# S$^2$D: Selective Spectral Decay for Quantization-Friendly Conditioning of Neural Activations

Arnav Chavan[1*] Nahush Lele[1*] Udbhav Bamba[1*]
Sankalp Dayal[1] Aditi Raghunathan[1,2] Deepak Gupta[1]
[1]Amazon [2]Carnegie Mellon University

## Abstract

*Activation outliers in large-scale transformer models pose a fundamental challenge to model quantization, creating excessively large ranges that cause severe accuracy drops during quantization. We empirically observe that outlier severity intensifies with pre-training scale (e.g., progressing from CLIP to the more extensively trained SigLIP and SigLIP2). Through theoretical analysis as well as empirical correlation studies, we establish the direct link between these activation outliers and dominant singular values of the weights. Building on this insight, we propose Selective Spectral Decay ($S^2D$), a geometrically-principled conditioning method that surgically regularizes only the weight components corresponding to the largest singular values during fine-tuning. Through extensive experiments, we demonstrate that $S^2D$ significantly reduces activation outliers and produces well-conditioned representations that are inherently quantization-friendly. Models trained with $S^2D$ achieve up to 7% improved PTQ accuracy on ImageNet under W4A4 quantization and 4% gains when combined with QAT. These improvements also generalize across downstream tasks and vision-language models, enabling the scaling of increasingly large and rigorously trained models without sacrificing deployment efficiency.*

## 1. Introduction

Modern transformer models exhibit an increasingly prominent phenomenon: *activation outliers*, or extremely large values in specific dimensions of neural network activations. These outliers, which can be orders of magnitude larger than typical activation values, occur more severely as models undergo more extensive pre-training [1]. Although initially observed primarily in large language models, recent evidence shows this pattern extends broadly across model families and architectures [4]. Activation outliers can severely degrade affine quantization performance, due to inefficient bit allocation.

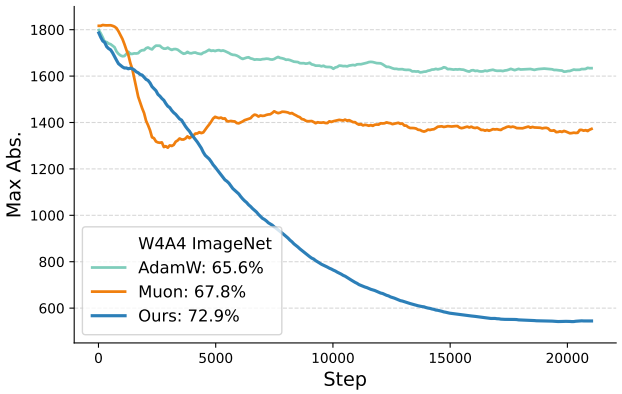



Figure 1. **Activation Outlier Suppression.**Comparison of the absolute maximum activation value (Max Abs.) of the Layer-9-FC1 output in the SigLIP-2 Base model. AdamW and Muon produce large activation outliers, whereas $S^2D$ substantially suppresses them, leading to improved downstream quantization performance as shown for W4A4.

Understanding the nature of these outliers is essential before developing mitigation strategies. Outliers severely compromise quantization by inflating activation ranges. For example, a single extreme value can force nearly all activations close to zero to be allotted in the same quantization bin. One could argue that outliers are functionally necessary features essential to the representation space, and removing them will be detrimental to model capability. However recent research on orthogonal optimizers [17] shows that outliers are an artifact of AdamW's biased optimization [3].

In this work, we empirically demonstrate that this problem escalates with the scale and duration of AdamW pre-training. Through a comparative analysis of the widely-used CLIP [20], SigLIP [34], and the more extensively trained SigLIP2 [25], we reveal a clear trend: the severity of activation outliers progressively increases with more extensive pre-training (see Figure 2). We posit that this phe-

*Equal contribution. First co-author order has been decided by a coin toss.

nomenon is a direct consequence of prolonged optimization with AdamW, whose core mechanism of adaptive, per-parameter gradient scaling is inherently anisotropic [10]. Over millions of training iterations, these anisotropic updates introduce a *privileged basis* in the model's representation space, where certain axes are preferentially amplified [6], leading to the runaway magnitudes that characterize activation outliers.

The precise geometric mechanism that leads to these outliers is still not understood well. This paper establishes the direct link: the root cause of activation outliers is the uncontrolled growth of the spectral norm of the weight matrices (see Section 3). A linear layer's capacity to amplify its input is fundamentally bounded by its spectral norm. We move beyond correlation and provide a diagnostic we term as *Principal Component Dominance Ratio (PCDR)*. This metric quantifies what fraction of an activation's absolute magnitude comes from the top-$k$ singular components of the weight matrix. Our analysis reveals that activation outliers have a substantially higher $\text{PCDR}_k$, proving that these extreme values are generated by the inflated dominant singular components in the preceding weight matrix, while normal activations have significantly lower $\text{PCDR}_k$ values.

To mitigate the occurrence of large outliers and stabilize training, orthogonal optimizers such as Muon [9] have recently been proposed. However, these approaches are designed to train models from scratch, and when applied on an AdamW pre-trained model, the benefits are not too significant (see Figure 1). We propose Selective Spectral Decay ($S^2D$), a spectral conditioning method for correcting activation outliers in AdamW pre-trained models. One of the key advantages of $S^2D$ is that it works directly on existing pre-trained models without requiring expensive retraining from scratch. Figure 1 shows that $S^2D$ is able to reduce outliers substantially compared to AdamW or Muon, improving downstream quantization performance. Using Singular Value Decomposition, $S^2D$ selectively regularizes only the largest singular values, the specific components causing outliers, while standard $L2$ weight decay uniformly shrinks all parameters. $S^2D$ can be applied during downstream fine-tuning or as a standalone post-processing step, producing well-conditioned models with maintained accuracy and improved robustness to quantization.

**Our contributions are as follows.**

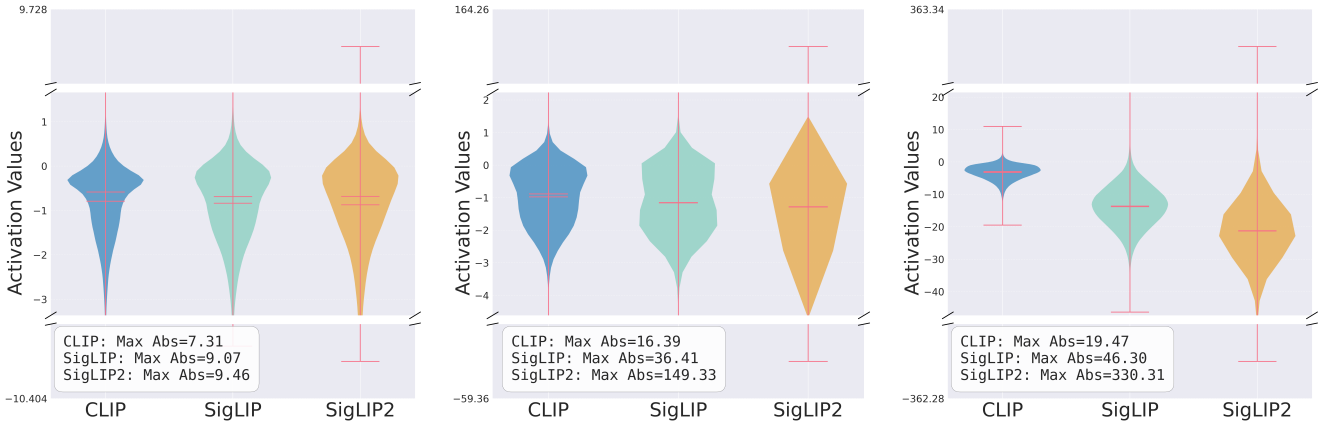
- We demonstrate that activation outlier severity escalates with pre-training scale and duration across vision-language models (*e.g.*, CLIP $\rightarrow$ SigLIP $\rightarrow$ SigLIP2), establishing outliers as an inherent artifact of prolonged optimization with traditional optimizers such as AdamW.
- We establish the direct link between inflated dominant singular values of weight matrices and activation outliers, and introduce the top-$k$ Principal Component Dominance Ratio ($\text{PCDR}_k$) as a diagnostic metric.
- We propose Selective Spectral Decay ($S^2D$), a geometrically-principled regularizer that selectively penalizes largest singular values during fine-tuning, suppressing the spectral pathologies responsible for outliers while preserving useful model capacity.
- We demonstrate through extensive experiments that $S^2D$ produces well-conditioned, quantization-ready models and push the performance of existing state-of-the-art quantization methods.

