SNECV-Muon: Energy-Aware Adaptive Orthogonalization for Blockwise Muon in Shard MoE Pretraining

Shen Jiarun
The Chinese University of Hong Kong, Shenzhen
225045001@link.cuhk.edu.cn

Li Xiao
The Chinese University of Hong Kong, Shenzhen
lixiao@cuhk.edu.cn

Abstract

Muon-style optimizers improve optimization by orthogonalizing matrix-valued updates, but their distributed cost becomes significant under tensor parallelism [4]. Existing blockwise variants reduce communication by orthogonalizing local shards independently, and periodic correction methods recover part of the lost global geometry [5]. In sparse Mixture-of-Experts (MoE) training, however, routing creates strong cross-shard heterogeneity [12, 2], making a fixed periodic schedule suboptimal. We propose **SNECV-Muon**, a simple adaptive controller for blockwise Muon. The method monitors the coefficient of variation of shard energies and standardizes this signal with an exponential moving average per matrix. The resulting score acts as an online estimate of when the block-diagonal approximation is reliable and when a global full Muon update is needed. SNECV-Muon performs local orthogonalization by default, damps updates in mildly abnormal regimes, and triggers global full orthogonalization only when shard imbalance becomes statistically significant. We evaluate Adam, Dion, Muon, MuonBP, and SNECV-Muon on several variants of Dense and MoE language models in Megatron-LM with tensor parallelism and expert parallelism [13, 10]. Under matched communication budgets, SNECV-Muon consistently improves the throughput–quality tradeoff over MuonBP, either reaching lower loss or matching loss with higher throughput. Additional analysis shows that the trigger tracks cross-shard geometric imbalance and correlates with the failure of purely local orthogonalization.

1 Introduction

Orthogonalized updates have emerged as a practical alternative to coordinate-wise preconditioning for large language model pretraining [3]. Recent work shows that Muon can scale to large language models and improve training efficiency relative to AdamW [4, 9, 6], while several follow-up methods address its communication and normalization issues in distributed settings. In particular, MuonBP reduces the cost of global orthogonalization by alternating local blockwise steps with periodic full steps [5], Dion redesigns orthonormalized updates for communication efficiency [1], and NorMuon combines orthogonalization with neuron-wise normalization [8]. These results establish that orthogonalized optimization is viable at scale, but they also show that communication overhead remains a central bottleneck once model parallelism is introduced [10, 11].

This problem becomes sharper in Mixture-of-Experts (MoE) training [12, 7, 2]. Under expert parallelism and tensor parallelism, expert weights are sharded across devices, and the routing mechanism induces strong heterogeneity in activation frequency, token load, and gradient scale. In dense models, a local blockwise orthogonalization is often a reasonable approximation to the global update. In MoE models, that assumption breaks more easily: some shards accumulate much larger update energy than others, and the block-diagonal approximation can deviate substantially from the global Muon geometry.

This paper focuses on the following question:

> *Can we decide, online and at negligible overhead, when local blockwise orthogonalization is adequate and when a global full Muon update is necessary?*

Our answer is **SNECV-Muon**, an energy-aware adaptive controller for blockwise Muon. The method is built on a simple observation: in TP-sharded MoE training, cross-shard update energies become highly non-uniform when local blockwise orthogonalization is least trustworthy. We quantify this effect with the coefficient of variation (CV) of shard energies and normalize it with an exponential moving average and variance estimate for each matrix. The resulting standardized score is cheap to compute, robust across heterogeneous layers, and directly usable as a trigger for communication.

The contribution is not a new orthogonalization kernel. Instead, SNECV-Muon is a *dynamic communication controller* for distributed orthogonalized optimization. It keeps the zero-communication path of blockwise Muon in normal regimes, attenuates local updates when the block-diagonal approximation begins to degrade, and invokes a global full Muon step only when the anomaly score is sufficiently large. This yields a better Pareto tradeoff between optimization quality and throughput than a fixed periodic schedule.

