27	23 / 27	0.740	6.517
Always-LAMP + ASR-q50	7 / 27	16 / 27	19 / 27	1.577	8.167

Uniform and Global allocations serve as useful controls for this interpretation. Uniform applies the same sparsity to every repairable convolution, while Global is more selective and is strong on ResNet-34/CIFAR-100 but weaker than LAMP in several ResNet-50 transition settings. These controls reinforce that recoverability depends on the activation signal left by the allocation, not only on nominal sparsity. A supplementary ResNet-50/ImageNet-1K stress test gives a larger-scale collapse case where none of the repairs restores substantial accuracy.

6.4 Calibration-Based Recoverability Diagnostics

We next ask whether unlabeled calibration activations can rank allocations before running the full repair suite. For each architecture, dataset, and sparsity, we score the four candidate sparse models using dense–pruned activation statistics and compare the selected allocation with the allocation that achieves the highest repaired accuracy.

Table 3 shows that dense–pruned activation geometry is highly predictive of post-repair allocation quality. LDMS-σ matches the best repaired allocation in 25 of 27 settings and remains within one point in 26 settings. Its two misses occur at the highest ResNet-50 sparsity: one gives 0.11 points of regret on CIFAR-10, and the other gives the maximum regret of 2.684 points on Imagenette. With fixed ASR-q50 repair, LDMS-σ also improves over always using LAMP. Component rows show that layerwise dispersion gives more exact matches than Channel CV but also larger failures, while Channel CV captures complementary within-layer heterogeneity. Supplementary tests cover VGG16-BN, DenseNet-121, ImageNet-100, and ViT-Tiny (Touvron et al., 2021).

7 Conclusion

This work frames high-sparsity post-training pruning as a joint allocation–repair problem. At fixed global sparsity, allocation changes repaired accuracy, the preferred label-free repair method, and the recoverability regime reached by the sparse model. Across the main ResNet evaluation, these interactions are most pronounced in repair-sensitive transition settings, where stronger activation-statistic correction can recover substantial accuracy only when the allocation preserves usable activation structure.

We further introduce LDMS-σ, a calibration-only diagnostic that ranks candidate allocations from dense–pruned activation statistics before the full repair suite. LDMS-σ selects the best repaired allocation in 25 of 27 main settings and remains near-best in almost all others, showing that unlabeled calibration data can guide allocation choice. Overall, the results argue against treating sparsity allocation as a preprocessing detail before repair: allocation determines the activation geometry on which repair operates, making recoverability a property of the allocation–repair pair. This perspective suggests a practical route toward cheaper post-pruning deployment without exhaustive allocation–repair sweeps when labels or fine-tuning are unavailable.

Reproducibility Statement

The main experimental protocol, repair methods, allocation strategies, and LDMS-σ diagnostic are specified in Sections 3–5. Full repair matrices, sensitivity analyses, and additional architecture checks are provided in the appendix. The supplementary code package includes implementations of the allocation and repair methods, canonical machine-readable results, scripts for regenerating the main tables and figures, and automated consistency checks.

AI Use Statement

AI tools were used to assist with editing, code packaging, and formatting. All scientific claims, experimental values, tables, figures, and conclusions were reviewed by the authors.

Ethics Statement

This work studies post-training sparsification and label-free repair for vision models using standard image-classification benchmarks. It does not introduce new datasets, collect personal information, or involve human subjects. The main ethical considerations are transparent reporting and responsible deployment: sparse models should be evaluated in their intended application context, especially when compression may change failure modes, class-level performance, or robustness.

References

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.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Utku Evci, Trevor Gale, Jacob Menick, Pablo Samuel Castro, and Erich Elsen. Rigging the lottery: Making all tickets winners. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2943–2952. PMLR, 2020.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019.

Elias Frantar and Dan Alistarh. SparseGPT: Massive language models can be accurately pruned in one-shot. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10323–10337. PMLR, 2023.

Ruihao Gong, Yang Yong, Zining Wang, Jinyang Guo, Xiuying Wei, Yuqing Ma, and Xianglong Liu. Fast and controllable post-training sparsity: Learning optimal sparsity allocation with global constraint in minutes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(11): 12190–12198, 2024. doi: 10.1609/aaai.v38i11.29108.

Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. In *Advances in Neural Information Processing Systems*, volume 28, pp. 1135–1143, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Jeremy Howard. Imagenette: A smaller subset of 10 easily classified classes from ImageNet. https://github.com/fastai/imagenette, 2019. Accessed: 2026-07-25.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017.

Weizhong Huang, Yuxin Zhang, Xiawu Zheng, Fei Chao, and Rongrong Ji. Determining layer-wise sparsity for large language models through a theoretical perspective. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 26175–26201. PMLR, 2025.

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.

Keller Jordan, Hanie Sedghi, Olga Saukh, Rahim Entezari, and Behnam Neyshabur. REPAIR: REnormalizing permuted activations for interpolation repair. In *International Conference on Learning Representations*, 2023.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Ivan Lazarevich, Alexander Kozlov, and Nikita Malinin. Post-training deep neural network pruning via layer-wise calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 798–805, 2021.

Jaeho Lee, Sejun Park, Sangwoo Mo, Sungsoo Ahn, and Jinwoo Shin. Layer-adaptive sparsity for the magnitude-based pruning. In *International Conference on Learning Representations*, 2021.

Lujun Li, Peijie Dong, Zhenheng Tang, Xiang Liu, Qiang Wang, Wenhan Luo, Wei Xue, Qifeng Liu, Xiaowen Chu, and Yike Guo. Discovering sparsity allocation for layer-wise pruning of large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 141292–141317, 2024. doi: 10.52202/079017-4487.

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.

Markus Nagel, Rana Ali Amjad, Mart van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7197–7206. PMLR, 2020.

Dhananjay Saikumar and Blesson Varghese. Signal collapse in one-shot pruning: When sparse models fail to distinguish neural representations, 2025.

Gil Shomron and Uri Weiser. Post-training BatchNorm recalibration, 2020.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. A simple and effective pruning approach for large language models. In *International Conference on Learning Representations*, 2024.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10347–10357. PMLR, 2021.

Zining Wang, Jinyang Guo, Ruihao Gong, Yang Yong, Aishan Liu, Yushi Huang, Jiaheng Liu, and Xianglong Liu. PTSBench: A comprehensive post-training sparsity benchmark towards algorithms and models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9963–9972, 2024. doi: 10.1145/3664647.3680982.

Lu Yin, You Wu, Zhenyu Zhang, Cheng-Yu Hsieh, Yaqing Wang, Yiling Jia, Gen Li, Ajay Kumar Jaiswal, Mykola Pechenizkiy, Yi Liang, Michael Bendersky, Zhangyang Wang, and Shiwei Liu. Outlier weighed layerwise sparsity (OWL): A missing secret sauce for pruning LLMs to high sparsity. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57101–57115. PMLR, 2024.

Peiqi Yu, Jinhao Wang, Xinyi Sui, Nam Ling, Wei Wang, and Wei Jiang. Efficient post-training pruning of large language models with statistical correction, 2026.

Qishi Zhan, Ziheng Chen, and Minxuan Hu. Adaptive signal resuscitation: Channel-wise post-pruning repair for sparse vision networks, 2026. URL https://arxiv.org/abs/2605.21426.

Yingtao Zhang, Haoli Bai, Haokun Lin, Jialin Zhao, Lu Hou, and Carlo Vittorio Cannistraci. Plug-and-play: An efficient post-training pruning method for large language models. In *International Conference on Learning Representations*, 2024.

