When and Why Grouping Attention Heads Accelerates Muon Optimization

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2025

Abstract

Muon orthogonalizes matrix updates, but multi-head attention naturally operates at the level of heads. This granularity mismatch raises the question of whether Muon should be applied to the full attention projection, to individual heads, or to intermediate head groups. We study this question through a one-step descent comparison between full-matrix Muon and group-wise Muon. Our analysis reveals a trade-off between the **group-wise whitening gain** from group-wise updates and the **grouping-induced norm cost**, an additional update-norm cost caused by replacing full-matrix whitening with group-wise whitening. Motivated by this trade-off, we propose **Group Muon**, which treats head group size and grouping rule as optimizer hyperparameters. On GPT-2 Small trained on FineWeb, appropriate grouping improves validation loss over both full-QKV Muon and fully head-wise MuonSplit.

1. Introduction

Muon [7] orthogonalizes the momentum update of each 2-D parameter matrix, so its natural unit is a matrix. Multi-head attention, however, later splits the query, key, and value projections into heads that compute attention independently. Applying Muon to the full projection couples all heads through a single polar factor, whereas splitting the projection gives each head, or group of heads, an independent orthogonalized update. We ask:

Q: What is the right granularity for applying Muon to multi-head attention?

This question is already practically relevant. GLM-5 [6] uses Muon Split, applying orthogonalization to per-head blocks of attention projections for MLA-based attention. A record on the modded-nanoGPT speedrun [12] also recently improved its record by orthogonalizing Q and K in pairs of heads rather than across the full multi-head matrix. These examples suggest that Muon's matrix scope is an optimizer design choice, not a mere implementation detail. Appendix A provides a fuller related-work discussion.

Yet finer splitting is not always better. As shown in Figure 1, head-wise MuonSplit reduces validation loss faster early in training, but is later overtaken by full-matrix Muon. This suggests a stage-dependent trade-off: splitting can improve descent at one stage while degrading optimization at another.

To answer this granularity question, we make the following contributions:

- We provide a one-step descent analysis comparing full-matrix Muon and group-wise Muon. The resulting criterion identifies a trade-off between the **group-wise whiten-**

© .

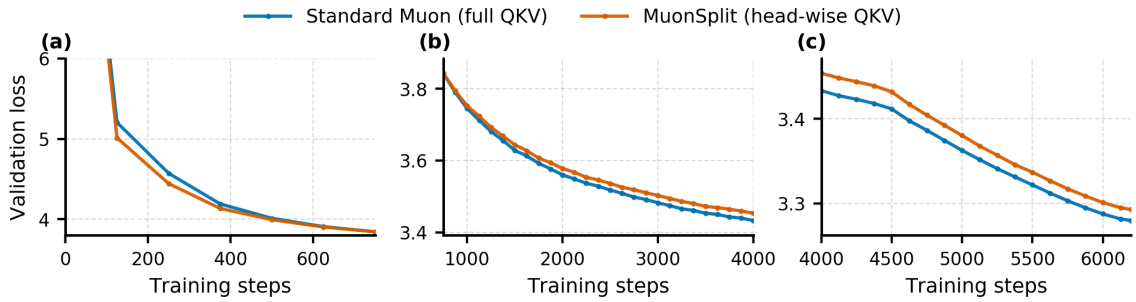

Figure 1: Validation loss of full-matrix Muon and head-wise MuonSplit in the GPT-2 Small speedrun setting. MuonSplit descends faster early, whereas full-matrix Muon performs better later.

ing gain, which increases the first-order descent term from $\|G\|_*$ to $\sum_i \|G_i\|_*$, and the **grouping-induced norm cost**, a second-order cost induced by the larger norm of group-wise whitened updates. This criterion explains the early/late reversal in Figure 1: finer splitting can help when the group-wise whitening gain outweighs the grouping-induced norm cost, but may underperform when this gain is insufficient to compensate for the norm cost introduced by grouping.

- We propose **Group Muon**, a practical head-grouped Muon variant that treats head group size and grouping rule as explicit optimizer hyperparameters. On GPT-2 Small trained on FineWeb, appropriate intermediate grouping improves validation loss over both Full-QKV Muon and fully head-wise MuonSplit.

2. Preliminaries

Block structure. Partition $W \in \mathbb{R}^{m\times d}$ into row groups $W = [W_1^\top \cdots W_M^\top]^\top$, where $W_i \in \mathbb{R}^{m_i\times d}$ and $\sum_{i=1}^M m_i = m$. For objective $\mathcal{L}(W)$, set $G := \nabla_W \mathcal{L}(W) \in \mathbb{R}^{m\times d}$ and $G = [G_1^\top \cdots G_M^\top]^\top$, where $G_i := \nabla_{W_i}\mathcal{L}(W) \in \mathbb{R}^{m_i\times d}$. Set $r := \mathrm{rank}(G)$ and $r_i := \mathrm{rank}(G_i)$.