Our main contributions are:

- We identify **cross-shard energy imbalance** as a useful online proxy for the failure of blockwise orthogonalization under TP-sharded MoE training.
- We propose **SNECV-Muon**, a self-normalized energy-based controller that adaptively switches between local blockwise and global full Muon updates.
- We present a large-scale study of Adam, Dion, Muon, MuonBP, and SNECV-Muon on 960M and 1.2B dense language models and 1.5B and 2B MoE language models in Megatron-LM with tensor parallelism and expert parallelism.
- We show that, under matched communication budgets, SNECV-Muon improves the throughput-quality tradeoff over MuonBP, with a more favorable scaling profile in sparse TP/EP training.

2 Background and Motivation

2.1 Muon and distributed orthogonalization

For a matrix-valued update G_t, Muon applies momentum and a matrix orthogonalization step to improve conditioning. In practice, the orthogonalization is approximated with Newton–Schulz iterations. On a single device, this is straightforward. Under tensor parallelism, however, a full orthogonalization requires communication to reconstruct the complete matrix across shards. This is the main source of Muon's distributed overhead.

MuonBP addresses this by applying orthogonalization locally on each shard and periodically invoking a full orthogonalization step. The method is simple and effective, but it uses a *fixed* correction schedule. In MoE settings, where routing makes shard statistics highly non-stationary, a fixed schedule is unlikely to be optimal for all layers and all training phases.

2.2 Failure of block-diagonal approximation

Let a TP-sharded update matrix be split across K ranks:

$$M_t = [M_{t,1}, M_{t,2}, \ldots, M_{t,K}]$$

for a column partition, or analogously for a row partition. Pure blockwise Muon replaces the global orthogonalization of M_t with independent orthogonalization of each shard $M_{t,k}$.

This approximation is accurate only when the shards are sufficiently similar in scale and spectral structure. In dense models, parameter partitioning is often regular enough that the approximation works reasonably well. In MoE models, routing induces a long-tailed distribution of expert activity, which leads to large differences in gradient magnitude and update energy across shards. Once this happens, the local blockwise approximation can drift away from the global Muon direction.

2.3 Rationale for Energy-Imbalance Monitoring

For each shard, define the local update energy

$$E_{t,k} = \|M_{t,k}\|_F^2.$$

We then define the cross-shard coefficient of variation:

$$\mu_{E,t} = \frac{1}{K}\sum_{k=1}^{K} E_{t,k}, \qquad \sigma_{E,t} = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(E_{t,k} - \mu_{E,t})^2},$$

$$c_t(W) = \frac{\sigma_{E,t}}{\mu_{E,t} + \epsilon}.$$

This statistic directly measures *relative dispersion of shard energy*. It does *not* directly measure stable rank. Stable rank is

$$\operatorname{srank}(A) = \frac{\|A\|_F^2}{\|A\|_2^2},$$

so it depends on both Frobenius norm and spectral norm. However, under Muon-style updates and spectral-norm-controlled orthogonalization, cross-shard Frobenius-energy imbalance becomes a useful proxy for cross-shard differences in effective update dimensionality. Empirically, we observe that the stable-rank trends of weights, gradients, and updates are closely aligned in our runs, and that large cross-shard energy imbalance correlates with larger disagreement between local blockwise and full Muon updates.

This leads to the main interpretation of our method:

> **SNECV does not estimate stable rank directly. It estimates when the local block-diagonal approximation is likely to fail.**

3 Method

3.1 Self-Normalized Energy-CV trigger

A raw energy-CV threshold is too brittle because different layers have different baseline levels of shard imbalance. Router layers, hot experts, cold experts, and dense blocks all live on different scales. We therefore normalize the energy signal with a per-matrix exponential moving average (EMA).

For each matrix W, let

$$c_t(W) = \frac{\sigma_{E,t}}{\mu_{E,t} + \epsilon}$$

be the instantaneous energy-CV. We maintain an EMA mean and variance:

$$m_t(W) = \beta m_{t-1}(W) + (1-\beta)c_t(W),$$

$$v_t(W) = \beta v_{t-1}(W) + (1-\beta)\big(c_t(W) - m_{t-1}(W)\big)\big(c_t(W) - m_t(W)\big).$$

The incremental variance form above is numerically more stable than a naive squared residual update.