APPENDIX CONTENTS

A Supplement Overview 14

B Mechanistic Interpretation of LDMS-σ 14

C Calibration-Budget Sensitivity 15

D Allocation-Dependent Repair Rankings 15

E Near-Tie Sensitivity of Repair Ranking 16

F Repair Gains and Regime Sensitivity 17

G Full Repair Comparisons 19

H ImageNet-100 and DenseNet-121 19

I Held-Out Architecture Stress Tests 20

J Exploratory Transformer Validation 21

K ImageNet-1K Single-Seed Stress Test 23

L Clipped ASR Sensitivity 24

M Runtime of Label-Free Repair 24

A Supplement Overview

This supplementary material provides the analyses and full results supporting the main paper. It first gives a mechanistic interpretation of LDMS-σ and examines sensitivity to the unlabeled calibration budget. It then reports the allocation-specific repair identities, near-tie and seed-bootstrap checks, allocation-level regime labels, and the complete three-seed ResNet repair (He et al., 2016) matrices underlying the main 27-setting analysis. Additional sections present ImageNet-100 (Deng et al., 2009), DenseNet-121 (Huang et al., 2017), VGG16-BN (Simonyan & Zisserman, 2014; Ioffe & Szegedy, 2015), ViT-Tiny (Touvron et al., 2021), and ImageNet-1K (Deng et al., 2009) stress tests, followed by clipping sensitivity and runtime measurements. Unless explicitly stated otherwise, supplementary analyses use the same pruning, calibration, and evaluation protocols as the main paper. The additional results document the robustness, scope, and limitations of the allocation–repair interaction; they do not alter the main aggregate counts.

B Mechanistic Interpretation of LDMS-σ

LDMS is intended as a diagnostic for the activation geometry left by an allocation, not as a theoretical bound on accuracy. Let $a^d_{\ell,c}(x)$ and $a^p_{\ell,c}(x)$ denote the dense and pruned activations for channel c in layer ℓ on calibration input x, and define the preserved standard-deviation ratio

$$q_{\ell,c} = \frac{\sigma_{x\in\mathcal{C}}(a^p_{\ell,c}(x))}{\sigma_{x\in\mathcal{C}}(a^d_{\ell,c}(x)) + \epsilon}.$$

Label-free repair methods can only use calibration statistics of the dense and pruned models. They can rescale or recalibrate existing activation signal, but they cannot restore channels whose post-pruning activation signal has effectively collapsed (Saikumar & Varghese, 2025). This is the activation-geometry view underlying our main results: recoverability depends not only on global sparsity, but also on how coherently activation scale is preserved across layers and channels.

The first LDMS term,

$$\frac{\sigma_\ell(\mathrm{median}_c\, q_{\ell,c})}{\mathrm{median}_{\ell,c}\, q_{\ell,c} + \epsilon},$$

measures imbalance in preserved activation scale across layers. Large values indicate that some layers remain near the dense activation scale while others are strongly compressed. This imbalance makes lightweight recalibration less reliable and limits repair rules that apply similar corrections across layers. Medians over channels and over all layer–channel pairs reduce the influence of outliers and unusually amplified ratios.

The second term,

$$\mathrm{median}_\ell\, \mathrm{CV}_c(q_{\ell,c}),$$

measures within-layer channel heterogeneity. High channel heterogeneity means that a layer contains channels with substantially different preserved activation ratios. In this case, raw channelwise matching can produce unstable correction factors for channels whose pruned signal is weak, while layerwise correction can over-correct or under-correct substantial subsets of channels. This motivates reliability-aware repair variants such as shrinkage and clipping, although their effectiveness remains an empirical property of the allocation and architecture.

LDMS therefore combines two activation mismatches directly relevant to label-free repair: cross-layer scale imbalance and within-layer channel instability. The variance-ratio version, $\mathrm{LDMS}_{\mathrm{var}}$, uses

$$r_{\ell,c} = \frac{\mathrm{Var}_{x\in\mathcal{C}}(a^p_{\ell,c}(x))}{\mathrm{Var}_{x\in\mathcal{C}}(a^d_{\ell,c}(x)) + \epsilon}$$

in place of $q_{\ell,c}$ and provides the counterpart most directly aligned with variance-preservation repair. The full $\mathrm{LDMS}_{\mathrm{var}}$ variant further adds median activation mean shift. Its competitive performance indicates that mean displacement carries some recoverability signal, but the scale-dispersion and channel-heterogeneity terms provide the clearest allocation-ranking signal in the main matrix.

Simple collapse statistics are therefore insufficient. Counting nearly dead channels captures extreme failure but misses transition regimes in which channels remain active while their geometry is difficult to repair.

Table 4: Calibration-budget sensitivity of LDMS-σ allocation selection on the 27 main ResNet settings. Results use seed 0 and are intended as a sensitivity check rather than a separate multi-seed benchmark. For each calibration budget, LDMS-σ ranks four candidate allocations within the same architecture, dataset, and sparsity setting. Regret is measured relative to the best repaired allocation under the same budget.

Cal. size	Exact	$\leq$1 pp	$\leq$1.5 pp	Mean regret
32	22 / 27	25 / 27	25 / 27	0.321
64	26 / 27	26 / 27	26 / 27	0.229
128	24 / 27	26 / 27	26 / 27	0.214
256	23 / 27	26 / 27	26 / 27	0.225

C Calibration-Budget Sensitivity

We examine whether the LDMS-σ allocation ranking depends strongly on the number of unlabeled calibration images. Using seed 0, we repeat the allocation selection analysis on the 27 main ResNet settings using 32, 64, 128, or 256 calibration images. For each architecture, dataset, and sparsity, LDMS-σ ranks the four candidate allocations. Regret is measured relative to the best repaired allocation under the same budget (Table 4).

The LDMS-σ ranking is largely stable across calibration budgets. The selected allocation is unchanged across all four budgets in 26 of the 27 main settings. Exact selection varies mildly because several settings have allocation gaps below one percentage point, but near-best selection is stable: for $n \geq 64$, LDMS-σ selects an allocation within 1.5 percentage points of the best in 26 of 27 settings. The remaining large-regret case is the same ResNet-50/Imagenette/95% boundary setting across budgets, suggesting a persistent diagnostic failure near a sharp transition-to-collapse boundary rather than instability caused by the calibration budget.

We also check whether the attainable repaired accuracy changes materially with calibration size. For each of the 27 settings, we compute the best repaired accuracy over allocations and repair protocols at each calibration budget and measure the range across $n \in \{32, 64, 128, 256\}$. The median range is only 0.14 percentage points, the mean range is 0.24 points, and the largest observed range is 1.83 points. Thus, in this sensitivity sweep, changing the calibration budget does not materially alter either the attainable repaired accuracy or the LDMS-σ allocation ranking.

D Allocation-Dependent Repair Rankings

The main paper reports that the strongest label-free repair protocol is not fixed across allocations. Table 5 lists the underlying identities for all 27 combinations of architecture, dataset, and sparsity. Each row fixes that combination, while the columns vary the allocation. If repair were independent of allocation, a row would tend to select the same protocol in all four columns. Instead, the preferred repair changes in most settings, showing that pruning-induced activation geometry changes which repair rule is most effective.

Several patterns are visible. Channelwise or ASR-style repair is often favored in ResNet-18 transition settings, while clipping becomes more common in fragile ResNet-50 settings. BN recalibration (Shomron & Weiser, 2020; Ioffe & Szegedy, 2015) is selected where stronger correction has little headroom or becomes unstable. These patterns support evaluating repair under more than one allocation.