Polar factor. Idealized Muon uses the gradient polar factor. For nonzero $X = U_X\Sigma_X V_X^\top$ with compact SVD, set $P(X) := U_X V_X^\top$ and $P(0) = 0$. Then $\langle X, P(X)\rangle = \|X\|_*, \|P(X)\|_F^2 = \mathrm{rank}(X)$, where $\|\cdot\|_*$ denotes the nuclear norm.

Update rules. *Full Matrix Muon* polar-whitens G globally:

$$O_{\text{all}} = P(G), \qquad W_{\text{all}}^+ = W - \eta O_{\text{all}}. \tag{1}$$

Group Muon polar-whitens each G_i separately:

$$O_{\text{grp}} = \begin{bmatrix} P(G_1)^\top & P(G_2)^\top & \cdots & P(G_M)^\top \end{bmatrix}^\top, \qquad W_{\text{grp}}^+ = W - \eta O_{\text{grp}}. \tag{2}$$

Experiment Setup **GPT-2 on FineWeb**: We train GPT-2 Small (124M) model on dataset using the highly optimized `nano-gpt` benchmark training recipe [11] [1]. Details are provided in Appendix C.

1. https://github.com/KellerJordan/modded-nanogpt

3. When and Why Grouping Helps: Gain versus Norm Cost

We first compare full matrix Muon and Group Muon through a one-step smoothness bound. Assume that $\mathcal{L}$ is β-smooth with respect to W, i.e.,

$$\mathcal{L}(W + \Delta) \leq \mathcal{L}(W) + \langle \nabla \mathcal{L}(W), \Delta \rangle + \frac{\beta}{2} \|\Delta\|_F^2. \tag{3}$$

Proposition 1 (One-step descent comparison) *Let $G = \nabla_W \mathcal{L}(W)$ and let G_i be the i-th group block of G. Denote $r := \operatorname{rank}(G), r_i := \operatorname{rank}(G_i)$. For the full matrix Muon and Group Muon the corresponding guaranteed one-step decrease lower bounds are*

$$D_{\text{all}} = \eta \|G\|_* - \frac{\beta \eta^2}{2} r, \qquad D_{\text{grp}} = \eta \sum_{i=1}^{M} \|G_i\|_* - \frac{\beta \eta^2}{2} \sum_{i=1}^{M} r_i.$$

Consequently, Group Muon has a larger guaranteed descent lower bound than full matrix Muon whenever

$$\boxed{\sum_{i=1}^{M} \|G_i\|_* - \|G\|_* > \frac{\beta \eta}{2} \left(\sum_{i=1}^{M} r_i - r \right).} \tag{4}$$

A detailed proof is provided in Appendix B.

Gain–cost trade-off. In (4), the left-hand side is the **group-wise whitening gain**, i.e., the first-order benefit of group-wise whitening. The right-hand side is the **grouping-induced norm cost**, which measures the additional update-norm cost caused by replacing full-matrix whitening with group-wise whitening.

Takeaway. This criterion gives a simple answer to when grouping helps: Group Muon is favored when the group-wise whitening gain outweighs the grouping-induced norm cost. Thus, the relative size of these two terms determines whether grouping improves over full matrix Muon. We first consider the small-cost regime.

Small-cost regime: near-full-rank gradients.

Observation 3.1 (Near-full-rank gradients) *Empirically, during the main training regime, the full gradient and the group gradients are close to full row rank. For a row-wise partition with $m \leq d$, this corresponds to $r \approx m, r_i \approx m_i, \sum_{i=1}^{M} r_i \approx r$.*

We provide supporting empirical evidence for this observation in Appendix D.1.

Zero-cost case: full row rank. If $\operatorname{rank}(G) = m$ and $\operatorname{rank}(G_i) = m_i$ for all groups, then $\sum_i r_i - r = 0$, so (4) reduces to $\sum_i \|G_i\|_* > \|G\|_*$. By nuclear-norm subadditivity, $\|G\|_* \leq \sum_i \|G_i\|_*$, with equality for the row-wise partition iff the block row spaces are mutually orthogonal. Thus, ideal Muon zero-cost grouping is generically favored.