## 2. Related Works

The phenomenon of activation outliers, reflected through extreme values that appear consistently in specific feature dimensions, had emerged as a critical challenge in deploying large-scale neural networks. Dettmers et al. [5] characterized this phenomenon in LLMs, demonstrating that outlier features can exhibit magnitudes up to 150,000 times larger than typical activations. Xiao et al. [30] showed across multiple transformer architectures that outlier dimensions are highly consistent across tokens and that outlier severity increases in deeper layers. Yao et al. [32] extended this analysis to show that outliers appear across different model families and scales, with severity generally increasing with model size. Wei et al. [26] further demonstrated that outlier patterns persist across different training runs and are reproducible, suggesting they arise from fundamental properties of the training process rather than random initialization effects.

While the majority of outlier research has focused on language models, a few recent works have reported it in the context of vision transformers and multimodal models [4]. Our work contributes to this line of research by providing the first comparative analysis varying pre-training durations (CLIP, SigLIP, SigLIP2) and demonstrating that outlier severity correlates with training duration rather than model capability or task complexity. This observation provides some evidence that outliers are optimization artifacts rather than functionally necessary features.

The impact of activation outliers on quantization has been extensively studied, as they posed a fundamental challenge to model compression. Dettmers et al. [5] demonstrated that even a single outlier dimension can catastrophically affect the process of uniform quantization. Extreme outlier values force the scale factor to be very large, causing the vast majority of normal-magnitude activations to be rounded to zero or very small integers, thereby leading to *quantization collapse*. They proposed to process outlier dimensions (approximately 0.1% of features) in FP16, and the remaining in INT8. This vector-wise quantization preserves accuracy but requires dynamic routing and specialized kernels. [16] presented QLLM that extends this with more sophisticated outlier detection and handling mechanisms. An alternative line of work attempts to reduce out-




(a) Layer 1 (b) Layer 5 (c) Layer 9

Figure 2. **Activation outlier severity escalates with pre-training scale.** Activation outlier severity escalates with increasing pre-training scale. The figure plots activation distributions from the feed-forward network (FFN) layers of ViT backbones across CLIP, SIGLIP, and SIGLIP2. A clear upward trend emerges: *the magnitude of activation outliers consistently increase as we move from CLIP → SigLIP → SigLIP2*, highlighting the heavier-tailed activation behavior induced by larger and more recent model families.

lier impact through mathematically equivalent transformations [18, 21, 24]. SmoothQuant [30] mitigated the quantization difficulty by migrating outlier issue from activations to weights through per-channel scaling. Outlier Suppression [26] proposes channel-wise shifting and scaling operations that equivalently transform the network to reduce outlier magnitudes.

Concurrently, a complementary line of work has developed PTQ methods tailored to the unique architectural challenges of Vision Transformers. RepQViT [12] introduced specialized handling for post-LayerNorm (using channel-wise quantization) and post-Softmax (using $\log\sqrt{2}$) quantization, later transforming them to simple quantizers through scale reparameterization. PTQ4ViT [33] addresses the *unbalanced* post-Softmax and *asymmetric* post-GELU outputs, by proposing a *twin uniform quantization* scheme that uses separate, hardware-friendly quantizers for different value ranges. ERQ [35] introduces an innovative two-step framework to sequentially reduce both activation and weight quantization errors. By formulating their minimization as a Ridge Regression problem, ERQ addresses the intricate interdependence between weight and activation errors, thereby significantly outperforming existing ViT and LLM-centric approaches like SpinQuant [18] and OmniQuant [21] on ViT models.

Our work differs from existing approaches in several key aspects. Unlike methods that work around outliers (mixed-precision, smoothing), we address their root cause by conditioning the weight matrices to reduce spectral imbalance. Unlike QAT approaches that require retraining, $S^2D$ can be applied to existing pre-trained models. Importantly, $S^2D$ is complementary to existing quantization methods: by producing well-conditioned models with reduced outliers, $S^2D$ creates better starting points for PTQ methods, and can be naturally combined with QAT during task-specific fine-tuning to achieve even better quantization robustness.

## 3. Motivation

A well-known barrier to effective quantization is the presence of activation outliers, which force most normal activations to be squished into a narrow dynamic range, leading to sub-optimal bin allocation and ultimately degrading model accuracy. Existing works have identified this as a critical problem, but the root causes remain poorly understood. Our goal is to understand: *where do these outliers come from, and can we eliminate them without retraining from scratch?* Answering these questions requires investigating the relationship between pre-training dynamics and outlier formation. We provide two key empirical observations that motivate the development of $S^2D$. First, we observed that the problem of activation outliers is not a static issue but rather one that escalates with the scale and duration of pre-training in foundational vision models. Second, moving beyond correlation we established a direct link, showing that these outliers are generated by the dominant singular components of the weight matrices they pass through.

***Outliers scale with pre-training.*** We demonstrate here that the problem of outliers intensifies systematically with pre-training scale and duration. To investigate this relationship, we analyze three foundational vision encoder models: CLIP, SigLIP, and SigLIP2. This progressions represent substantial increases in training compute and data scale, providing controlled observations of long-term AdamW pre-training effects. Figure 2 shows the progressive emergence of activation outliers across three representative lay-

Table 1. **Principal Component Dominance Ratios for Outlier Activations.** Top-1 and Top-3 PCDR for FFN layer activations, along with the maximum singular value of the corresponding FFN weights across CLIP, SigLIP and SigLIP2. $\sigma_{max}$ denotes largest singular value.

| Model | Layer | $\text{PCDR}_1$ | $\text{PCDR}_3$ | $\sigma_{max}$ |
|---|---|---|---|---|
| CLIP | Layer 1 | 0.08 | 0.09 | 3.78 |
| | Layer 5 | 0.05 | 0.14 | 3.06 |
| | Layer 9 | 0.18 | 0.19 | 4.81 |
| SigLIP | Layer 1 | 0.01 | 0.02 | 3.40 |
| | Layer 5 | 0.03 | 0.89 | 4.86 |
| | Layer 9 | 0.14 | 0.14 | 7.59 |
| SigLIP2 | Layer 1 | 0.02 | 0.06 | 4.53 |
| | Layer 5 | 0.93 | 0.95 | 10.8 |
| | Layer 9 | 0.88 | 0.99 | 8.19 |

ers of the CLIP family. Compared to CLIP, SigLIP shows significantly increased outlier severity represented through wider tails, and this phenomenon exponentially amplifies further for SigLIP2. Note that this degradation can be seen across layers and this is evident from the distributions shown for Layers 1, 5 and 9.