We then define the standardized anomaly score:

$$z_t(W) = \frac{c_t(W) - m_t(W)}{\sqrt{v_t(W) + \epsilon}}.$$

The key point is that $z_t(W)$ is measured relative to the recent history of the same matrix. This makes the trigger robust to the heterogeneous baselines of different layer types.

3.2 Adaptive update routing

SNECV-Muon routes each matrix update according to the anomaly score.

Normal regime: $z_t < z_{\text{low}}$. The shard imbalance is within the expected range. We apply local blockwise orthogonalization independently on each shard:

$$W_{t+1,k} = W_{t,k} - lr_{\text{block}} \cdot \text{Orth}\big(\nabla W_{t,k}\big).$$

Mild regime: $z_{\text{low}} \le z_t < z_{\text{high}}$. The block-diagonal approximation is degrading, but not yet catastrophically. We still use local blockwise orthogonalization, but we damp the step size:

$$lr_{\text{damp}} = \frac{lr_{\text{block}}}{1 + \gamma(z_t - z_{\text{low}})},$$

$$W_{t+1,k} = W_{t,k} - lr_{\text{damp}} \cdot \text{Orth}\big(\nabla W_{t,k}\big).$$

The purpose of this damping is not to "fix" the geometry, but to reduce the damage caused by taking a large step in a direction known to be less reliable.

Abnormal regime: $z_t \ge z_{\text{high}}$. The local approximation is deemed invalid. We gather the full update, apply global orthogonalization, and scatter the result back to shards:

$$G_t = \text{AllGather}\left(\nabla W_{t,1}, \ldots, \nabla W_{t,K}\right),$$

$$\tilde{G}_t = \text{Orth}(G_t),$$

$$W_{t+1,k} = W_{t,k} - lr_{\text{full}} \cdot \tilde{G}_{t,k}.$$

Algorithm 1 SNECV-Muon

Require: TP degree K, learning rates $lr_{\text{block}}, lr_{\text{full}}$, EMA decay β, thresholds $z_{\text{low}}, z_{\text{high}}$, damping coefficient γ, stability constant ϵ

for each training step t and each TP-sharded matrix W **do**
  for each rank $k \in \{1, \dots, K\}$ **do**
    Compute local energy $E_{t,k} \leftarrow \|M_{t,k}\|_F^2$
  end for
  Compute $c_t(W) = \sigma(E_{t,1}, \dots, E_{t,K}) / (\mu(E_{t,1}, \dots, E_{t,K}) + \epsilon)$
  Update EMA statistics:

$$m_t \leftarrow \beta m_{t-1} + (1-\beta) c_t, \qquad v_t \leftarrow \beta v_{t-1} + (1-\beta)(c_t - m_{t-1})(c_t - m_t)$$

  Compute anomaly score:

$$z_t \leftarrow \frac{c_t - m_t}{\sqrt{v_t + \epsilon}}$$

  if $z_t < z_{\text{low}}$ **then**
    Apply local blockwise orthogonalization with lr_{block}
  else if $z_{\text{low}} \le z_t < z_{\text{high}}$ **then**
    Set $lr_{\text{damp}} \leftarrow lr_{\text{block}} / (1 + \gamma(z_t - z_{\text{low}}))$
    Apply local blockwise orthogonalization with lr_{damp}
  else
    Gather full update across shards
    Apply global orthogonalization
    Scatter and apply update with lr_{full}
  end if
end for

3.3 Communication cost

The trigger is cheap. Each rank computes a local scalar $E_{t,k} = \|M_{t,k}\|_F^2$ and participates in a scalar reduction to obtain the mean and variance across shards. This is negligible compared with reconstructing the full matrix for global orthogonalization. In normal and mild regimes, SNECV-Muon adds essentially zero communication beyond local blockwise Muon. The expensive path is only activated when the energy statistic indicates that local orthogonalization is likely to be inaccurate.