Practical cost: effective rank and noise. For ideal Muon with the exact polar factor, the grouping-induced norm cost reduces to the rank-dependent term because $\|P(G)\|_F^2 = \operatorname{rank}(G)$. However, practical Muon computes an approximate whitening direction, e.g., through a finite number of Newton–Schulz iterations, and the gradients contain stochastic noise. Appendix D.2 derives this non-ideal form and empirically shows that the Frobenius-norm gap, so the grouping-induced norm cost should not be assumed to vanish in practical Muon implementations. In this case, the exact algebraic rank is not the most direct proxy for the second-order term in the descent bound.

Over-splitting risk: aligned low-rank gradients. In an aligned low-rank case with k nonzero groups, the grouping-induced norm cost is $\frac{\beta\eta}{2}(k-1)$, which increases monotonically with k. This indicates that the grouping-induced norm cost should not be assumed to vanish in practical Muon implementations. We provide more details in Appendix E.

Thus, Grouping can increase the group-wise whitening gain, but overly fine splitting can also increase the grouping-induced norm cost through larger whitened update norms. Therefore, the key design question is not simply whether to split the QKV, but how coarsely the attention heads should be grouped.

4. Group Muon Algorithm

The analysis above suggests that Group Muon's effectiveness depends on grouping structure: overly fine splitting may add grouping-induced norm cost, while intermediate groups can preserve group-wise whitening gain. We therefore introduce a practical Group Muon algorithm with tunable head group size and evaluate grouping rules.

We form groups over attention heads using a head group size g and a grouping rule $\mathcal{S}$. Here, g denotes the number of heads per group, and $\mathcal{S}$ specifies how heads are assigned. We consider adjacent (adj), interval (int), and random (rand) grouping, corresponding to consecutive heads, maximally separated heads, and randomly resampled size-g partitions, respectively. More details are provided in Appendix C.

Algorithm 1: GroupMuon

Data: parameter matrix W, gradient G, momentum buffer M, learning rate η, momentum coefficient μ, number of heads H, head group size g, grouping rule $\mathcal{S}$
Result: updated parameter matrix W
split G and M into head-wise blocks $\{G_h\}_{h=1}^{H}$ and $\{M_h\}_{h=1}^{H}$;
construct head groups $\mathcal{P} = \{P_1, \ldots, P_{H/g}\}$ according to $\mathcal{S}$;
foreach *group* $P \in \mathcal{P}$ **do**
  $\widetilde{G}_P \leftarrow \mathrm{Concat}(\{G_h : h \in P\})$;
  $\widetilde{M}_P \leftarrow \mathrm{Concat}(\{M_h : h \in P\})$;
  $\widetilde{M}_P \leftarrow \mu\widetilde{M}_P + \widetilde{G}_P$;
  $\widetilde{O}_P \leftarrow \mathrm{NewtonSchulz}(\widetilde{M}_P)$;
  split $\widetilde{O}_P$ back into head-wise updates $\{O_h : h \in P\}$;
end
merge all head-wise updates into O;
$W \leftarrow W - \eta O$;
return W;

Based on these definitions, we propose the Group Muon update described in Algorithm 1.

We apply Group Muon to attention projection matrices in three settings: QK groups the query and key projections, V groups the value projection, and Fixed-QK+V fixes QK to its best observed grouping while varying V.

As a representative case, we analyze the QK random setting to illustrate how the head group size affects the balance between whitening gain and norm cost. In the early stage(Fig 2(a)), the finest grouping $g = 1$ achieves the lowest validation loss, indicating that fine-grained head-wise whitening can be beneficial at the beginning of training. This is consistent with Observation 3.1: when the full gradient and group gradients are close to full row rank, the idealized grouping-induced norm cost in (4) is small, so the group-wise whitening gain can dominate the comparison.

However, this fine-splitting advantage does not persist throughout training. As training proceeds(Fig 2(b)), the coarser QK random grouping with $g = 6$ surpasses $g = 1$ and gradually becomes the best-performing setting in both Figure 2(c) and Table 1. This reversal is consistent with the gain–cost view: finer splitting may increase the group-wise whitening gain, but it also introduces a grouping-induced norm cost. When the additional

gain no longer compensates for this cost, a coarser grouping can become preferable. This also helps explain why fully head-wise Muon does not necessarily outperform Full-QKV Muon, despite its larger first-order group-wise whitening term.

Table 1: Validation loss under different Group Muon settings. In the Fixed-QK+V setting, QK is fixed to its best grouping configuration($g = 6$, rand.), while the V grouping is varied.