***Spectral decomposition of activation outliers***. The correlation between training scale and outlier severity raises an important question: which components of the weight matrices are responsible for generating extreme activation values? We hypothesize that outliers are not distributed across all spectral components but are disproportionately generated by the dominant singular values or spectral norm of weights. To test this, we perform a spectral decomposition analysis using what we term as the *Principal Component Dominance Ratio (PCDR$_k$)*, which quantifies how much of an individual activation's magnitude originates from the largest few (top-$k$) singular components.

For the $j^{\text{th}}$ data sample, given an input activation vector $\mathbf{x}_j$ to a layer with weight matrix $\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\intercal$, the output activation for neuron $i$ can be decomposed as:

$$A_{ij} = \sum_r \sigma_r u_{ir} \mathbf{v}_r^\intercal \mathbf{x}_j \tag{1}$$

where $r$ denotes the total number of singular directions.

To account for the activation mass contributed by the top-$k$ singular values, we define $\text{PCDR}_k$ as the fraction of the activation's magnitude that comes from the top $k$ singular components.

$$\text{PCDR}_k^{(i,j)} = \frac{\sum_{r=1}^{k} \left|\sigma_r u_{ir} \mathbf{v}_r^\intercal \mathbf{x}_j\right|}{\sum_r \left|\sigma_r u_{ir} \mathbf{v}_r^\intercal \mathbf{x}_j\right|}, \tag{2}$$

where a $\text{PCDR}_k$ value of close to 1 indicates that the activation value is almost entirely determined by the top-$k$ components, while values near $1/n$ (where $n$ is the total number of components) indicate contributions are uniformly distributed.

We compute $\text{PCDR}_k$ for the largest activations value $A_{ij}$ in the FFN layer of ViT models. Results are presented in Table 1. It can be seen that $\text{PCDR}_3$ increases and approaches values closer to 1 as we scale from CLIP to SigLIP to SigLIP2. This analysis establishes that outliers are not uniformly generated by the entire weight matrix but are specifically produced by inflated dominant singular values.

## 4. Mathematical Formulation

In this section, we first establish the link between the spectral properties of a layer's weight matrix and the magnitude of its output activations. This provides a formal basis for our central hypothesis that activation outliers are a direct consequence of an inflated spectral norm. Building on this, we formulate our proposed regularizer, Selective Spectral Decay ($S^2D$) and detail its mechanism along with an efficient implementation.

***Preliminaries.*** The fundamental building block of a neural network is the linear layer, which performs the transformation $\mathbf{y} = \mathbf{W}\mathbf{x}$ for a weight matrix $\mathbf{W} \in \mathbb{R}^{m\times n}$. The geometric properties of this transformation are characterized by the Singular Value Decomposition (SVD) of $\mathbf{W}$:

$$\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\intercal = \sum_{r=1}^{N} \sigma_r \mathbf{u}_r \mathbf{v}_r^\intercal \tag{3}$$

where $N = \min(m, n)$, the columns of $\mathbf{U} \in \mathbb{R}^{m\times m}$ and $\mathbf{V} \in \mathbb{R}^{n\times n}$ are the orthonormal left and right singular vectors, respectively, and $\Sigma \in \mathbb{R}^{m\times n}$ is a rectangular diagonal matrix containing the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_N \geq 0$.

The most common form of regularization, L2 weight decay, penalizes the squared Frobenius norm of the weight matrix:

$$\mathcal{L}_2 = \frac{\lambda}{2}\|\mathbf{W}\|_F^2 = \frac{\lambda}{2}\sum_{i=1}^{N} \sigma_i^2 \tag{4}$$

This penalty applies a uniform decay pressure to all singular values, regardless of their magnitude. While effective for general-purpose regularization, it is not specifically designed to target the spectral artifacts that we have shown, are responsible for activation outliers.

***The Spectral Origin of Activation Outliers.*** We now formalize the link between the spectral norm of a weight matrix and its capacity to generate large-magnitude activations.

**Theorem 1.** *Let $\mathbf{y} = \mathbf{W}\mathbf{x}$ be the output of a linear layer for an input vector $\mathbf{x} \in \mathbb{R}^n$ and weight matrix $\mathbf{W} \in \mathbb{R}^{m\times n}$. The Euclidean norm of the output vector is bounded by the spectral norm of the weight matrix $\sigma_{\max}(\mathbf{W})$, as follows:*

$$\|\mathbf{y}\|_2 \leq \sigma_{\max}(\mathbf{W}) \cdot \|\mathbf{x}\|_2 \tag{5}$$

This establishes that a large spectral norm is a necessary condition for a layer to produce a large-magnitude output from a reasonably scaled input, providing a direct mechanism for the amplification of activation magnitude.

### 4.1. Selective Spectral Decay ($S^2D$)

Having established that the inflation of the largest singular values is the primary mechanism behind activation outliers, we introduce a regularizer that specifically targets this specific behavior.

**Definition 1.** *Given a weight matrix $\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\intercal$, we define $\mathcal{W}^{(n)} = \mathbf{U}\mathbf{\Sigma}^n\mathbf{V}^\intercal$ for a real exponent $n > 1$. The Selective Spectral Decay regularizer is then defined as*

$$\mathcal{L}_{\mathrm{S^2D}}^{(n)}(\mathbf{W}) = \frac{\lambda}{n+1}\,\mathrm{tr}\left((\mathcal{W}^{(n)})^\intercal \mathbf{W}\right) \tag{6}$$

where $\mathrm{tr}$ denotes the trace. By orthogonality of $\mathbf{V}$ and $\mathbf{U}$

$$\mathrm{tr}\big((\mathcal{W}^{(n)})^\intercal \mathbf{W}\big) = \mathrm{tr}(\mathbf{V}\mathbf{\Sigma}^{n+1}\mathbf{V}^\intercal),$$

and by cyclicity of trace,

$$\mathrm{tr}(\mathbf{V}\mathbf{\Sigma}^{n+1}\mathbf{V}^\intercal) = \mathrm{tr}(\mathbf{\Sigma}^{n+1}) = \sum_{i=1}^{N} \sigma_i^{n+1}.$$

Thus, we obtain

$$\mathcal{L}_{\mathrm{S^2D}}^{(n)}(\mathbf{W}) = \frac{\lambda}{n+1}\sum_{i=1}^{N} \sigma_i^{n+1}.$$

By choosing $n > 1$, the penalty $\sigma_i^{n+1}$ disproportionately affects larger singular values while having a little to negligible effect on smaller ones. This provides a directed mechanism for suppressing the spectral inflation compared to standard L2 decay (which corresponds to $n = 1$). This allows us to penalize those $W_{ij}$ proportionately which are not just large in value, but specifically large due to the influence of a larger singular components in the system.