3.4 Algorithm

Algorithm 1 summarizes the full training-time routing procedure, including energy-statistic computation, self-normalized triggering, and the switch between local, damped-local, and global full updates.

4 Model Configurations and Training Infrastructure

4.1 Model scales and distributed setup

Our evaluation spans four pretraining configurations: 960M and 1.2B dense language models, together with 1.5B and 2B MoE language models. All runs are implemented in Megatron-LM

with tensor parallelism and expert parallelism [13, 10]. To ensure comparability across optimizers, we keep the tokenizer, data pipeline, and learning-rate schedule family fixed, and we report each model under its primary TP/EP layout together with additional parallel configurations for scaling analysis.

4.2 Baseline optimizers

We compare SNECV-Muon against representative coordinate-wise and orthogonalized baselines: Adam as a standard adaptive optimizer, Dion as a communication-oriented orthonormalized optimizer, full Muon with global orthogonalization at every step, and MuonBP with periodic global correction on top of blockwise local orthogonalization.

4.3 Tuning protocol

To avoid an unfair comparison, each optimizer is tuned separately. For Adam and Dion, we sweep the base learning rate. For Muon, we sweep the global learning rate. For MuonBP and SNECV-Muon, we tune both lr_{block} and lr_{full}. For SNECV-Muon, we tune $(\beta, z_{\text{low}}, z_{\text{high}})$ after fixing the best learning-rate pair. The pressure-based controller discussed in earlier development is included only as an appendix ablation because it adds several hyperparameters without a consistent gain over the simpler trigger.

4.4 Evaluation metrics

Performance is evaluated with training and validation loss, throughput (tokens/s or samples/s), wall-clock time to a fixed loss target, and the realized fraction of global full updates. To characterize mechanism-level behavior, we additionally report trigger statistics by layer type and stable-rank/energy-distribution diagnostics at selected checkpoints.

5 Main Results

5.1 Optimization quality and throughput

Table 1 summarizes the main results. Across all four model scales, SNECV-Muon improves the throughput–quality tradeoff relative to MuonBP. Under matched communication budgets, it either achieves lower validation loss or reaches the same loss with lower wall-clock time. Relative to full Muon, it retains most of the optimization benefit while substantially reducing orthogonalization overhead.

5.2 Scaling behavior

Figure 1 compares the scaling trends across model sizes and parallel layouts. The gain of SNECV-Muon becomes more pronounced as model scale and communication cost grow. This is consistent with the design of the method: the larger the distributed orthogonalization overhead, the more valuable it is to reserve global full steps for moments when local blockwise orthogonalization is most likely to fail.

Table 1: Main results across dense and MoE pretraining scales. Throughput is reported in tokens/s and time-to-loss in wall-clock hours.

Scale	Optimizer	Val Loss ↓	Throughput ↑	Time-to-Loss ↓	Full Ratio	Notes
960M Dense	Adam	2.94	332k	10.6h	0.00	coordinate-wise adaptive
960M Dense	Dion	-	-	-	-	orthonormalized baseline
960M Dense	Muon	2.71	214k	8.1h	1.00	full every step
960M Dense	MuonBP	-	-	-	0.25	periodic full
960M Dense	SNECV-Muon	2.73	306k	6.8h	0.20	adaptive full
1.2B Dense	Adam	2.88	278k	10.9h	0.00	coordinate-wise adaptive
1.2B Dense	Dion	-	-	-	-	orthonormalized baseline
1.2B Dense	Muon	2.66	178k	8.7h	1.00	full every step
1.2B Dense	MuonBP	-	-	-	0.25	periodic full
1.2B Dense	SNECV-Muon	2.68	258k	7.3h	0.21	adaptive full
1.5B MoE	Adam	2.63	205k	11.8h	0.00	coordinate-wise adaptive
1.5B MoE	Dion	-	-	-	-	orthonormalized baseline
1.5B MoE	Muon	2.43	121k	9.2h	1.00	full every step
1.5B MoE	MuonBP	-	-	-	0.25	periodic full
1.5B MoE	SNECV-Muon	2.45	182k	7.7h	0.23	adaptive full
2B MoE	Adam	2.58	168k	12.6h	0.00	coordinate-wise adaptive
2B MoE	Dion	-	-	-	-	orthonormalized baseline
2B MoE	Muon	2.39	98k	9.9h	1.00	full every step
2B MoE	MuonBP	-	-	-	0.25	periodic full
2B MoE	SNECV-Muon	2.41	152k	8.3h	0.25	adaptive full