Baselines		Validation loss				
Full-QKV Muon		3.2780				
Head-wise QKV MuonSplit		3.2799				
GroupMuon(**ours**)	Rule	$g = 1$	$g = 2$	$g = 3$	$g = 4$	$g = 6$
QK	adj.	3.2787	3.2774	3.2770	3.2757	3.2743
	int.	3.2778	3.2776	3.2767	3.2745	3.2750
	rand.	**3.2759**	**3.2758**	**3.2746**	**3.2731**	**3.2722**
V	adj.	3.2806	3.2765	3.2780	3.2743	3.2781
	rand.	3.2806	3.2792	3.2769	3.2765	3.2769
Fixed-QK+V	adj.	3.2799	3.2770	3.2762	3.2761	3.2746
	rand.	3.2788	3.2782	3.2740	3.2744	3.2752

By contrast, $g = 6$ provides a more stable grouping scale in the later stage, retaining more structure than Full-QKV Muon while avoiding excessive head-wise fragmentation. Accordingly, Q/K random grouping with $g = 6$ achieves the best validation loss among the settings, whereas additional V grouping brings no consistent improvement. This suggests that moderately coarse Q/K grouping can be preferable for sustained training performance.

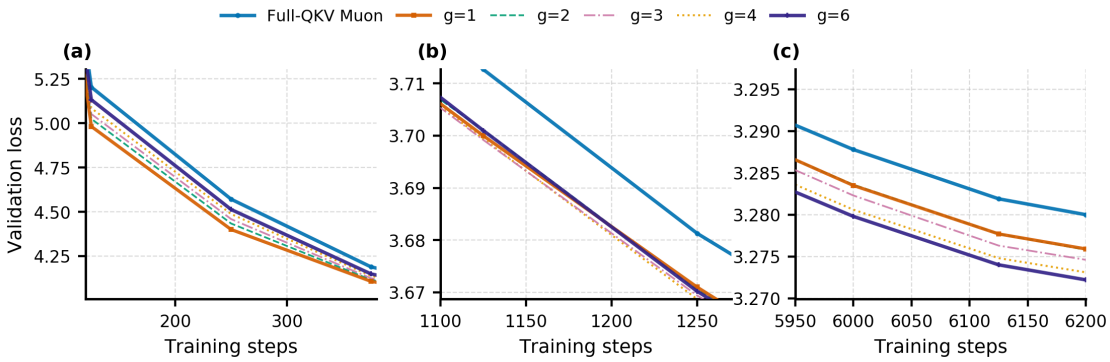

Figure 2: Validation loss of Full-QKV Muon and random Group Muon. In the early stage (a), smaller group sizes perform better, with $g = 1$ achieving the lowest validation loss. In the middle and late stages, (b) and (c), larger group sizes become more favorable, with $g = 6$ surpassing $g = 1$ and performing best.

5. Conclusion

Grouping attention heads can accelerate Muon when the group-wise whitening gain outweighs the grouping-induced norm cost. This gain–cost view motivates intermediate grouping scales, and our experiments find Q/K random grouping with $g = 6$ to be the best among the tested settings.

References

[1] Kwangjun Ahn, Noah Amsel, and John Langford. Dion2: A simple method to shrink matrix in muon, 2025. URL https://arxiv.org/abs/2512.16928.

[2] Kwangjun Ahn, Byron Xu, Natalie Abreu, Ying Fan, Gagik Magakyan, Pratyusha Sharma, Zheng Zhan, and John Langford. Dion: Distributed orthonormalized updates, 2025. URL https://arxiv.org/abs/2504.05295.

[3] Essential AI, :, Ishaan Shah, Anthony M. Polloreno, Karl Stratos, Philip Monk, Adarsh Chaluvaraju, Andrew Hojel, Andrew Ma, Anil Thomas, Ashish Tanwer, Darsh J Shah, Khoi Nguyen, Kurt Smith, Michael Callahan, Michael Pust, Mohit Parmar, Peter Rushton, Platon Mazarakis, Ritvik Kapila, Saurabh Srivastava, Somanshu Singla, Tim Romanski, Yash Vanjani, and Ashish Vaswani. Practical efficiency of muon for pre-training, 2025. URL https://arxiv.org/abs/2505.02222.

[4] Valentyn Boreiko, Zhiqi Bu, and Sheng Zha. Towards understanding orthogonalization in muon. In *ES-FoMo III: 3rd Workshop on Efficient Systems for Foundation Models*, 2025. URL https://openreview.net/forum?id=HhEWl5ovTX.

[5] Kushal Chakrabarti and Nirmal Balachundar. Multi-head attention is a multi-player game, 2026. URL https://arxiv.org/abs/2602.00861.