Based on the above formulation, the standard partial gradients of the $L2$ regularizer can be modified from:

$$\frac{\partial}{\partial W_{ij}}\left(\frac{\lambda}{2}\|\mathbf{W}\|_F^2\right) = \lambda W_{ij} = \lambda \sum_{k=1}^{N} U_{ik}\,\sigma_k\,V_{jk}.$$

and rewritten for $S^2D$ as:

$$\frac{\partial}{\partial W_{ij}}\left(\frac{\lambda}{n+1}\,\mathrm{tr}\left((\mathcal{W}^{(n)})^\intercal \mathbf{W}\right)\right) = \lambda W_{ij}^{(n)}$$
$$= \lambda \sum_{k=1}^{N} U_{ik}\,\sigma_k^n\,V_{jk}.$$

This formulation focuses the regularization pressure on the top few $\sigma_i$ - the singular values directly responsible for the worst-case amplification of activations as shown in Theorem 1.

### 4.2. $S^2D$ in Action

$S^2D$ regularizer provides a powerful tool for penalizing dominant singular values. However, a naive implementation that computes a full SVD and applies the gradient to all singular values of all layers at every training step is computationally prohibitive and unnecessary. As our analysis in Section 3 demonstrated, activation outliers are a pathological phenomenon driven by a few dominant components in a subset of layers.

Therefore, a practical and efficient implementation of $S^2D$ must be both selective and computationally amortized. We achieve this through a $\mathrm{PCDR}_k$-based criterion and a staggered update schedule.

***$\mathrm{PCDR}_k$-based Selection.*** We directly use $\mathrm{PCDR}_k$ to identify which layers require regularization and, for those layers, which singular components to target. This selection process is governed by two hyperparameters: (1) $\tau$ – The minimum PCDR contribution that signifies a pathological concentration of mass. (2) $K_{max}$ – The maximum number of dominant singular values to consider.

For a given layer, we find the smallest rank $k_{target}$ such that $1 \le k_{target} \le K_{max}$ and $\mathrm{PCDR}_{k_{target}} \ge \tau$. If such a $k_{target}$ exists, the layer is marked for regularization, and the $S^2D$ penalty is applied only to its top $k_{target}$ singular components. If $\mathrm{PCDR}_{K_{max}} < \tau$, the layer is considered healthy, and no $S^2D$ gradient is applied.

***Amortized SVD Computation.*** The primary computational bottleneck of $S^2D$ is the SVD computation itself. Instead of re-computing the SVD at every step, we perform a full SVD on all network layers only once every $m$ iterations. This step identifies the target layers and their corresponding $k_{target}$ ranks, and caches the singular vectors ($U$, $V$) and singular values ($\Sigma$) for those layers. For the subsequent $m$ iterations, we apply the $S^2D$ gradient using these *stale* cached components. While this introduces a minor approximation (as the weight matrix $\mathbf{W}$ evolves during these steps), it amortizes the high cost of SVD over $m$ steps, making the algorithm highly efficient.

## 5. Experiments

Our experimental evaluation focuses on validating the effectiveness of $S^2D$ in producing quantization-friendly models across diverse settings. The experiments presented in this paper are designed around three important questions: (1) Does $S^2D$ effectively reduce activation outliers that hinder quantization? (2) Does $S^2D$ improve quantization performance across both PTQ and QAT regimes? (3) Do these quantization improvements generalize to downstream vision tasks?

**Hyperparameters for $S^2D$.** We use the following hyperparameters across all experiments: $\tau = 0.95$, $K_{\max} = 3$, $m = 100$, $n = 2$, and $\lambda = 5 \times 10^{-4}$. Here, $\tau$ is

Table 2. **SigLIP2 quantization performance on ImageNet1k**. Comparing ImageNet1k performance on AdamW and AdamW+$S^2D$ (Ours) across various weight (W) / activation (A) precision settings and post-training quantization (PTQ) methods. The table demonstrates that $S^2D$ shows consistent improvements for W4A4, W5A5, W6A6, and W8A8 configurations, when using ERQ, PTQ4ViT, and RepQ-ViT to perform PTQ.

| Res. | Method | FP16 | RepQ-ViT | | | | PTQ4ViT | | | | ERQ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | W4A4 | W5A5 | W6A6 | W8A8 | W4A4 | W5A5 | W6A6 | W8A8 | W4A4 | W5A5 | W6A6 | W8A8 |
| 384 | AdamW | 85.0 | 37.4 | 46.0 | 58.5 | 83.0 | 3.4 | 44.4 | 65.3 | 83.4 | 65.6 | 78.5 | 81.1 | 83.8 |
| | Ours | 85.0 | **41.4** | **78.0** | **80.0** | **83.5** | **3.9** | **62.0** | **81.5** | **84.1** | **73.0** | **81.9** | **82.7** | **84.0** |
| 512 | AdamW | 85.3 | 16.3 | 43.3 | 53.9 | 83.3 | **4.2** | 46.9 | 80.1 | 84.5 | 58.6 | 77.6 | 80.1 | 78.6 |
| | Ours | 85.4 | **17.2** | **79.2** | **82.7** | **84.3** | 3.8 | **77.0** | **83.0** | **84.9** | **68.1** | **82.3** | **83.5** | **84.1** |

the PCDR threshold indicating concentrated spectral mass, $K_{\max}$ is the maximum number of dominant singular values considered, $m$ controls the interval (in steps) between $S^2D$ updates, $n$ is the power used in $S^2D$, and $\lambda$ is the decay strength. Additional algorithmic details, sensitivity analyses, and full training settings are provided in the Supplementary Material.

### 5.1. Outlier Severity Across Model Scale

We intiated our analysis by establishing the empirical foundation: activation outliers intensify with pre-training scale, motivating the need for conditioning methods like $S^2D$. Building on the motivational analysis presented in Section 3, we quantify outlier severity across CLIP, SigLIP, and SigLIP2 vision models using metrics including maximum activation magnitude and the $\text{PCDR}_k$.

Figures 2 demonstrate a clear monotonic trend that outlier severity systematically increases from CLIP to SigLIP to SigLIP2, showing a correlation with training duration and compute. All three models use the exact same ViT-Base [28] architecture which establishes outliers as a consequence of prolonged pre-training rather than architecture-specific artifacts. This observation motivates our focus on SigLIP2 for the various experiments in the paper. As outliers are most prominent in heavily pre-trained models, the quantization challenge are greatest in this regime.

### 5.2. Post-Training Quantization (PTQ)

To evaluate $S^2D$'s capability in producing quantization-friendly models, we conduct PTQ experiments on ImageNet-1k classification. We initialize from the pre-trained SigLIP2 backbone and fine-tune using AdamW optimizer as the baseline method and AdamW+$S^2D$. Using both the approaches, the pre-trained model is fine-tuned for 10 epochs with similar hyperparamters and augmentations as outlined in [27] to produce full-precision checkpoints. These fine-tuned models serve as inputs for post-training quantization. For PTQ, we use the current state-of-the-art PTQ method, ERQ [35]; a SOTA vision transformer PTQ method, PTQ4ViT [33]; and a re-parameterization based quantization method RepQ-ViT [12] and quantize the full-precision models to different bit sizes. ERQ is competitive with all exisiting PTQ methods across vision and language transformers and hence serves as a strong PTQ baseline. Evaluation results for this experimental setup are reported in Table 2.