The table also shows that repair non-invariance is not restricted to a single architecture or dataset. ResNet-18, ResNet-34, and ResNet-50 all contain rows in which the best repair protocol changes across allocations. This reinforces that the allocation–repair interaction is a recurring phenomenon in the main matrix rather than an artifact of one particular model.

Table 5: Best label-free repair protocol under each allocation for the 27 main ResNet architecture–dataset–sparsity settings. Entries are selected by exact three-seed mean accuracy within each allocation. The exact mean-argmax repair identity changes across allocations in 21 of 27 settings. Because several top methods are close, Table 6 reports a near-tie sensitivity analysis.

Setting	ERK	LAMP	Global	Uniform
R18/C10/90.0	Channelwise	ASR-q50	Channelwise	ASR-q50
R18/C10/92.5	Channelwise	Channelwise	ASR-q50	ASR-q50
R18/C10/95.0	Channelwise	Channelwise	Channelwise	Channelwise
R18/C100/90.0	Channelwise	Clip	BN	Clip
R18/C100/92.5	Channelwise	Channelwise	BN	BN
R18/C100/95.0	Channelwise	Channelwise	ASR-q50	Channelwise
R18/Img/90.0	Channelwise	Channelwise	Clip	Clip
R18/Img/92.5	Channelwise	Channelwise	Clip	BN
R18/Img/95.0	ASR-q50	Channelwise	Clip	Clip
R34/C10/90.0	ASR-q50	ASR-q50	Clip	Clip
R34/C10/92.5	ASR-q50	ASR-q50	ASR-q50	Clip
R34/C10/95.0	ASR-q50	ASR-q50	ASR-q50	Clip
R34/C100/90.0	ASR-q50	ASR-q50	ASR-q50	ASR-q50
R34/C100/92.5	ASR-q50	ASR-q50	ASR-q50	Clip
R34/C100/95.0	ASR-q50	ASR-q50	ASR-q50	Clip
R34/Img/90.0	ASR-q50	Channelwise	Clip	Clip
R34/Img/92.5	ASR-q50	ASR-q50	ASR-q50	ASR-q50
R34/Img/95.0	ASR-q50	ASR-q50	ASR-q50	ASR-q50
R50/C10/90.0	ASR-q50	ASR-q50	Clip	Clip
R50/C10/92.5	ASR-q50	ASR-q50	Clip	Clip
R50/C10/95.0	ASR-q50	ASR-q50	Clip	Clip
R50/C100/90.0	Clip	Layerwise	Clip	Clip
R50/C100/92.5	Clip	Clip	Clip	Clip
R50/C100/95.0	Clip	Clip	Clip	Clip
R50/Img/90.0	ASR-q50	ASR-q50	Layerwise	Layerwise
R50/Img/92.5	ASR-q50	ASR-q50	Layerwise	Layerwise
R50/Img/95.0	ASR-q50	ASR-q50	ASR-q50	Clip

Table 6: Near-tie sensitivity of repair-ranking changes across allocations. A setting is counted as having no allocation-invariant near-best repair when no single repair protocol lies within the stated tolerance of the best repair for all four allocations.

Tolerance	Count	Interpretation
Exact argmax	21 / 27	Different exact best
0.5 pp	18 / 27	No common near-best
1.0 pp	12 / 27	No common near-best
1.5 pp	7 / 27	No common near-best

E Near-Tie Sensitivity of Repair Ranking

The exact identities in Table 5 summarize the allocation–repair matrix but can be sensitive when top methods have similar three-seed means. We therefore treat all protocols within a fixed tolerance of the allocation-specific best as tied. For each architecture, dataset, and sparsity, we ask whether one protocol is near-best for all four allocations. If none exists, the setting retains allocation-dependent repair preference under that tolerance.

The interaction is strongest under exact argmax, but it does not disappear when small gaps are treated as ties. With a one-percentage-point tolerance, 12 of 27 settings still have no repair protocol that is near-best under every allocation. The seed bootstrap in Table 7 gives the same qualitative picture: the median counts are 23 exact repair-non-invariant settings and 12 one-point near-tie settings across 10,000 resamples. We therefore use the 21 of 27 exact count as a descriptive summary, while the near-tie and bootstrap analyses give more conservative views of repair non-invariance.

Table 7: Seed-bootstrap sensitivity of repair-ranking non-invariance across the 27 main ResNet settings. We resample the three seed runs with replacement within each allocation–repair cell, recompute the allocation-specific mean repair ranking, and repeat the exact and one-point near-tie counts over 10,000 bootstrap resamples.

Statistic	Mean	Median	95% interval
Exact non-invariance	22.64	23	21–25
No common near-best (1 pp)	12.13	12	10–15

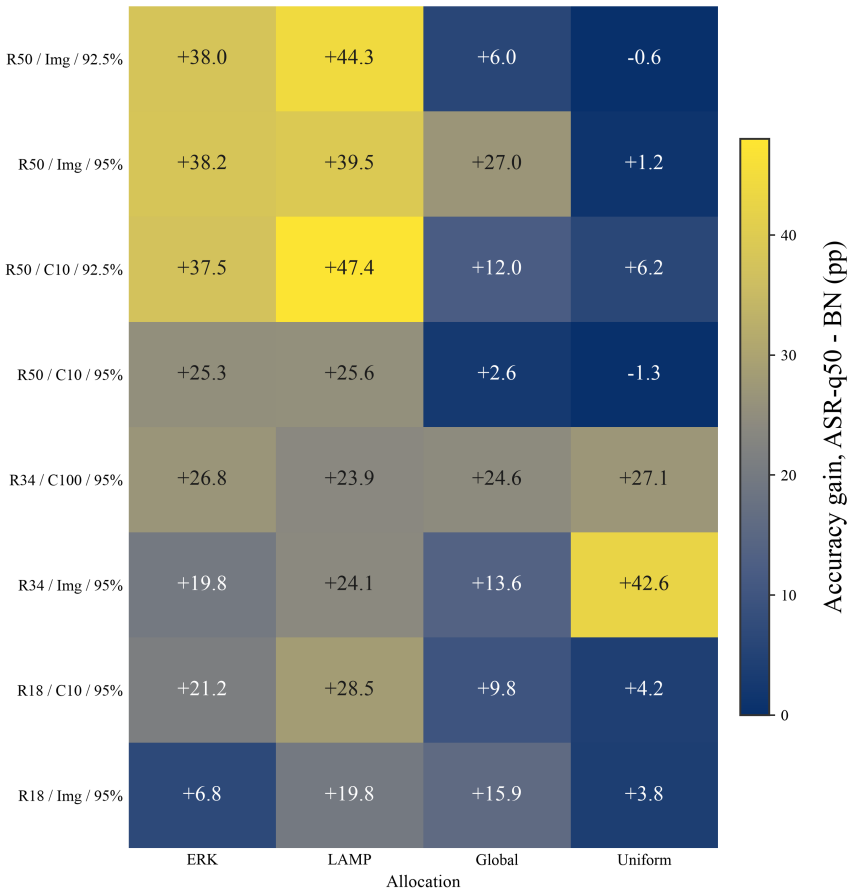


Figure 5: Gain of ASR-q50 over BN recalibration in representative transition settings. Each cell reports the difference in mean test accuracy (ASR-q50 minus BN, percentage points) for a fixed architecture, dataset, sparsity level, and allocation. The gains vary substantially across allocations, indicating that stronger label-free repair does not provide a constant offset over BN recalibration but depends on the activation structure left by pruning.

F Repair Gains and Regime Sensitivity

The allocation–repair interaction appears both quantitatively, through the gain of stronger repair over BN recalibration, and qualitatively, through the recoverability regime reached under each allocation. Figure 5 shows representative transition settings in which ASR-q50 provides substantially different gains across allocations at the same architecture, dataset, and global sparsity. The gain is therefore not a constant correction applied after pruning; it depends on the activation structure preserved by the allocation.