6 Why the trigger matters

6.1 Trigger statistics by layer type

The trigger is not uniformly active. Router layers and hot experts show much stronger and more volatile anomaly scores than ordinary dense layers. Figure 2 illustrates this behavior. This supports the central claim of the paper: the communication controller is reacting to real structural heterogeneity in MoE training rather than random noise.

6.2 Local versus full geometry

We compare local blockwise updates and full Muon updates on matched checkpoints and mini-batches. When the anomaly score is small, the two updates are close. When the anomaly score is large, the cosine similarity between local and full updates drops, and the stable-rank dispersion across shards increases. This supports the interpretation of SNECV as an error-aware detector for blockwise approximation failure.

7 Stable-rank analysis

Stable rank is not the trigger itself, but it is useful for understanding the effect of cross-shard imbalance. For a matrix A,

$$\operatorname{srank}(A) = \frac{\|A\|_F^2}{\|A\|_2^2}.$$

A larger stable rank indicates that the singular-value mass is less concentrated in a small number of directions. In our experiments, the stable-rank trends of weights, gradients, and updates are qualitatively similar. More importantly, we observe that when cross-shard energy imbalance is high, stable-rank dispersion across shards also tends to increase. This is consistent with the view that local blockwise updates become less faithful when different shards operate at very different effective scales.

We stress that this is an empirical relationship, not an identity. ECV measures Frobenius-energy dispersion directly; stable rank also depends on the spectral norm. We therefore use stable rank only as a diagnostic tool, not as the formal trigger.

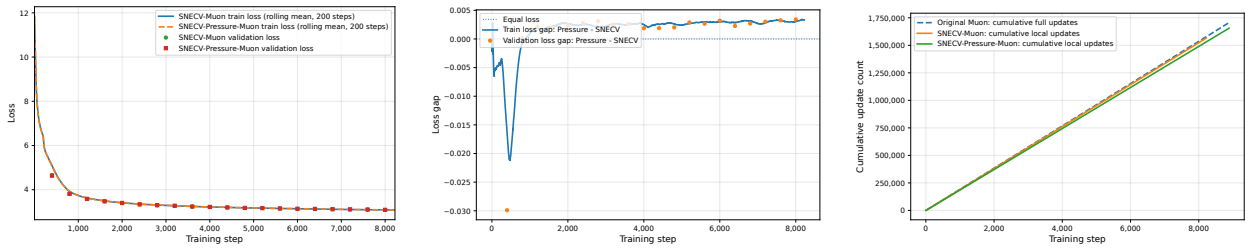
8 Ablations

8.1 Raw ECV versus self-normalized ECV

A raw energy-CV threshold is unstable across heterogeneous layers. Self-normalization with EMA statistics substantially improves transferability across router, expert, and dense blocks.

8.2 Pressure-based controller

We also evaluated a pressure-based extension that accumulates medium-regime anomalies before triggering a full update. To analyze this mechanism directly, we vary the realized number of global full orthogonalization steps while keeping the remaining training setup fixed. Figure 4 reports the resulting loss trajectories, loss-gap dynamics, and cumulative update counts. Across these diagnostics, increasing full-update frequency does not provide a meaningful improvement in training performance; instead, it primarily increases communication and synchronization cost. This explains why the pressure-based variant does not consistently outperform the simpler z-score trigger despite its additional control parameters.



(a) Validation loss trajectories. (b) Loss-gap trends over training. (c) Cumulative full-update counts.

Figure 4: Ablation of Muon full-update frequency. Across all diagnostics, more frequent full Muon updates do not improve optimization outcomes, while incurring higher distributed overhead.