Based on Table 2, it is clear that $S^2D$–fine-tuned models substantially outperform standard AdamW across all PTQ methods and bit-settings. For SigLIP2-Base-384 under the ERQ W4A4 configuration, $S^2D$ achieves 72.99% versus 65.58% for AdamW, a sizable 7.41-point improvement. Even larger gains of 17.52 and 16.18 points are observed with PTQ4ViT under the W5A5 and W6A6 settings, respectively. The trend extends to RepQ-ViT, where $S^2D$ improves W5A5 accuracy from 46.04% to 78.07% and W6A6 from 58.49% to 79.98%. This consistent improvement across diverse PTQ strategies strongly suggests that the benefits of $S^2D$ arise from fundamentally better weight conditioning rather than method-specific interactions. Importantly, full-precision accuracy is essentially preserved, confirming that $S^2D$ reshapes the weight geometry specifically for quantization robustness without diminishing the model's inherent representational capacity. Overall, these results indicate that $S^2D$'s spectral regularization selectively suppresses the pathological components introduced by prolonged AdamW pre-training while preserving the useful information encoded in the weight matrices. This allows $S^2D$ to function as a pure conditioning method that produces well-conditioned models ready for deployment across both full-precision and quantized settings.

To validate the link between spectral concentration and quantization performance, we analyze the per-layer distribution of $\text{PCDR}_1$ metrics in models fine-tuned with and without $S^2D$. Table 3 demonstrates the aggregate improvements in spectral concentration for models trained using AdamW and AdamW+$S^2D$, respectively. The AdamW baseline exhibits a wide range of activation magnitudes, whereas $S^2D$ significantly reduces large-magnitude activations without negatively affecting performance. This effect is crucial for quantization, as a single poorly conditioned

Table 3. **Improved Layer Conditioning.** $\text{PCDR}_1$, maximum absolute activation, and maximum singular value for the FC1 layers of SigLIP2 after fine-tuning with AdamW and AdamW+$S^2D$ (Ours). The results show that $S^2D$ consistently reduces $\text{PCDR}_1$ compared to AdamW, indicating better spectral concentration. We also observe notable decreases in absolute max activation (Max Abs.) and leading singular values.

| Metric | Optimizer | Layer 5 | Layer 9 |
|---|---|---|---|
| $\text{PCDR}_1$ | AdamW | 0.91 | 0.77 |
| | Ours | 0.46 | 0.09 |
| Max Abs. | AdamW | 176.4 | 1166.2 |
| | Ours | 59.7 | 614.7 |
| $\sigma_{\max}$ | AdamW | 10.4 | 7.9 |
| | Ours | 3.2 | 3.9 |

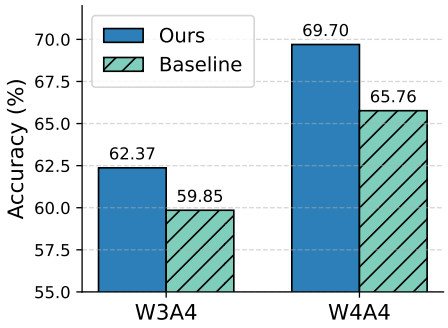



Figure 3. **QAT Performance Gains.** Quantization-Aware Training (QAT) results for W3A4 and W4A4 on ImageNet1K using AdamW and AdamW+$S^2D$ (Ours). The bar plot shows that pairing QAT with $S^2D$ improves downstream accuracy over vanilla QAT..

layer can dominate the activation range and force suboptimal scaling decisions. By explicitly regularizing large dominant singular values, $S^2D$ reduces the maximum singular value in the affected layers, directly improving their conditioning. Additional analysis on DINO [22] is provided in the supplementary material.

### 5.3. Quantization-Aware Training

We combine $S^2D$ with QAT in a challenging low-bit regime. We implement W3A4 and W4A4 quantization, a harder setup where the impact of activation outliers is particularly acute. In low-bit settings, even modest range imbalance can cause catastrophic quantization errors, making outlier mitigation critical for maintaining accuracy. We compare two QAT training regimes: a standard AdamW-based QAT baseline, and the same setup enhanced with $S^2D$ regularization. Both start from the pre-trained SigLIP2-Base-384 checkpoint and are trained for 10 epochs on ImageNet with simulated quantization in the forward pass and Straight Through Estimators (STEs) in backward propagation, using identical learning schedules and hyperparameter. We use symmetric per-channel quantization for weights and asymmetric per-tensor quantization for activations. The goal is to evaluate whether $S^2D$'s conditioning persists and provides benefits in a learnable quantization regime. $S^2D$ provides substantial benefits: an absolute improvement of 2.5% and 3.9% for W3A4 and W4A4 respectively. The results presented in Figure 3 show that $S^2D$'s conditioning can be combined with custom quantization learning schemes and is not limited to full precision fine-tuning regime.

### 5.4. Downstream Task Adaptation

***Object Detection and Instance Segmentation.*** To validate the performance of $S^2D$ on downstream vision tasks, we focus on object detection and instance segmentation, tasks that are fundamentally different from the pre-training objective and require significant model adaptation. We fine-tune the pre-trained SigLIP2-Base-384 backbone on MS-COCO [13] dataset. For both tasks, we initialize the encoder from the pre-trained checkpoint and fine-tune using the AdamW baseline and AdamW combined with $S^2D$. We leverage Detectron2 [29] library and employ Generalized RCNN network with a FPN head. Fine-tuning is performed for 270K iterations until convergence and identical learning schedules and hyperparameters across both methods. After fine-tuning, the resulting models are quantized to W4A4 using ERQ and PTQ4ViT. Table 4 presents the quantization performance on downstream tasks. Consistent with the ImageNet results, $S^2D$-fine-tuned models substantially outperform AdamW baselines across both tasks and quantization methods. On object detection, $S^2D$ achieves 29.88 points AP50 improvement with W5A5 ERQ quantization. Baseline PTQ4ViT approaches random performance at lower bits but $S^2D$ is able to retain relatively more performance across quantization bits. For semantic segmentation, the improvements are equally consistent demonstrating task-agnostic benefits. Importantly, full-precision accuracy is slightly better across both tasks, confirming that $S^2D$ functions purely as a better conditioning method without sacrificing model capacity.

***Vision-Language Models.*** We evaluate $S^2D$ on LLaVA-1.5 [15] training setting, combining a SigLIP2-Base-384 vision encoder with a Qwen2.5-0.5B language model [31]. This architecture presents a unique opportunity to study whether $S^2D$ can effectively condition heterogeneous model components with different pre-training dynamics and architectural properties. We apply $S^2D$ regularization to both components of LLaVA 1.5 during fine-tuning. Following standard practice, we first pre-train the projector using Llava-Pretrain [14] dataset to align the vision en-

Table 4. **Downstream Task Improvements.** Quantization results for object detection and instance segmentation on MS-COCO. The table reports ERQ (W4A4, W5A5, W6A6, W8A8) and PTQ4ViT (W6A6, W8A8) performance, showing that AdamW+$S^2D$ (Ours) delivers consistent performance gains across PTQ settings.

| Task | Method | FP16 | ERQ | | | | PTQ4ViT | |
|---|---|---|---|---|---|---|---|---|
| | | | W4A4 | W5A5 | W6A6 | W8A8 | W6A6 | W8A8 |
| **Object Detection** | AdamW | 50.1 | 1.8 | 10.8 | 45.8 | 48.2 | 2.2 | 42.1 |
| | Ours | 50.3 | **16.6** | **40.7** | **47.8** | **49.7** | **24.7** | **48.2** |
| **Instance Segmentation** | AdamW | 43.6 | 0.50 | 11.7 | 39.9 | 41.9 | 1.9 | 36.5 |
| | Ours | 43.8 | **13.0** | **34.4** | **41.4** | **43.2** | **20.7** | **42.8** |