The stable, transition, and collapse regimes are assigned at the level of an architecture–dataset–sparsity–allocation tuple. This allocation-level definition is important because a fixed architecture,

Table 8: Allocation-specific regime labels for the 27 main ResNet settings. Labels are assigned separately for each architecture–dataset–sparsity– allocation combination. *Stable* indicates that BN recalibration is within 5 percentage points of the best evaluated label-free repair under the same allocation. *Collapse* indicates that the best evaluated label-free repair recovers less than 20% of the dense-to-chance accuracy gap. All remaining cases are labeled *transition*. A stable label does not necessarily imply high absolute accuracy; it indicates that stronger evaluated repair methods have limited headroom beyond BN recalibration under that allocation.

Setting	ERK	LAMP	Global	Uniform
R18/C10/90.0	Transition	Transition	Transition	Stable
R18/C10/92.5	Transition	Transition	Transition	Stable
R18/C10/95.0	Transition	Transition	Transition	Collapse
R18/C100/90.0	Transition	Transition	Stable	Stable
R18/C100/92.5	Transition	Transition	Collapse	Collapse
R18/C100/95.0	Collapse	Collapse	Collapse	Collapse
R18/Img/90.0	Stable	Stable	Stable	Stable
R18/Img/92.5	Stable	Transition	Stable	Stable
R18/Img/95.0	Transition	Transition	Transition	Transition
R34/C10/90.0	Transition	Transition	Stable	Transition
R34/C10/92.5	Transition	Transition	Transition	Transition
R34/C10/95.0	Transition	Transition	Transition	Transition
R34/C100/90.0	Transition	Transition	Transition	Transition
R34/C100/92.5	Transition	Transition	Transition	Transition
R34/C100/95.0	Transition	Transition	Transition	Transition
R34/Img/90.0	Stable	Stable	Stable	Transition
R34/Img/92.5	Transition	Transition	Transition	Transition
R34/Img/95.0	Transition	Transition	Transition	Transition
R50/C10/90.0	Transition	Transition	Transition	Transition
R50/C10/92.5	Transition	Transition	Transition	Collapse
R50/C10/95.0	Transition	Transition	Collapse	Collapse
R50/C100/90.0	Transition	Transition	Collapse	Stable
R50/C100/92.5	Transition	Transition	Collapse	Collapse
R50/C100/95.0	Collapse	Collapse	Collapse	Collapse
R50/Img/90.0	Transition	Transition	Stable	Transition
R50/Img/92.5	Transition	Transition	Transition	Transition
R50/Img/95.0	Transition	Transition	Transition	Transition

Table 9: Sensitivity of allocation-dependent regime changes to the descriptive stable and collapse thresholds. Each entry counts the number of main architecture–dataset–sparsity settings, out of 27, in which the four allocations span at least two regime labels.

Stable threshold	15%	20%	25%	30%
3 pp	9	10	11	12
5 pp	12	13	14	15
7 pp	14	14	15	15

dataset, and global sparsity can receive different regime labels depending only on how sparsity is distributed across layers. Table 8 reports the resulting labels for the 27 main ResNet settings.

The table should be read row-wise. When all four allocations share the same label, recoverability is largely determined by the architecture, dataset, and global sparsity. When labels differ within a row, changing only the allocation changes the qualitative behavior of the sparse model. These within-row differences show that allocation affects not only the final repaired accuracy, but also whether the model remains stable, becomes repair-sensitive, or enters an empirical collapse regime.

Because these regimes are empirical descriptors rather than universal classes, we examine their sensitivity to the two defining thresholds. Table 9 varies the stable threshold from 3 to 7 percentage points and the collapse threshold from 15% to 30% of the recovered dense-to-chance accuracy gap.

Under the default thresholds, 13 of the 27 settings contain at least two different regime labels across allocations. Thus, the same architecture, dataset, and global sparsity can exhibit stable, transition, or collapse behavior solely because the sparsity allocation changes. This conclusion is not tied to one exact threshold choice. Across stable thresholds of 3, 5, and 7 percentage points and collapse thresholds of 15–30% of the recovered dense-to-chance gap, the four allocations span at least two regimes in 9–15 of the 27 settings. The regime analysis therefore provides a qualitative summary of the same allocation–repair interaction measured quantitatively by repair ceilings and LDMS-σ regret in the main paper.

G Full Repair Comparisons

The 27-setting analysis uses sparsities of 90.0%, 92.5%, and 95.0%. The ResNet-50/CIFAR-100 table also includes 95.5% as a diagnostic point; this extra row is excluded from every 27-setting count. Tables 10–12 report the full three-seed repair matrix used to derive the main results. Each table fixes an architecture and varies dataset, sparsity, allocation, and repair protocol. These tables serve three purposes. First, they show the absolute accuracy levels behind the repair ceilings reported in the main text. Second, they make clear that the best repaired allocation is not determined by global sparsity alone. Third, they show that the strongest repair protocol can change with architecture, dataset, sparsity, and allocation.

The most informative comparisons are within each block defined by dataset and sparsity. All four allocations then share the global sparsity and repair suite. Differences across rows therefore reflect where sparsity is placed rather than how much is applied. This structure is the empirical basis for treating allocation and repair as coupled choices.

Taken together, the full matrices show that allocation effects are largest in repair-sensitive transition settings. In stable settings, many repair protocols approach similar accuracy because BN recalibration already restores most of the available signal. In collapse settings, all methods remain low. The largest allocation-dependent gaps occur between these extremes, where stronger repair can exploit preserved activation signal only when the allocation has left that signal in a coherent form.

H ImageNet-100 and DenseNet-121

We use two supplementary studies to test the main interpretation beyond the primary ResNet matrix. The ImageNet-100 (Deng et al., 2009) experiment uses ResNet-18 (He et al., 2016) with ERK and LAMP allocations (Evci et al., 2020; Lee et al., 2021) and therefore serves as a data-scale check rather than a complete allocation-selection benchmark. The DenseNet-121 experiments (Huang et al., 2017) provide an architecture check with a different connectivity pattern and activation ordering.

Table 13 shows that clipped activation-statistic repair consistently gives the best or tied-best result among the evaluated repair methods on ImageNet-100, but the margins over BN recalibration are modest. This suggests that the repair-sensitive region is narrower in this setting: stronger repair helps, but the allocation-dependent gaps are less dramatic than in the hardest CIFAR transition cases. Since this experiment evaluates only ERK and LAMP, we use it as supporting evidence for data-scale consistency rather than as a full LDMS allocation-selection test.

Tables 14–16 show a different kind of boundary behavior. DenseNet-121 remains recoverable at lower sparsities, but the transition to collapse occurs earlier than in the main ResNet experiments. On CIFAR-10 and CIFAR-100, useful recovery is visible at 70–80% sparsity, while many 85–90% settings have much lower repaired accuracy. On Imagenette, Global magnitude pruning often leaves the most recoverable model among the evaluated allocations. These results reinforce that recoverability regimes are architecture-dependent and that the best allocation can change when the activation geometry changes.

The DenseNet results are a scope check rather than a repeat of the ResNet matrix. They show that allocation matters and that the transition boundary depends on architecture. The next section examines LDMS-σ transfer.

Table 10: Full repair comparison for ResNet-18 across all four allocations (He et al., 2016; Krizhevsky, 2009; Howard, 2019). Entries report mean±standard deviation test accuracy (%) over three seeds. The best repair method for each allocation setting is shown in **bold**.