[6] GLM-5-Team, :, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, Huilong Chen, Jiajie Zhang, Jian Jiao, Jiaqi Guo, Jingsen Wang, Jingzhao Du, Jinzhu Wu, Kedong Wang, Lei Li, Lin Fan, Lucen Zhong, Mingdao Liu, Mingming Zhao, Pengfan Du, Qian Dong, Rui Lu, Shuang-Li, Shulin Cao, Song Liu, Ting Jiang, Xiaodong Chen, Xiaohan Zhang, Xuancheng Huang, Xuezhen Dong, Yabo Xu, Yao Wei, Yifan An, Yilin Niu, Yitong Zhu, Yuanhao Wen, Yukuo Cen, Yushi Bai, Zhongpei Qiao, Zihan Wang, Zikang Wang, Zilin Zhu, Ziqiang Liu, Zixuan Li, Bojie Wang, Bosi Wen, Can Huang, Changpeng Cai, Chao Yu, Chen Li, Chengwei Hu, Chenhui Zhang, Dan Zhang, Daoyan Lin, Dayong Yang, Di Wang, Ding Ai, Erle Zhu, Fangzhou Yi, Feiyu Chen, Guohong Wen, Hailong Sun, Haisha Zhao, Haiyi Hu, Hanchen Zhang, Hanrui Liu, Hanyu Zhang, Hao Peng, Hao Tai, Haobo Zhang, He Liu, Hongwei Wang, Hongxi Yan, Hongyu Ge, Huan Liu, Huanpeng Chu, Jia'ni Zhao, Jiachen Wang, Jiajing Zhao, Jiamin Ren, Jiapeng Wang, Jiaxin Zhang, Jiayi Gui, Jiayue Zhao, Jijie Li, Jing An, Jing Li, Jingwei Yuan, Jinhua Du, Jinxin Liu, Junkai Zhi, Junwen Duan, Kaiyue Zhou, Kangjian Wei, Ke Wang, Keyun Luo, Laiqiang Zhang, Leigang Sha, Liang Xu, Lindong Wu, Lintao Ding, Lu Chen, Minghao Li, Nianyi Lin, Pan Ta, Qiang Zou, Rongjun Song, Ruiqi Yang, Shangqing Tu, Shangtong Yang, Shaoxiang Wu, Shengyan Zhang, Shijie Li, Shuang Li, Shuyi Fan, Wei Qin, Wei Tian, Weining Zhang, Wenbo Yu, Wenjie Liang, Xiang Kuang, Xiangmeng Cheng, Xiangyang Li, Xiaoquan Yan, Xiaowei Hu, Xiaoying Ling, Xing Fan, Xingye Xia, Xinyuan Zhang, Xinze Zhang, Xirui Pan, Xu Zou, Xunkai Zhang, Yadi Liu, Yandong Wu, Yanfu Li, Yidong Wang, Yifan Zhu, Yijun Tan, Yilin Zhou, Yiming Pan, Ying Zhang, Yinpei Su, Yipeng Geng, Yong Yan,

Yonglin Tan, Yuean Bi, Yuhan Shen, Yuhao Yang, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yurong Wu, Yutao Zhang, Yuxi Duan, Yuxuan Zhang, Zezhen Liu, Zhengtao Jiang, Zhenhe Yan, Zheyu Zhang, Zhixiang Wei, Zhuo Chen, Zhuoer Feng, Zijun Yao, Ziwei Chai, Ziyuan Wang, Zuzhou Zhang, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. Glm-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

[7] Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

[8] Ahmed Khaled, Kaan Ozkara, Tao Yu, Mingyi Hong, and Youngsuk Park. Muonbp: Faster muon via block-periodic orthogonalization, 2025. URL https://arxiv.org/abs/2510.16981.

[9] 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. URL https://arxiv.org/abs/2502.16982.

[10] Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

[11] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI*, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. Accessed: 2024-11-15.

[12] samacqua. Modded-nanogpt speedrun pr 253: Pair-head muon, April 2026. URL https://github.com/KellerJordan/modded-nanogpt/pull/253.

[13] Egor Shulgin, Sultan AlRashed, Francesco Orabona, and Peter Richtárik. Beyond the ideal: Analyzing the inexact muon update, 2025. URL https://arxiv.org/abs/2510.19933.

[14] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[16] Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

Appendix A. Related Work

Muon for large-scale language-model training. Muon applies matrix orthogonalization to optimizer updates and has rapidly become a practical alternative to AdamW for hidden layers in neural-network training [7]. Recent large-scale studies show that Muon can improve the compute–loss trade-off in LLM pretraining [3, 9], while complementary analyses investigate learning-rate transfer, spectral regularization, weight decay, and the effect of approximate Newton–Schulz orthogonalization [4, 13]. Our work is orthogonal to these scaling and approximation studies: we keep the Muon update family fixed and ask which attention-head granularity should define the matrix blocks being orthogonalized.