Table 5. **VLM Quantization.** Evaluation of LLaVA-1.5 (SigLIP2-Base-384 + Qwen2.5-0.5B) under AdamW and AdamW+$S^2D$ (Ours) fine-tuning. $S^2D$ provides gains in both full-precision and quantized settings across GQA, TextVQA, POPE, and DocVQA, demonstrating its effectiveness as a spectral conditioning method for multimodal VLM pipelines.

| Benchmark | Method | FP16 | W4A4 | W5A5 | W6A6 |
|---|---|---|---|---|---|
| **GQA** | AdamW | 47.6 | 35.3 | 38.5 | 43.9 |
| | Ours | 55.4 | 40.1 | 47.6 | 52.8 |
| **TextVQA** | AdamW | 31.8 | 6.79 | 10.4 | 19.2 |
| | Ours | 36.7 | 9.8 | 17.6 | 28.0 |
| **POPE** | AdamW | 69.8 | 66.6 | 67.4 | 69.8 |
| | Ours | 70.3 | 66.6 | 66.6 | 69.4 |
| **DocVQA** | AdamW | 14.3 | 5.7 | 6.0 | 8.8 |
| | Ours | 17.4 | 6.1 | 7.6 | 12.4 |

coder and language decoder. Post this, fine-tuning is performed on Llava-Instruct [15] using identical hyperparameters across AdamW and Adam$S^2D$ approaches. Following fine-tuning, both full-precision and quantized versions are evaluated on standard VLM benchmarks including GQA [8], TextVQA [23], POPE [11] and DocVQA [19] covering visual question answering, fine-grained OCR and hallucination benchmarks. Table 5 demonstrates that $S^2D$ provides consistent improvements across diverse VLM evaluation benchmarks. It is interesting to note that the full-precision performance of $S^2D$ conditioned model is better than the baseline. Across the benchmarks $S^2D$ shows strong quantization performance except in the case of POPE where the differences are insignificant. The consistent improvements demonstrates that $S^2D$'s spectral conditioning benefits the entire VLM pipeline. The improvement in full-precision performance shows that $S^2D$ may serve as a better conditioning method for multi-modal models.

### 5.5. Latency analysis

There is a natural increase in training time due to the SVD computations required by $S^2D$. In our setup, a full SVD pass over all layers takes approximately 18 seconds, while the corresponding gradient pass requires about 6 seconds on 8×NVIDIA A100 GPUs. Over 10 epochs of training (25K steps), we perform roughly 250 SVD updates. However, we can effectively hide this latency by parallelizing the SVD computation and triggering it 3 iterations before it is needed. As shown in Section 4.2, we can also approximate the SVD by computing it every 100 steps without any loss in performance, making this amortized strategy both feasible and efficient. As a result, the overall overhead introduced by $S^2D$ is negligible.

## 6. Conclusions and Future Work

***Conclusions.*** This work addresses the emergence of activation outliers that arise from prolonged optimization with AdamW, establishing that these outliers are optimization artifacts rather than functionally meaningful features, and that their severity escalates with pre-training scale. To diagnose and quantify this phenomenon, we introduced PCDR$_k$, an effective spectral diagnostic, and proposed $S^2D$, a geometrically principled regularization method that selectively suppresses dominant singular components while preserving useful model capacity. Extensive experiments show that $S^2D$ yields substantial and consistent improvements across PTQ and QAT pipelines while maintaining full-precision accuracy across architectures and tasks. Together, our findings demonstrate that spectral conditioning is a powerful and general mechanism for mitigating activation outliers and enabling robust low-precision deployment.

***Future Work.*** There are several promising directions building on this work. First, exploring the interaction between $S^2D$ and alternative optimizers may reveal whether spectral conditioning remains necessary under fundamentally different optimization dynamics. Second, applying $S^2D$ during large-scale pre-training, rather than only during downstream fine-tuning, could suppress outlier formation at its source and potentially improve both stability and generalization. Finally, extending our analysis to multimodal architectures would help assess the universality of the spectral mechanisms uncovered here.

## References


[1] Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Understanding and overcoming the challenges of efficient transformer quantization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7947–7969, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 1

[2] Diego Caples and Rob Neuhaus. Adam optimizer causes privileged basis in transformer lm residual stream. LessWrong, 2024. 1

[3] Diego Caples and Rob Neuhaus. Adam optimizer causes privileged basis in transformer lm, 2024. 1

[4] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2024. 1, 2

[5] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Llm.int8(): 8-bit matrix multiplication for transformers at scale, 2022. 2

[6] Nelson Elhage, Robert Lasenby, and Christopher Olah. Privileged bases in the transformer residual stream. *Transformer Circuits Thread*, 24, 2023. 2, 1

[7] Bobby He, Lorenzo Noci, Daniele Paliotta, Imanol Schlag, and Thomas Hofmann. Understanding and minimising outlier features in neural network training (2024). *URL https://arxiv. org/abs/2405.19279*. 1

[8] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 8

[9] Keller Jordan. Muon: A new optimizer for training neural networks. `kellerjordan.github.io/posts/muon/`, 2024. 2

[10] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 2

[11] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 292–305, 2023. 8

[12] Zhikai Li, Junrui Xiao, Lianwei Yang, and Qingyi Gu. Repq-vit: Scale reparameterization for post-training quantization of vision transformers, 2023. 3, 6

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 7

[14] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 7

[15] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024. 7, 8

[16] Jing Liu, Ruihao Gong, Xiuying Wei, Zhiwei Dong, Jianfei Cai, and Bohan Zhuang. Qllm: Accurate and efficient low-bitwidth quantization for large language models, 2024. 2

[17] Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for llm training, 2025. 1

[18] Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. Spinquant: Llm quantization with learned rotations, 2025. 3

[19] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 8

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1

[21] Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. Omniquant: Omnidirectionally calibrated quantization for large language models, 2024. 3

[22] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 7, 2

[23] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 8

[24] Fuwen Tan, Royson Lee, Łukasz Dudziak, Shell Xu Hu, Sourav Bhattacharya, Timothy Hospedales, Georgios Tzimiropoulos, and Brais Martinez. Mobilequant: Mobile-friendly quantization for on-device language models, 2024. 3

[25] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. 1

[26] Xiuying Wei, Yunchen Zhang, Yuhang Li, Xiangguo Zhang, Ruihao Gong, Jinyang Guo, and Xianglong Liu. Outlier suppression+: Accurate quantization of large language models by equivalent and optimal shifting and scaling, 2023. 2, 3

[27] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR, 2022. 6

[28] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph

Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision, 2020. 6

[29] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 7

[30] Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models, 2024. 2, 3

[31] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 7

[32] Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. Zeroquant: Efficient and affordable post-training quantization for large-scale transformers, 2022. 2

[33] Zhihang Yuan, Chenhao Xue, Yiqi Chen, Qiang Wu, and Guangyu Sun. Ptq4vit: Post-training quantization for vision transformers with twin uniform quantization, 2024. 3, 6, 2

[34] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. 1

[35] Yunshan Zhong, You Huang, Jiawei Hu, Yuxin Zhang, and Rongrong Ji. Towards accurate post-training quantization of vision transformers via error reduction, 2025. 3, 6, 2

# Appendix

## A. Theoretical Analysis: Spectral Bounds

In this section, we provide the formal proof for Theorem 1 stated in the main text and expand upon the connection between spectral norms and the propagation of activation outliers in deep networks.

### A.1. Proof of Activation Magnitude Bound

We first restate the bound regarding the relationship between the spectral norm of a weight matrix and the magnitude of the output activations.