9 Related Work

Muon introduced orthogonalized momentum for matrix-valued parameters and demonstrated strong empirical gains in language model training [4]. Later work established its scalability to large LLM pretraining [10]. MuonBP addressed the communication overhead of full distributed orthogonalization by alternating local blockwise and periodic full steps [5]. Dion redesigned orthonormalized updates for large distributed training [1]. NorMuon showed that orthogonalization and neuron-wise adaptive normalization can be complementary [8]. Our method is closest in spirit to MuonBP, but

it replaces a fixed periodic correction schedule with a cheap online error-aware controller tailored to TP-sharded MoE training.

10 Conclusion

We presented SNECV-Muon, an adaptive communication controller for blockwise Muon in tensor-parallel MoE pretraining. The method uses a self-normalized energy-CV trigger to decide when local blockwise orthogonalization is sufficient and when a global full Muon step is necessary. The signal is cheap to compute, robust across heterogeneous layer types, and directly tied to the failure of the block-diagonal approximation in distributed orthogonalization. Across model scales and parallel layouts, SNECV-Muon improves the throughput–quality tradeoff relative to MuonBP and retains much of the optimization benefit of full Muon at a lower communication cost. More broadly, the results suggest that in sparse distributed training, communication should be allocated by online estimates of approximation error rather than fixed schedules.

References

[1] Kwangjun Ahn, Byron Xu, Natalie Abreu, and John Langford. Dion: Distributed orthonormalized updates. Microsoft Research publication and arXiv preprint, 2025.

[2] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

[3] Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning (ICML)*, 2018.

[4] Keller Jordan. Muon: An optimizer for hidden layers in neural networks. Technical blog post and open-source implementation, 2024.

[5] Ahmed Khaled, Kaan Ozkara, Tao Yu, Mingyi Hong, and Youngsuk Park. Muonbp: Faster muon via block-periodic orthogonalization. *arXiv preprint arXiv:2510.16981*, 2025.

[6] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015.

[7] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Chen, Guillaume Wenzek, Adam Roberts, et al. GShard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations (ICLR)*, 2021.

[8] Zichong Li, Liming Liu, Chen Liang, Weizhu Chen, and Tuo Zhao. Normuon: Making muon more efficient and scalable. *arXiv preprint arXiv:2510.05491*, 2025.

[9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR)*, 2019.

[10] Deepak Narayanan, Mohammad Shoeybi, Jared Casper, Mostofa Patwary, Patrick LeGresley, Bryan Catanzaro, and Erich Elsen. Efficient large-scale language model training on gpu clusters using megatron-lm. *SC21: International Conference for High Performance Computing, Networking, Storage and Analysis*, 2021.

[11] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: Memory optimizations toward training trillion parameter models. *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, 2020.

[12] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

[13] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

A Appendix: Additional implementation details

A.1 Distributed overhead of the trigger

The trigger computes one local scalar per shard and requires only a scalar collective for each matrix. This is negligible compared with reconstructing the full matrix for a global orthogonalization step.

A.2 Appendix: Pressure-based SNECV controller

This appendix reports the pressure-based variant explored during development:

$$C_t = \gamma C_{t-1} + \mathbf{1}[z_t \in \text{medium}] \left(1 + \alpha \cdot \frac{z_t - z_{\text{low}}}{z_{\text{high}} - z_{\text{low}} + \epsilon}\right).$$

We found this controller to be less attractive than the simpler z-score trigger once tuning cost is taken into account.