Block-wise and distributed orthogonalization. Several recent methods reduce the cost of orthogonalized updates by changing the matrix on which orthogonalization is performed. Dion replaces dense Newton–Schulz orthogonalization with distributed low-rank orthonormalized updates designed for sharded weights [2], and Dion2 further simplifies this idea by sampling rows or columns to shrink the orthogonalization problem [1]. Block-wise Muon variants also split weight matrices into tiles to make tensor-parallel training more efficient [4]. MuonBP applies block-wise orthogonalization on device-local matrix shards and periodically performs full orthogonalization to recover the stability of baseline Muon [8]. The key distinction from MuonBP is that our blocks are *attention-head groups*, not primarily device shards: MuonBP is motivated by distributed communication and throughput, whereas Group Muon treats the head group size and grouping rule as optimizer hyperparameters and analyzes the trade-off between group-wise whitening gain and grouping-induced norm cost.

Attention-head structure. Our focus on head grouping is motivated by the internal multi-head structure of attention. Recent work models multi-head attention as an interacting multi-player system, emphasizing that different heads can compete, coordinate, and induce cross-head externalities [5]. This perspective is consistent with treating attention heads as meaningful optimizer units, but our goal is different: rather than regularizing head interactions, we study how the orthogonalization scope of Muon should be chosen over heads or groups of heads.

Appendix B. Proof of Proposition 1

Proof By β-smoothness(3), for any update direction O and stepsize $\eta > 0$, taking $\Delta = -\eta O$ gives

$$\mathcal{L}(W - \eta O) \leq \mathcal{L}(W) - \eta\langle G, O\rangle + \frac{\beta\eta^2}{2}\|O\|_F^2. \tag{5}$$

We first apply (5) to the full matrix Muon direction

$$O_{\text{all}} = P(G).$$

Using the polar-factor identities

$$\langle G, P(G)\rangle = \|G\|_*, \qquad \|P(G)\|_F^2 = \operatorname{rank}(G) = r,$$

we obtain

$$\mathcal{L}(W_{\text{all}}^{+}) \leq \mathcal{L}(W) - \eta\|G\|_{*} + \frac{\beta\eta^{2}}{2}r. \tag{6}$$

Therefore the guaranteed one-step decrease lower bound for full matrix Muon is

$$D_{\text{all}} = \eta\|G\|_{*} - \frac{\beta\eta^{2}}{2}r. \tag{7}$$

Next consider the Group Muon direction

$$O_{\text{grp}} = \begin{bmatrix} P(G_1)^{\top} & P(G_2)^{\top} & \cdots & P(G_M)^{\top} \end{bmatrix}^{\top}.$$

Since G and O_{grp} share the same block structure, we have

$$\langle G, O_{\text{grp}} \rangle = \sum_{i=1}^{M} \langle G_i, P(G_i) \rangle = \sum_{i=1}^{M} \|G_i\|_{*}. \tag{8}$$

Similarly, because the blocks are disjoint,

$$\|O_{\text{grp}}\|_{F}^{2} = \sum_{i=1}^{M} \|P(G_i)\|_{F}^{2} = \sum_{i=1}^{M} \operatorname{rank}(G_i) = \sum_{i=1}^{M} r_i. \tag{9}$$

Substituting these two identities into (5) gives

$$\mathcal{L}(W_{\text{grp}}^{+}) \leq \mathcal{L}(W) - \eta \sum_{i=1}^{M} \|G_i\|_{*} + \frac{\beta\eta^{2}}{2} \sum_{i=1}^{M} r_i. \tag{10}$$

Thus the guaranteed one-step decrease lower bound for Group Muon is

$$D_{\text{grp}} = \eta \sum_{i=1}^{M} \|G_i\|_{*} - \frac{\beta\eta^{2}}{2} \sum_{i=1}^{M} r_i. \tag{11}$$

Group Muon has a larger guaranteed descent lower bound than full matrix Muon when

$$D_{\text{grp}} > D_{\text{all}}.$$

Substituting the two expressions above, this condition becomes

$$\eta \sum_{i=1}^{M} \|G_i\|_{*} - \frac{\beta\eta^{2}}{2} \sum_{i=1}^{M} r_i > \eta\|G\|_{*} - \frac{\beta\eta^{2}}{2} r.$$

Rearranging and dividing by $\eta > 0$ yields

$$\sum_{i=1}^{M} \|G_i\|_{*} - \|G\|_{*} > \frac{\beta\eta}{2} \left(\sum_{i=1}^{M} r_i - r \right),$$

which proves the comparison criterion. ■

Appendix C. Experimental Setting

GPT-2 Small on FineWeb.