**Theorem 1** (Restated)**.** *Let $\mathbf{y} = \mathbf{W}\mathbf{x}$ be the output of a linear layer for an input vector $\mathbf{x} \in \mathbb{R}^n$ and weight matrix $\mathbf{W} \in \mathbb{R}^{m\times n}$. The Euclidean norm of the output vector is bounded by the spectral norm of the weight matrix $\sigma_{\max}(\mathbf{W})$, such that:*

$$\|\mathbf{y}\|_2 \leq \sigma_{\max}(\mathbf{W}) \cdot \|\mathbf{x}\|_2 \tag{7}$$

*Proof.* Let $\|\cdot\|_2$ denote the Euclidean norm on vectors. The matrix norm induced by the vector Euclidean norm (the spectral norm) is defined as:

$$\|\mathbf{W}\|_2 := \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{W}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \sigma_{\max}(\mathbf{W}) \tag{8}$$

where $\sigma_{\max}(\mathbf{W})$ is the largest singular value of $\mathbf{W}$. By the definition of the supremum, for any specific $\mathbf{x} \in \mathbb{R}^n$, it must hold that:

$$\frac{\|\mathbf{W}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \leq \sigma_{\max}(\mathbf{W}) \tag{9}$$

Multiplying both sides by $\|\mathbf{x}\|_2$ (assuming $\mathbf{x} \neq \mathbf{0}$; the trivial case holds for $\mathbf{x} = \mathbf{0}$) yields:

$$\|\mathbf{y}\|_2 = \|\mathbf{W}\mathbf{x}\|_2 \leq \sigma_{\max}(\mathbf{W})\|\mathbf{x}\|_2 \tag{10}$$

□

This result establishes that a large spectral norm is a necessary condition for a linear layer to amplify a reasonably scaled input into a large-magnitude output outlier.

## B. Algorithm

Algorithm 1 presents the complete training procedure for Selective Spectral Decay ($S^2D$). The algorithm operates by periodically computing singular value decompositions and selectively penalizing dominant spectral components responsible for activation outliers.

**Key steps:**

1. **Periodic SVD updates (Lines 6–17):** Every $k$ training steps, the algorithm computes the SVD of each layer's weight matrix $\mathbf{W}^{(l)} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, identifying the spectral structure of the weights.
2. **Outlier detection (Line 9):** Using the Principal Component Dominance Ratio (PCDR), the algorithm identifies the minimum rank $k$ where PCDR $\geq \tau$. This determines how many dominant singular values are responsible for creating outliers.
3. **Penalty matrix construction (Lines 11–12):** For layers with identified outliers, a penalty matrix $\mathbf{G}_{\text{reg}}$ is constructed by raising the top-$k$ singular values to power $n$ and reconstructing the partial matrix. This targets only the problematic spectral components.
4. **Gradient update (Lines 18–21):** During each training step, the standard task gradient is augmented with the cached penalty $\lambda\mathbf{G}_{\text{reg}}$, applying selective regularization pressure to the weight components aligned with the largest singular values while leaving other components largely unaffected.

## C. Analysis of Outlier Origins

In this section, we expand upon the connection between adaptive optimizers and the formation of activation outliers in transformer models. While the dominant singular *directions* $(U, V)$ encode semantically meaningful representations, their *extreme magnitudes* $(\Sigma)$ are predominantly optimization artifacts rather than functionally necessary features.

**Evidence from Prior Work.** Several independent lines of evidence support this characterization. [2] show that Adam-trained models exhibit rapid growth in excess kurtosis ($> 100$), indicating the emergence of significant outlier channels, whereas SGD-trained models maintain substantially lower kurtosis throughout training. [6] demonstrate that Adam's component-wise normalization privileges the training basis; when this basis is rotated to decorrelate the model, outliers disappear without performance loss, confirming they are not functionally necessary. Furthermore, [7] link outlier features to large diagonal adaptive learning rates in Adam, showing that reducing adaptivity minimizes outlier formation.

**Implications for $S^2D$.** These findings establish that outlier magnitudes are preventable artifacts of AdamW's basis preference and anisotropic update dynamics. $S^2D$ acts as a targeted counter-force to this spectral amplification. The fact that $S^2D$ maintains or improves full-precision accuracy (e.g., +1.2% on LLaVA, Table 5 in the main text) confirms that suppressing these extreme magnitudes is benign to the model's semantic capacity.

**Algorithm 1** Selective Spectral Decay ($S^2D$)

**Input:** Weights $\mathbf{W}$
**Hyperparams:** Power $n$, Reg. strength $\lambda$, Learning rate $\eta$, Update frequency $k$, PCDR threshold $\tau$
Initialize step counter $t \leftarrow 0$
Initialize penalty matrices $\mathbf{G}_{\text{reg}}^{(l)} \leftarrow \mathbf{0}$ for all layers $l$
**while training do**
  **if** $t \bmod k = 0$ **then** ▷ Periodic spectral update
    **for** layer $l = 1$ **to** $L$ **do**
      $\mathbf{U}, \mathbf{\Sigma}, \mathbf{V} \leftarrow \text{SVD}(\mathbf{W}^{(l)})$
      $\hat{k} \leftarrow \min\{k' : \text{PCDR}(\mathbf{\Sigma}, k') \geq \tau\}$ ▷ Identify outlier rank cutoff
      **if** $\hat{k}$ is defined **then**
        $\mathbf{\Sigma}_n \leftarrow \text{diag}(\sigma_1^n, \ldots, \sigma_{\hat{k}}^n)$
        $\mathbf{G}_{\text{reg}}^{(l)} \leftarrow \mathbf{U}_{:,1:\hat{k}}\, \mathbf{\Sigma}_n\, (\mathbf{V}_{:,1:\hat{k}})^\top$ ▷ Cache penalty matrix
      **else**
        $\mathbf{G}_{\text{reg}}^{(l)} \leftarrow \mathbf{0}$ ▷ No significant outlier rank found
      **end if**
    **end for**
  **end if**
  $\nabla_{\mathbf{W}} \mathcal{L}_{\text{task}} \leftarrow \text{Backward}(\mathcal{L}_{\text{task}}(\text{batch}))$ ▷ Standard task loss gradient
  **for** layer $l = 1$ **to** $L$ **do** ▷ Apply regularized update
    $\mathbf{W}^{(l)} \leftarrow \mathbf{W}^{(l)} - \eta \left( \nabla_{\mathbf{W}^{(l)}} \mathcal{L}_{\text{task}} + \lambda\, \mathbf{G}_{\text{reg}}^{(l)} \right)$
  **end for**
  $t \leftarrow t + 1$
**end while**

## D. Comparison with Alternative Regularization Approaches

**$S^2D$ vs. Rotation Methods (SpinQuant, QuIP).** Rotation-based methods mitigate outliers by *redistributing* activation magnitudes through a learned or analytically computed basis transformation $W' = RW$. In contrast, $S^2D$ *suppresses* the spectral cause directly by penalizing the dominant singular values in $\Sigma$. These two strategies are thus orthogonal and potentially complementary. Additionally, $S^2D$ avoids the online inference overhead of rotation methods, producing standard weights compatible with vanilla deployment kernels.

**$S^2D$ vs. Standard Spectral Regularization.** Standard spectral regularization applies uniform pressure to every singular component across the network. As shown in Table 6, applying spectral regularization without the PCDR diagnostic collapses W4A4 accuracy to 40.1% on ImageNet (SigLIP2-Base-384). PCDR acts as a surgical guide, targeting only the spectral components identified to cause pathological activation concentration. Without this selectivity, regularization indiscriminately suppresses both harmful and beneficial spectral components, degrading model capacity.