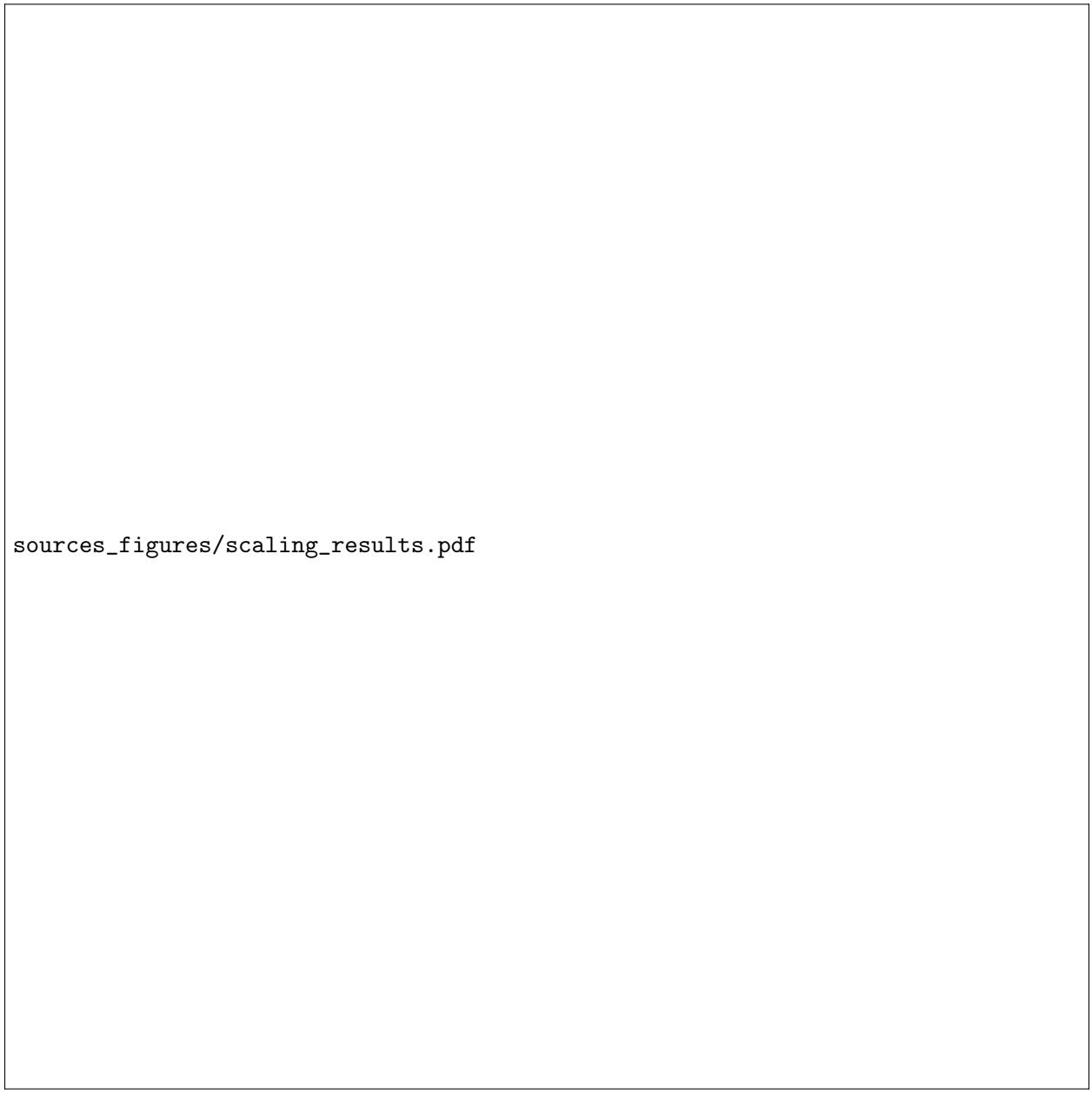


Figure 1: Scaling trends across 960M and 1.2B dense models and 1.5B and 2B MoE models. Left: validation loss. Right: wall-clock time to fixed loss. SNECV-Muon preserves most of Muon's quality while improving time-to-loss over MuonBP at matched full-update budgets.