Models and Architecture. We conduct all head-grouping experiments on a GPT-2 Small style decoder-only Transformer [11, 15] trained from scratch on FineWeb. The model uses $n_{\text{layer}} = 12$, $d_{\text{model}} = 768$, and $n_{\text{head}} = 12$, giving a head dimension of $d_{\text{head}} = 64$. Following the modded NanoGPT-style implementation, the architecture uses tied input/output embeddings, RMSNorm[16], Rotary Positional Embeddings (RoPE)[14], bias-free linear layers, and a decoder-only causal attention stack.

Dataset and Tokenization. All experiments use the FineWeb10B binary shards [10] with the GPT-2 BPE vocabulary of size 50,257. We use a context length of $L = 1024$ tokens. Validation loss is evaluated every 125 optimization steps on a fixed validation budget of 10,485,760 tokens.

Shared Hyperparameters. All runs use 4-GPU distributed data parallel training with a global batch size of 512 sequences, corresponding to 524,288 tokens per optimization step. Each experiment is trained for 6,200 steps, or approximately 3.25B training tokens. We use a constant learning-rate phase followed by a linear warmdown over the final 1,800 steps, with no warmup.

Optimizer. We use a split optimization strategy. The language-model head and tied token embeddings are optimized with AdamW using learning rate 3.6×10^{-3}, betas $(0.9, 0.95)$, and zero weight decay. The Transformer blocks are optimized with Muon using learning rate 3.6×10^{-4} and momentum 0.95. Unless otherwise specified, Muon orthogonalization uses the Newton-Schulz backend; grouped head updates use the batched Newton-Schulz backend.

Optimizer. We use a split optimization strategy. The language-model head and tied token embeddings are optimized with AdamW using learning rate 3.6×10^{-3}, betas $(0.9, 0.95)$, and zero weight decay. The Transformer blocks are optimized with Muon using base learning rate 3.6×10^{-4} and momentum 0.95. For head-wise and group-wise Muon variants, splitting a Q/K/V matrix changes the effective matrix shape on which the polar-factor update is applied. We therefore apply a shape-aware learning-rate transfer rule[2]: the Muon learning rate is transferred from the original full-matrix shape to the corresponding grouped matrix shape, so that the update scale remains comparable across Full-QKV Muon, Head-wise QKV Muon, and Group Muon variants. Unless otherwise specified, Muon orthogonalization uses the Newton–Schulz backend, and grouped head updates use the batched Newton–Schulz backend.

Table 2: Shared hyperparameters for the GPT-2 Small head-grouping experiments.

Model	Steps	Tokens	Seq. Len.	Global Batch	Head LR	Body LR	WD	Schedule
GPT-2 Small	6,200	∼3.25B	1024	512	3.6e-3	3.6e-4	0.0	Linear warmdown, 1,800 steps

2. We follow the Muon learning-rate transfer practice described in https://kexue.fm/archives/11416.

Compared Methods. We organize the experiments according to the same categories used in Table 1. The two baselines are not parameterized by the group size g. **Full-QKV Muon** applies the standard Muon update to the packed QKV projection matrix. **Head-wise QKV Muon** first splits the packed QKV projection into Q, K, and V blocks, then further splits each block into individual attention heads and applies Muon independently to each head.

For grouped variants, we use the attention head as the basic unit. Since GPT-2 Small has $H = 12$ attention heads, a group size $g \in \{1, 2, 3, 4, 6\}$ partitions the heads into H/g groups. We consider three grouped settings:

- **QK**: apply Group Muon only to the Q and K projections;
- **V**: apply Group Muon only to the V projection;
- **Fixed-QK+V**: fix Q and K to the best QK configuration, namely $g = 6$ with random grouping, and sweep the grouping configuration of V.

Table 3: Experiment settings corresponding to Table 1. For grouped variants, g denotes the number of heads per group. In random grouping, the head partition is resampled at every optimizer step.

Setting	**Grouped target**	**Grouping rule**	**Sweep**
Full-QKV Muon	QKV	–	–
Head-wise QKV Muon	QKV	per-head	–
QK	Q, K	adj.	$g \in \{1, 2, 3, 4, 6\}$
QK	Q, K	int.	$g \in \{1, 2, 3, 4, 6\}$
QK	Q, K	rand.	$g \in \{1, 2, 3, 4, 6\}$
V	V	adj.	$g_V \in \{1, 2, 3, 4, 6\}$
V	V	rand.	$g_V \in \{1, 2, 3, 4, 6\}$
Fixed-QK+V	Q, K, V	adj. for V	$g_{QK} = 6$ with rand., $g_V \in \{1, 2, 3, 4, 6\}$
Fixed-QK+V	Q, K, V	rand. for V	$g_{QK} = 6$ with rand., $g_V \in \{1, 2, 3, 4, 6\}$

Grouping Rules. We index the 12 attention heads by $\{0, 1, \ldots, 11\}$. For a group size g, the number of groups is $K = 12/g$. We use the following grouping rules.