Dataset	Sparsity	Alloc.	No repair	BN	Layerwise	Channelwise	ASR-q50	Clip
CIFAR-10	90.0%	ERK	10.10±0.17	62.84±1.48	74.40±2.77	**78.26±1.30**	77.21±1.51	76.31±1.72
		LAMP	11.18±1.03	66.28±1.84	79.43±1.03	79.26±0.93	**79.77±1.09**	78.72±0.69
		Global	9.85±0.19	58.03±0.80	24.40±6.08	**66.77±1.39**	66.36±1.10	64.90±0.83
		Uniform	9.92±0.13	53.50±0.37	23.85±4.15	54.80±0.88	**57.85±1.17**	57.34±0.71
	92.5%	ERK	10.05±0.08	36.28±2.56	45.58±5.25	**66.89±2.13**	64.32±2.87	61.87±2.76
		LAMP	10.09±0.16	37.24±4.23	60.44±4.25	**69.82±1.36**	68.58±1.66	65.20±2.36
		Global	9.98±0.03	32.06±2.56	20.21±5.32	44.13±5.06	**44.81±4.38**	41.12±4.18
		Uniform	10.00±0.01	32.33±2.23	14.84±4.33	33.10±0.89	**36.60±2.19**	35.74±2.22
	95.0%	ERK	10.00±0.00	15.79±2.54	16.52±2.62	**45.44±2.63**	36.98±2.57	34.64±2.60
		LAMP	10.00±0.00	15.62±1.74	24.53±0.66	**50.59±1.73**	44.08±2.68	37.65±1.36
		Global	10.00±0.00	15.15±2.34	15.15±1.60	**27.99±3.73**	24.97±4.38	20.76±2.30
		Uniform	10.00±0.00	18.05±0.69	11.63±2.24	**24.06±0.62**	22.28±1.00	21.68±1.39
CIFAR-100	90.0%	ERK	1.43±0.29	32.73±1.76	35.42±1.87	**39.59±0.63**	38.75±1.22	38.89±1.12
		LAMP	1.31±0.36	33.47±1.40	39.71±1.18	40.03±0.96	40.50±0.87	**40.61±1.23**
		Global	0.90±0.17	**17.16±0.49**	3.20±0.88	10.75±1.85	15.85±0.50	16.03±0.52
		Uniform	0.96±0.21	22.48±0.86	3.37±0.78	19.60±0.30	22.97±1.28	**23.24±1.21**
	92.5%	ERK	1.08±0.21	17.23±1.19	14.08±1.76	**25.58±0.30**	24.69±0.94	24.52±1.03
		LAMP	1.15±0.21	18.17±1.06	22.54±1.86	**27.13±0.25**	26.85±0.63	26.67±0.49
		Global	1.01±0.07	**9.20±0.07**	1.66±0.28	5.45±0.90	8.75±0.82	8.76±0.60
		Uniform	1.15±0.24	**9.67±0.49**	1.57±0.42	7.77±0.38	9.12±0.69	9.46±0.50
	95.0%	ERK	1.02±0.03	5.22±1.33	1.57±0.58	**11.44±1.15**	9.96±0.20	9.94±0.35
		LAMP	1.20±0.10	4.55±1.23	4.85±2.22	**12.24±1.01**	11.52±0.32	11.51±0.54
		Global	1.00±0.01	2.12±0.19	0.95±0.10	2.57±0.23	**3.20±0.15**	3.17±0.25
		Uniform	0.98±0.04	2.76±0.57	1.07±0.08	**4.06±0.25**	3.90±0.43	3.83±0.48
Imagenette	90.0%	ERK	18.78±4.75	89.28±1.36	90.24±0.23	**91.36±0.62**	90.78±0.56	91.03±0.51
		LAMP	20.70±2.85	90.14±1.57	92.29±0.37	**92.50±0.65**	92.16±0.48	92.06±0.50
		Global	14.97±2.66	90.45±1.00	90.91±0.61	91.14±1.28	91.18±0.87	**91.34±0.97**
		Uniform	12.64±1.37	72.35±3.04	52.78±5.32	72.72±1.57	75.46±1.04	**75.77±0.64**
	92.5%	ERK	13.47±3.08	81.23±2.50	79.16±1.70	**85.56±1.45**	84.51±0.99	84.85±1.56
		LAMP	17.76±4.36	82.15±2.32	86.91±0.83	**88.06±1.33**	87.24±0.98	87.34±1.24
		Global	9.72±0.85	84.90±1.68	85.61±1.19	87.18±1.30	86.79±1.16	**87.30±1.30**
		Uniform	10.17±0.01	**50.82±1.45**	27.15±3.88	47.04±1.28	49.55±1.67	50.11±0.39
	95.0%	ERK	10.25±0.32	58.02±3.34	33.08±5.70	64.30±2.33	**64.82±1.62**	64.67±1.75
		LAMP	12.12±3.34	52.46±4.24	65.66±3.23	**73.18±3.28**	72.22±2.17	72.25±2.54
		Global	9.44±0.60	57.03±3.14	64.10±1.73	72.85±1.61	72.95±0.70	**73.59±1.15**
		Uniform	10.07±0.01	30.57±1.61	15.80±2.84	33.86±1.99	34.32±0.67	**35.63±0.91**

I Held-Out Architecture Stress Tests

We next test allocation–repair coupling and LDMS-σ beyond ResNets. The VGG16-BN (Simonyan & Zisserman, 2014; Ioffe & Szegedy, 2015) and DenseNet-121 (Huang et al., 2017) experiments ask whether recovery still depends on allocation under different activation geometries and whether the same calibration-only diagnostic ranks the candidates. LDMS-σ is not tuned on these settings.

Table 17 reports LDMS-σ selection. Regret is measured against the best repaired allocation for the same model, dataset, and sparsity.

The stress tests show useful transfer beyond ResNets. For VGG16-BN, exact selection is lower than in the main matrix, but the chosen allocation is usually near-best with small regret. DenseNet-121 shows partial transfer, with a larger failure near the CIFAR-100 transition boundary.

Overall, LDMS-σ captures a useful label-free recoverability signal across related activation geometries, although architecture-specific structure can affect selector strength.

Table 11: Full repair comparison for ResNet-34 across all four allocations. Entries report mean±standard deviation test accuracy (%) over three seeds. The best repair method for each allocation setting is shown in **bold**.