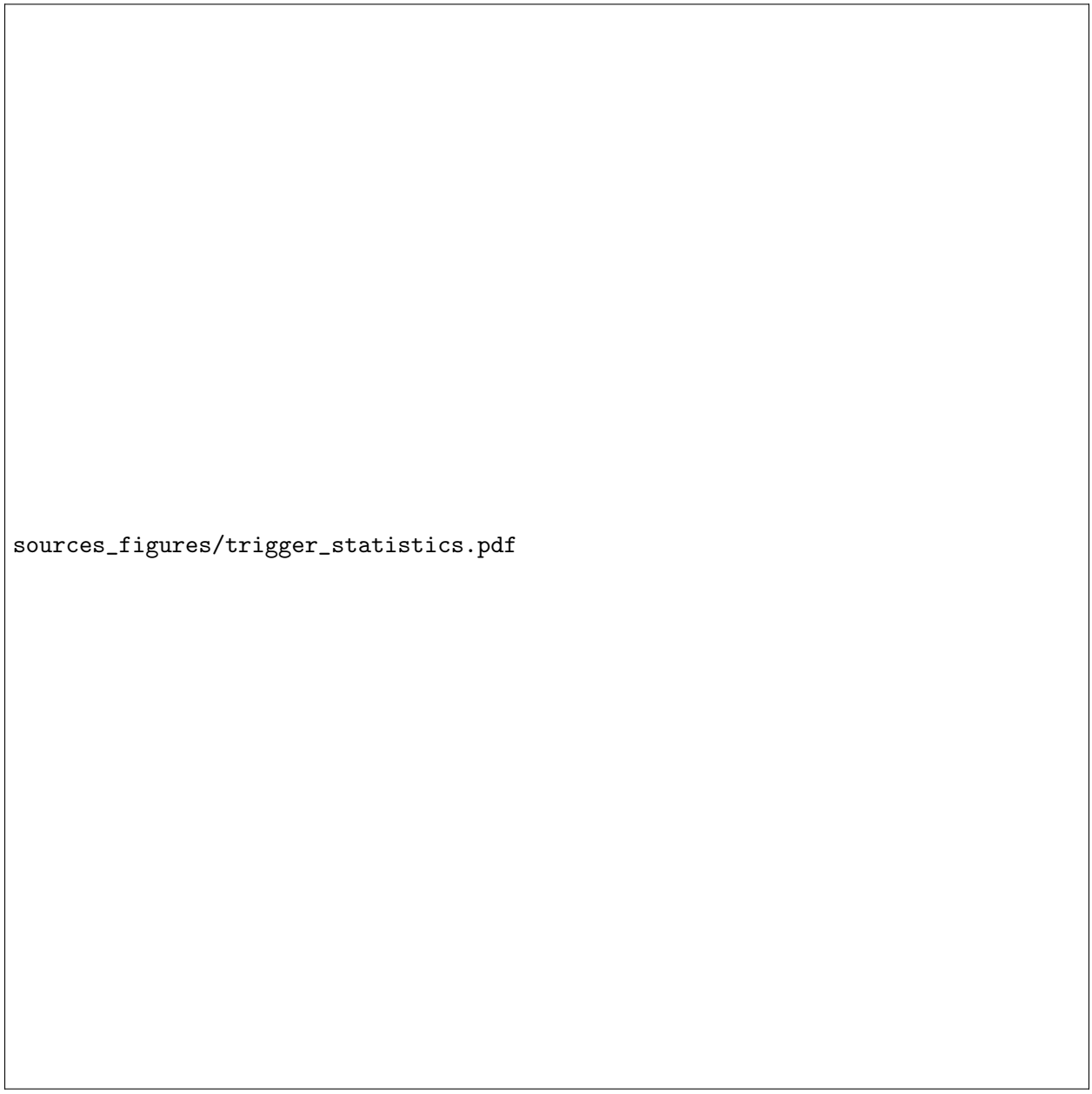


Figure 2: Trigger statistics over training. Router layers exhibit the highest anomaly magnitude and volatility, hot experts remain intermediate, and dense layers stay mostly in the low-anomaly regime.

Figure 3: Geometry diagnostics at saved checkpoints. Left: local-vs-full update cosine similarity versus trigger score. Middle: shard-energy CV over training. Right: shard stable-rank dispersion. Higher anomaly levels coincide with lower local-global alignment and increased cross-shard dispersion.