Table 6. **Impact of PCDR-guided layer selection.** Comparison of $S^2D$ with and without PCDR targeting on ImageNet classification using ERQ quantization (SigLIP2-Base-384). Without PCDR, uniform spectral regularization severely degrades low-bit performance.

| | No PCDR | $S^2D$ |
|---|---|---|
| W4A4 | 40.1 | **73.0** |
| W5A5 | 77.1 | **81.9** |
| W6A6 | 81.2 | **82.7** |

## E. Self-Supervised Backbone

We extend our experiments to DINOv3 [22], a recently trained self-supervised vision backbone. We attempted to quantize DINOv3 using the official ERQ codebase [35]; however, all ERQ configurations yielded near-random accuracy regardless of bit-width, suggesting an incompatibility with the self-supervised feature distribution. We therefore report results exclusively with PTQ4ViT [33] in Table 8. We also present spectral statistics for DINOv3 in Table 7. Both the PTQ results and spectral patterns are consistent with the SigLIP2 findings, confirming that $S^2D$ improves quantization of modern self-supervised models.

Table 7. **Comparison of FFN activation and weight statistics for SigLIP2 and DINOv3.** We report the $PCDR_1$ and maximum absolute activation of the FFN layers, along with the maximum singular value of their corresponding weights, after fine-tuning with AdamW and AdamW+$S^2$D.

| | | $PCDR_1$ | | Max Abs. Activation | | Max Singular Value | |
|---|---|---|---|---|---|---|---|
| **Model** | **Layer** | **AdamW** | **$S^2$D** | **AdamW** | **$S^2$D** | **AdamW** | **$S^2$D** |
| SigLIP2 | Layer 5 | 0.91 | 0.46 | 176.40 | 59.68 | 10.38 | 3.22 |
| | Layer 9 | 0.77 | 0.09 | 1166.2 | 614.7 | 7.87 | 3.96 |
| DINOv3 | Layer 2 | 0.47 | 0.11 | 1048.41 | 440.38 | 37.29 | 17.94 |
| | Layer 4 | 0.45 | 0.22 | 115.55 | 67.56 | 8.67 | 3.79 |

Table 8. **PTQ4ViT quantization results on DINOv3-Base** with and without $S^2$D regularization. ImageNet top-1 accuracy (%) is reported.

| **Method** | **W8A8** | **W6A6** | **W5A5** |
|---|---|---|---|
| AdamW | 58.03 | **57.04** | 28.52 |
| AdamW + $S^2$D | **76.53** | 56.10 | **30.69** |

## F. Extension to Language Models

To evaluate the generality of $S^2$D beyond vision and vision-language tasks, we conduct a preliminary experiment on a pure language model. We fine-tune Qwen2.5-0.5B using supervised fine-tuning (SFT) on the Dolci dataset, applying $S^2$D with the same hyperparameters used in the vision experiments (no task-specific tuning). We then evaluate on GSM8K (0-shot) under round-to-nearest (RTN) quantization at various bit-widths. Results are presented in Table 9.

Despite using fewer than 1B training tokens and no hyperparameter adaptation for the language domain, $S^2$D consistently improves quantized performance across W8A8, W7A7, and W6A6 settings, with gains of +2.2, +2.6, and +2.0 percentage points respectively. The slight reduction in full-precision performance ($-1.0$) reflects the regularization trade-off, which is more than compensated by the quantization gains. We expect that a learning rate sweep and longer training schedule would further improve both full-precision and quantized results.

Table 9. **GSM8K 0-shot results with RTN quantization for Qwen2.5-0.5B.** $S^2$D improves quantized accuracy across all tested bit-widths using the same hyperparameters from the vision experiments.

| | **$S^2$D** | **Baseline** |
|---|---|---|
| FP | 28.2 | 29.2 |
| W8A8 | **26.2** | 24.0 |
| W7A7 | **24.7** | 22.1 |
| W6A6 | **19.8** | 17.8 |

Table 10. **Hyperparameter sensitivity analysis for $S^2$D on ImageNet.** We vary the SVD computation frequency ($k$), number of targeted singular values (topk), and PCDR threshold. The **default** configuration is shown in bold.

| $k$ | topk | $\tau$ | **FP16 (%)** | **W4A4 (%)** |
|---|---|---|---|---|
| *Vary SVD update frequency $k$* | | | | |
| 10 | 3 | 0.95 | 85.0 | 72.2 |
| **100** | **3** | **0.95** | **85.0** | **73.0** |
| *Vary number of targeted singular values* | | | | |
| 100 | 5 | 0.95 | 85.0 | 72.5 |
| 100 | 10 | 0.95 | 84.9 | 72.6 |
| *Vary PCDR threshold $\tau$* | | | | |
| 100 | 3 | 0.90 | 84.9 | 72.4 |
| 100 | 3 | 0.80 | 84.8 | 72.8 |

## G. Hyperparameter Sensitivity

We analyze the robustness of $S^2$D by varying its key hyperparameters. As shown in Table 10, the default configuration ($k$=100, topk=3, threshold=0.95) yields the highest W4A4 performance at 73.0%.

We observe that a larger update interval ($k$=100) outperforms frequent updates ($k$=10), suggesting that accumulating statistics over a longer horizon improves stability. Interestingly, increasing topk from 3 to 10 results in a marginal performance drop, indicating that outlier mitigation is most effective when targeting only the few most dominant singular directions. Finally, a stricter PCDR threshold of 0.95 proves optimal compared to lower values.

### G.1. Sensitivity to Power Exponent $n$

The power exponent $n$ in $S^2$D controls the degree of non-uniformity in the penalty applied to singular values: larger $n$ concentrates regularization pressure more aggressively on

Table 11. **Sensitivity to power exponent $n$.** ImageNet accuracy (%) using ERQ quantization on SigLIP2-Base-384. $n$=2 provides the best balance across bit-widths.

| | Baseline | $\boldsymbol{n=2}$ | $n$=3 | $n$=4 |
|---|---|---|---|---|
| W4A4 | 65.6 | **73.0** | 68.3 | 69.6 |
| W5A5 | 78.5 | **81.9** | 79.0 | 78.7 |
| W6A6 | 81.1 | **82.7** | 81.7 | 82.4 |

the dominant singular values. We chose $n$=2 (yielding a $\sigma^3$ penalty) to exert stronger regularization on the singular values contributing to outliers while preserving smaller components. Table 11 confirms that while higher orders ($n$=3, 4) still outperform the baseline, $n$=2 offers the optimal trade-off between outlier suppression and capacity preservation.

## G.2. Amortized SVD Stability

A potential concern with the amortized SVD computation (every $m$=100 steps) is whether the cached singular vectors $(U, V)$ become stale and lead to inaccurate gradient updates. To validate this design choice, we analyzed the stability of the $S^2$D gradient penalty computed using cached versus freshly computed SVD factors. The cosine similarity between the two gradient signals remains above $0.99$ over the $m$=100 step caching interval, confirming that the singular vector subspaces evolve slowly relative to the caching frequency. This justifies the computational amortization and explains why $k$=100 outperforms more frequent updates ($k$=10) in Table 10: the additional noise from frequent recomputation slightly destabilizes training without meaningful accuracy benefit.