Dataset	Sparsity	Alloc.	No repair	BN	Layerwise	Channelwise	ASR-q50	Clip
CIFAR-10	90.0%	ERK	19.44±3.77	83.54±1.87	89.49±0.79	89.68±0.54	**90.08±0.74**	90.00±0.72
		LAMP	19.18±3.25	80.50±2.10	89.10±0.85	89.70±0.33	**89.87±0.60**	89.59±0.49
		Global	17.18±2.64	85.89±2.46	90.06±0.67	89.93±0.61	90.50±0.56	**90.53±0.54**
		Uniform	11.37±1.48	76.54±4.58	87.46±1.53	86.98±1.09	87.87±0.78	**87.98±0.89**
	92.5%	ERK	16.97±3.22	72.40±2.69	87.29±0.84	87.80±0.39	**88.25±0.59**	87.76±0.69
		LAMP	14.07±0.67	64.77±2.63	86.48±0.34	87.85±0.43	**87.99±0.34**	87.08±0.37
		Global	14.69±2.25	77.32±4.65	88.25±0.99	88.70±0.45	**89.13±0.41**	88.72±0.59
		Uniform	10.96±0.86	54.27±7.19	83.72±2.01	82.68±1.02	83.84±1.03	**84.65±0.95**
	95.0%	ERK	12.25±1.40	41.60±6.46	79.83±1.14	82.10±0.67	**82.96±0.95**	81.06±0.71
		LAMP	10.49±0.55	32.40±4.27	77.46±0.25	82.68±1.04	**83.25±0.92**	78.11±0.93
		Global	12.67±2.80	52.33±7.46	81.87±1.35	84.79±0.19	**85.64±0.45**	83.15±0.64
		Uniform	10.02±0.03	21.57±4.26	74.06±2.31	71.33±1.80	73.55±1.12	**74.97±1.19**
CIFAR-100	90.0%	ERK	4.92±0.81	51.02±0.51	60.45±0.19	61.24±1.01	**61.81±0.63**	61.13±0.48
		LAMP	4.95±0.80	46.18±1.30	57.90±0.40	59.71±0.62	**60.20±0.47**	58.63±0.45
		Global	2.71±1.14	55.94±1.15	62.34±0.68	61.93±1.05	**63.31±0.79**	63.06±0.63
		Uniform	1.87±0.48	45.07±0.61	58.65±0.52	57.08±1.24	**59.14±0.70**	59.00±0.73
	92.5%	ERK	2.36±0.75	36.21±1.69	52.80±0.82	54.15±1.04	**54.94±0.81**	53.53±0.86
		LAMP	2.56±1.08	30.22±2.39	49.04±1.26	51.48±0.64	**52.11±0.60**	50.05±1.10
		Global	1.81±0.68	43.51±1.01	56.76±0.21	55.82±1.22	**57.53±0.61**	57.17±0.75
		Uniform	1.37±0.50	24.90±0.51	51.55±0.49	48.80±1.36	51.02±0.79	**51.76±0.34**
	95.0%	ERK	1.52±0.54	15.22±1.16	38.72±1.36	39.27±1.10	**41.99±1.12**	39.54±1.58
		LAMP	1.80±0.67	12.86±1.02	33.18±2.00	35.28±1.00	**36.78±1.03**	33.80±2.05
		Global	1.28±0.46	20.40±1.22	43.45±0.53	42.40±1.45	**44.95±0.88**	43.72±0.86
		Uniform	1.15±0.26	6.82±0.31	35.54±0.91	31.51±1.32	33.90±1.51	**35.85±1.06**
Imagenette	90.0%	ERK	27.84±11.19	85.95±0.91	88.84±0.86	89.38±1.05	**89.39±0.88**	89.02±0.59
		LAMP	30.50±8.06	84.95±0.35	88.74±0.93	**89.27±1.00**	89.17±0.97	89.02±0.75
		Global	36.82±10.79	87.06±0.38	89.38±0.66	89.65±0.76	89.73±0.74	**89.74±0.71**
		Uniform	15.63±7.00	75.96±4.77	85.44±0.64	85.37±0.10	86.62±0.50	**86.74±0.37**
	92.5%	ERK	19.50±6.41	81.27±1.30	87.80±0.81	88.42±1.01	**88.55±0.81**	87.96±0.72
		LAMP	21.63±7.05	78.89±1.23	87.36±0.94	88.05±0.97	**88.25±0.82**	87.62±0.77
		Global	26.17±5.01	83.79±0.23	88.21±0.86	88.82±1.05	**88.86±0.90**	88.69±0.91
		Uniform	16.60±3.49	60.83±7.72	79.75±2.05	79.97±1.30	**82.95±0.51**	82.82±0.53
	95.0%	ERK	13.54±4.79	65.65±4.95	83.24±0.80	84.78±1.24	**85.47±0.62**	84.44±0.30
		LAMP	16.06±1.41	61.43±3.29	82.65±1.46	85.27±1.65	**85.56±0.94**	83.62±0.98
		Global	20.25±5.20	72.76±1.94	84.66±0.80	85.98±0.95	**86.40±0.76**	85.52±0.56
		Uniform	11.67±0.93	31.73±11.54	59.92±6.17	65.56±4.60	**74.35±1.55**	74.35±2.20

J Exploratory Transformer Validation

We additionally examine whether the LDMS-σ allocation-ranking signal extends beyond convolutional architectures to a LayerNorm-based Transformer. This pilot evaluates ViT-Tiny (Touvron et al., 2021) on CIFAR-10 (Krizhevsky, 2009) at 70% and 80% global sparsity using ERK, LAMP, and Uniform allocations. Because LayerNorm does not maintain running activation statistics, BN recalibration is not applicable in this setting. The *No repair* baseline therefore evaluates the pruned model directly without any calibration-based adjustment.

To retain a comparable label-free repair setting, ASR-q50 is applied using dense–pruned activation statistics collected from the same unlabeled calibration set. The final hidden dimension is treated as the channel axis, with statistics aggregated over the remaining calibration and token dimensions. LDMS-σ is computed from the corresponding dense–pruned activation ratios and selects one allocation at each sparsity level (Table 18).

LDMS-σ selects LAMP, the allocation with the highest repaired accuracy, at both evaluated sparsity levels. The selected accuracies are 66.8% at 70% and 49.3% at 80%, yielding zero selection regret. Their margins over the second-best allocations are 1.6 and 1.8 percentage points. Repair identity

Table 12: Full repair comparison for ResNet-50 across all four allocations. Entries report mean±standard deviation test accuracy (%) over three seeds. The best repair method for each allocation setting is shown in **bold**.

Dataset	Sparsity	Alloc.	No repair	BN	Layerwise	Channelwise	ASR-q50	Clip
CIFAR-10	90.0%	ERK	10.00±0.01	34.26±3.93	38.20±5.23	63.47±1.56	**69.95±1.57**	66.04±2.23
		LAMP	12.13±2.82	40.72±1.58	78.60±2.28	76.11±0.82	**79.30±1.30**	77.94±1.80
		Global	10.00±0.01	29.20±2.38	19.96±4.85	17.28±2.73	42.19±2.79	**44.76±3.06**
		Uniform	9.77±0.40	31.98±1.81	17.71±3.84	31.10±2.36	48.73±4.39	**48.81±3.73**
	92.5%	ERK	9.97±0.06	19.75±5.07	24.02±4.46	45.70±4.13	**57.22±0.57**	52.35±2.06
		LAMP	10.61±1.51	19.29±6.04	53.28±3.46	58.40±2.59	**66.71±1.93**	60.28±3.42
		Global	10.00±0.00	18.20±4.79	14.17±2.66	13.58±1.52	30.19±4.06	**34.89±4.00**
		Uniform	10.00±0.00	16.67±3.56	11.49±2.23	14.48±1.11	22.92±2.53	**26.57±1.29**
	95.0%	ERK	10.53±1.39	13.47±4.33	10.31±0.25	24.05±2.99	**38.81±0.53**	33.67±1.35
		LAMP	10.00±0.00	13.33±5.41	12.13±2.52	27.35±3.75	**38.92±0.31**	30.17±1.95
		Global	10.00±0.00	14.28±5.17	12.02±2.65	10.36±1.37	16.88±1.13	**23.52±1.34**
		Uniform	9.96±0.07	12.33±2.43	10.09±0.13	10.65±1.14	11.06±1.29	**17.21±1.94**
CIFAR-100	90.0%	ERK	1.24±0.35	18.15±0.92	9.61±0.63	25.68±0.89	32.62±0.78	**33.05±1.41**
		LAMP	1.11±0.13	18.65±1.11	**44.06±0.79**	32.26±1.58	40.85±1.07	42.32±0.34
		Global	0.98±0.28	4.67±0.78	1.60±0.62	1.37±0.29	3.23±0.54	**5.58±0.97**
		Uniform	1.08±0.13	15.69±0.24	3.24±1.54	5.97±1.07	18.93±1.73	**19.69±1.28**
	92.5%	ERK	1.01±0.02	7.78±0.83	5.39±0.69	12.34±0.98	20.04±0.51	**21.26±1.32**
		LAMP	1.00±0.00	5.99±2.13	19.29±0.83	13.70±0.27	22.99±0.59	**25.31±1.21**
		Global	1.19±0.32	3.25±1.24	2.03±0.93	1.25±0.21	1.60±0.09	**3.69±0.60**
		Uniform	1.00±0.00	4.46±0.69	1.78±0.10	1.57±0.06	4.98±0.43	**6.61±0.62**
	95.0%	ERK	1.02±0.04	3.27±0.94	1.82±0.65	4.09±0.09	9.04±0.62	**9.96±0.92**
		LAMP	0.99±0.01	3.05±0.80	1.61±0.18	3.50±0.22	7.23±0.24	**8.68±0.67**
		Global	0.98±0.03	2.31±0.99	1.65±0.59	1.00±0.16	1.31±0.19	**3.24±0.62**
		Uniform	1.00±0.00	2.17±0.73	1.64±0.56	1.21±0.15	1.65±0.27	**2.53±0.09**
	95.5%	ERK	1.00±0.00	3.13±0.73	1.40±0.21	3.01±0.30	7.17±1.04	**8.29±1.24**
		LAMP	1.00±0.00	3.04±0.94	1.58±0.42	2.81±0.23	5.57±0.31	**6.61±0.41**
		Global	1.02±0.03	2.20±0.88	1.84±0.37	1.02±0.10	1.32±0.22	**2.98±0.53**
		Uniform	1.00±0.00	**2.16±0.89**	1.47±0.37	1.12±0.23	1.28±0.30	1.96±0.12
Imagenette	90.0%	ERK	14.06±3.07	62.50±9.84	87.28±1.55	83.42±1.22	**88.35±0.24**	87.80±0.56
		LAMP	14.09±6.56	70.46±9.77	92.20±0.69	90.25±0.13	**92.43±0.13**	91.91±0.57
		Global	14.96±4.28	85.03±0.59	**88.76±1.50**	67.83±2.33	84.65±0.41	83.77±1.06
		Uniform	11.22±1.00	55.63±4.08	**78.58±2.59**	20.39±1.95	44.70±2.86	46.45±3.07
	92.5%	ERK	14.94±4.23	40.93±12.63	77.45±2.55	66.76±2.13	**78.89±1.06**	77.14±1.43
		LAMP	10.95±2.42	41.52±17.68	85.51±1.08	77.77±2.02	**85.78±0.36**	81.84±1.68
		Global	12.64±3.31	65.43±7.61	**80.67±2.32**	43.14±3.35	71.48±1.39	68.47±2.17
		Uniform	11.78±1.49	22.79±6.09	**57.19±5.55**	12.44±1.24	22.18±1.97	29.97±1.52
	95.0%	ERK	12.60±3.98	19.49±7.01	44.08±4.09	33.99±1.93	**57.65±2.80**	50.41±2.51
		LAMP	9.83±0.65	20.83±10.21	58.64±2.49	42.50±2.52	**60.33±2.46**	45.14±7.22
		Global	10.40±0.61	28.48±12.49	51.40±8.59	27.03±2.69	**55.48±3.08**	52.01±1.67
		Uniform	10.16±0.14	12.99±3.94	20.83±4.70	11.33±1.28	14.23±2.22	**30.81±2.59**

Table 13: ResNet-18/ImageNet-100 validation. Entries are test accuracies (%); bold marks the best repair for each allocation.

	90.0%		92.5%		95.0%	
Repair	ERK	LAMP	ERK	LAMP	ERK	LAMP
No repair	2.02	2.78	1.63	1.97	1.00	1.15
BN	48.96	49.75	31.59	30.99	13.15	11.95
ASR-q50	49.01	49.75	32.75	32.13	13.88	13.10
Clip	**49.48**	**50.17**	**33.00**	**32.71**	**14.01**	**13.20**

also varies: ASR-q50 is strongest for LAMP at both sparsities, whereas the unadjusted model is strongest for Uniform at both sparsities and for ERK at 80%.

Table 14: DenseNet-121 seed-0 validation on CIFAR-10 under the clean pipeline. Dense accuracy is 92.80%. CIFAR-10 uses lower sparsity points because the DenseNet transition boundary occurs earlier than in the main ResNet matrix. Best result in each row is **bold**.

Sparsity	Alloc.	No repair	BN	Layerwise	Channelwise	ASR-q50	Clip
70.0%	ERK	15.77	75.04	69.64	76.60	77.09	**77.18**
70.0%	LAMP	25.21	**80.01**	79.73	76.06	79.08	79.41
70.0%	Global	52.56	**87.65**	87.28	84.60	86.25	86.47
70.0%	Uniform	39.65	**87.89**	87.48	84.91	86.90	86.97
80.0%	ERK	12.77	43.34	10.79	23.41	42.87	**46.39**
80.0%	LAMP	14.80	56.89	**57.51**	45.43	52.49	53.26
80.0%	Global	18.70	**72.94**	72.63	58.49	67.30	68.15
80.0%	Uniform	14.99	73.51	**74.75**	58.43	70.45	71.39
85.0%	ERK	9.89	24.02	10.02	19.60	28.27	**30.73**
85.0%	LAMP	10.34	33.63	**38.42**	23.54	31.13	32.24
85.0%	Global	13.07	**43.98**	22.46	27.59	40.38	43.26
85.0%	Uniform	18.87	43.92	**50.04**	33.78	48.14	48.74
90.0%	ERK	10.00	11.01	9.99	12.46	12.62	**12.65**
90.0%	LAMP	10.31	15.54	12.15	16.19	21.34	**21.44**
90.0%	Global	12.86	14.95	10.24	13.92	20.07	**20.35**
90.0%	Uniform	13.75	18.89	10.01	14.86	19.68	**19.90**

Table 15: DenseNet-121 seed-0 validation on CIFAR-100 under the clean pipeline. Dense accuracy is 74.52%. Best result in each row is **bold**.

Sparsity	Alloc.	No repair	BN	Layerwise	Channelwise	ASR-q50	Clip
70.0%	ERK	2.31	47.69	45.45	47.06	**49.11**	48.71
70.0%	LAMP	5.63	**51.82**	50.69	47.60	51.05	50.79
70.0%	Global	11.51	**63.86**	63.30	60.00	62.45	62.66
70.0%	Uniform	10.41	**64.01**	63.46	59.47	63.13	62.96
80.0%	ERK	1.06	**18.65**	2.29	6.10	16.56	16.95
80.0%	LAMP	1.00	23.15	**25.06**	16.62	22.93	22.75
80.0%	Global	1.72	**44.51**	42.83	24.63	37.57	37.59
80.0%	Uniform	1.99	44.46	**45.48**	27.98	41.96	41.95
85.0%	ERK	0.93	5.40	1.24	4.32	**7.40**	6.21
85.0%	LAMP	1.30	10.56	**13.19**	6.86	10.87	10.81
85.0%	Global	1.04	**16.51**	7.47	5.34	15.58	16.06
85.0%	Uniform	1.15	22.08	**23.50**	7.92	21.11	21.02
90.0%	ERK	1.15	1.59	1.04	1.48	1.96	**2.04**
90.0%	LAMP	1.00	3.95	3.47	2.64	**4.70**	4.67
90.0%	Global	1.02	1.37	1.03	2.14	**3.31**	3.10
90.0%	Uniform	1.53	3.99	2.08	2.90	6.06	**6.10**

These results provide initial evidence that LDMS-σ remains informative for a LayerNorm-based Transformer. The pilot covers one model, one dataset, two sparsities, and three allocations; Global is not evaluated. It is therefore a three-way exploratory test, is excluded from the 27-setting ResNet aggregates, and does not establish broad Transformer transfer or isolate LayerNorm from other architectural differences.

K IMAGENET-1K SINGLE-SEED STRESS TEST

We evaluate ResNet-50 (He et al., 2016) on ImageNet-1K (Deng et al., 2009) at 92.5% sparsity using one pruning and calibration seed. This experiment is not included in the main aggregate counts because it is single-seed, but it provides a larger-scale stress test for the collapse regime (Table 19).

Table 16: DenseNet-121 seed-0 validation on Imagenette under the clean pipeline. Dense accuracy is 98.65%. Best result in each row is **bold**.

Sparsity	Alloc.	No repair	BN	Layerwise	Channelwise	ASR-q50	Clip
90.0%	ERK	10.17	12.82	10.57	15.87	**25.25**	24.41
90.0%	LAMP	9.12	**50.32**	48.20	16.36	24.59	26.29
90.0%	Global	10.11	**68.71**	67.26	19.29	39.49	37.50
90.0%	Uniform	10.17	46.14	**57.83**	15.41	22.78	22.42
92.5%	ERK	13.61	9.89	10.09	11.75	15.80	**16.46**
92.5%	LAMP	9.10	11.87	**31.85**	11.64	18.98	19.08
92.5%	Global	9.94	19.47	**35.62**	21.22	31.90	32.20
92.5%	Uniform	10.37	14.37	**26.11**	11.52	16.23	19.11
95.0%	ERK	9.12	10.17	10.17	11.06	12.41	**15.44**
95.0%	LAMP	9.35	9.76	10.22	16.15	15.08	**16.31**
95.0%	Global	9.30	10.39	11.24	13.22	20.51	**25.50**
95.0%	Uniform	10.17	10.04	9.66	13.04	13.73	**15.54**

Table 17: LDMS-σ allocation selection on held-out architecture stress tests. Regret is measured relative to the best repaired allocation in the corresponding setting.

Architecture	Scope	Exact	Mean regret	Max regret
VGG16-BN	all settings	11 / 24	0.333	2.004
VGG16-BN	main sparsities	5 / 9	0.126	0.500
DenseNet-121	held-out	7 / 11	0.811	6.990

This setting falls in the collapse regime. Even the best repaired model reaches only 7.38% top-1 accuracy, far below the 73.69% dense model. The strongest repair is layerwise scale matching followed by BN recalibration for all four allocations, while ASR-q50 remains below BN recalibration in every allocation. Thus, unlike the transition settings emphasized in the main text, this large-scale stress case leaves too little coherent activation signal for stronger channelwise repair to recover accuracy.

L Clipped ASR Sensitivity

Figure 6 reports a sensitivity analysis for the clipping bounds used by clipped ASR. The goal of this ablation is not to tune a dataset-specific hyperparameter, but to check whether the clipped variant is stable across a reasonable range of correction limits. Moderate clipping bounds preserve most of the gain over BN recalibration while remaining close to ASR-q50, whereas overly wide bounds can reintroduce unstable channelwise corrections in fragile settings.

The sensitivity results indicate that clipping acts primarily as a stability mechanism. Very narrow bounds can under-correct channels that still contain recoverable signal, while very wide bounds approach unstable raw channelwise matching. Moderate bounds preserve most of the benefit while limiting extreme correction factors. This supports the interpretation that repair aggressiveness must be matched to the reliability of the activation statistics left by the allocation.

M Runtime of Label-Free Repair

We report mean wall-clock times over the 27 main ResNet settings and three seeds. The timings depend on the implementation and hardware and are not complexity measurements. All methods are training-free and use neither labels nor gradient updates (Table 20).

Activation-statistic repair adds little cost beyond BN recalibration and evaluation. ASR-q50 and clipped ASR require about 0.1 seconds for the closed-form correction, and their total runtime differs from BN-only recalibration by at most 0.03 seconds. This supports the intended use of these methods as label-free, training-free post-pruning repair procedures.

Table 18: Exploratory ViT-Tiny results on CIFAR-10. For each allocation, the table reports the strongest evaluated repair and its resulting accuracy. Bold values identify the highest repaired accuracy at each sparsity level, and check marks identify the allocation selected by LDMS-σ.

Sparsity	Allocation	Best repair	Best acc. (%)	Selected
70%	ERK	ASR-q50	65.2	
	LAMP	ASR-q50	**66.8**	✓
	Uniform	No repair	58.1	
80%	ERK	No repair	47.5	
	LAMP	ASR-q50	**49.3**	✓
	Uniform	No repair	42.0	

Table 19: Single-seed ResNet-50/ImageNet-1K stress test at 92.5% target sparsity. Entries report top-1 accuracy (%). Dense accuracy is 73.69%. The highest repaired accuracy is shown in bold.

Allocation	BN	Layer+BN	ASR-q50	Best repair	Best
ERK	2.46	4.35	1.63	Layer+BN	4.35
Global	6.12	**7.38**	2.35	Layer+BN	**7.38**
LAMP	3.39	5.53	2.30	Layer+BN	5.53
Uniform	1.15	1.30	0.28	Layer+BN	1.30

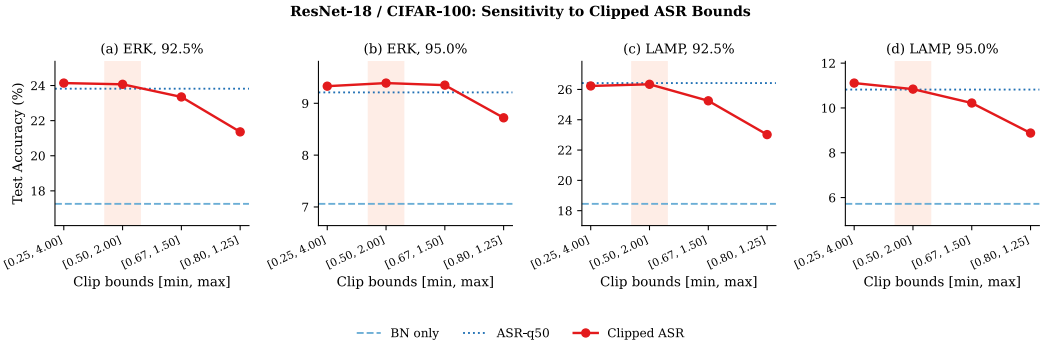


Figure 6: Sensitivity of clipped activation-statistic repair to clip bounds on ResNet-18/CIFAR-100. Each panel compares BN recalibration, ASR-q50, and clipped ASR under different clipping intervals. The results indicate that clipping mainly serves as a stability constraint for fragile transition settings rather than as a narrowly tuned hyperparameter.

Table 20: Runtime of label-free repair protocols on the 27 main ResNet settings. Times are wall-clock seconds averaged over architecture, dataset, sparsity, and seed. The repair step denotes the closed-form correction before BatchNorm recalibration.

Method	Repair	BN recal.	Eval.	Total
No repair	0.000	0.000	8.179	8.179
BN	0.000	3.902	8.107	12.009
Layerwise + BN	0.176	3.876	8.076	12.128
Channelwise + BN	0.112	3.869	8.051	12.032
ASR-q50 + BN	0.115	3.833	8.085	12.034
Clipped ASR + BN	0.118	3.822	8.079	12.019