Adjacent grouping partitions consecutive heads into the same group:

$$\{0, \ldots, g-1\},\ \{g, \ldots, 2g-1\},\ \ldots,\ \{12-g, \ldots, 11\}.$$

For example, when $g = 3$, adjacent grouping gives

$$\{0, 1, 2\}, \quad \{3, 4, 5\}, \quad \{6, 7, 8\}, \quad \{9, 10, 11\}.$$

Interval grouping distributes nearby heads into different groups by using a fixed stride $K = 12/g$. The j-th group is

$$\{j,\ j+K,\ j+2K,\ \ldots,\ j+(g-1)K\}, \qquad j = 0, \ldots, K-1.$$

For example, when $g = 3$, we have $K = 4$, so interval grouping gives

$$\{0, 4, 8\}, \quad \{1, 5, 9\}, \quad \{2, 6, 10\}, \quad \{3, 7, 11\}.$$

When $g = 6$, we have $K = 2$, so interval grouping gives

$$\{0, 2, 4, 6, 8, 10\}, \quad \{1, 3, 5, 7, 9, 11\}.$$

Random grouping first samples a random permutation of the 12 heads and then splits the permuted list into contiguous groups of size g. For example, if $g = 3$ and the sampled permutation is

$$(7, 2, 10, 0, 5, 9, 1, 4, 11, 3, 6, 8),$$

then the random groups are

$$\{7, 2, 10\}, \quad \{0, 5, 9\}, \quad \{1, 4, 11\}, \quad \{3, 6, 8\}.$$

In all random grouping experiments, the random partition is resampled at every optimizer step.

Fixed-QK+V and V-only Controls. The **Fixed-QK+V** experiments test whether additionally applying group-wise whitening to V improves upon the best QK-only configuration. In these runs, Q and K are fixed to random grouping with $g_{QK} = 6$, while the group size and grouping rule of V are varied. The **V** experiments apply Group Muon only to the V projection, leaving Q and K on their full matrix Muon paths. These controls are used to separate the effect of grouping Q/K from the effect of grouping V.

Appendix D. Experimental Observation

D.1. Nearly Full Rank

We empirically examine the rank structure of the attention(QK)-projection gradients and momentum during training. For each recorded step, we compute the rank of the full WQ/WK matrices and of their row-wise grouped submatrices. Ranks are normalized by the corresponding row dimension, so a value close to 1 indicates near-full row rank.

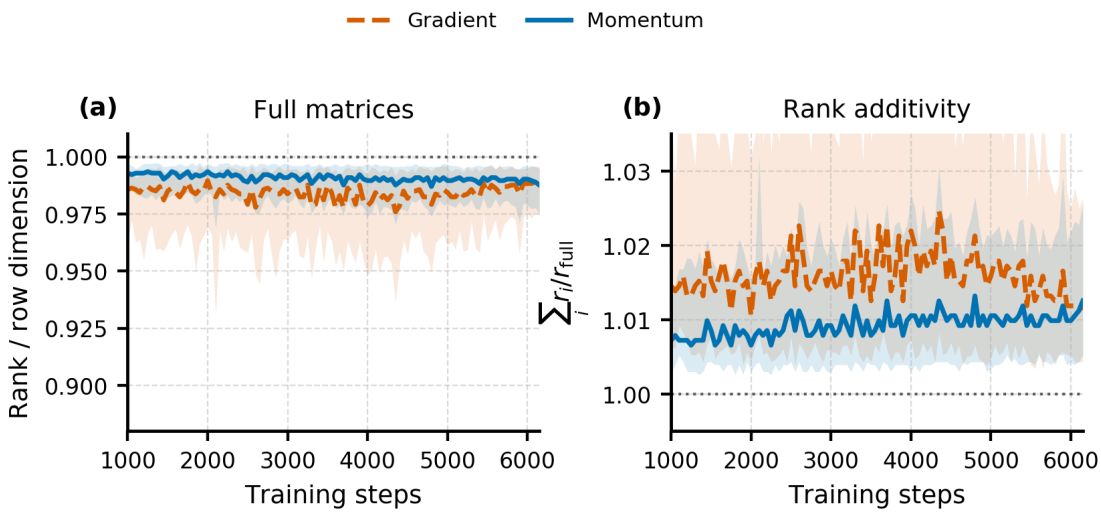

Figure 3: Rank behavior of gradients and momentum during training. Both full gradients and momentum remain close to full row rank, with momentum exhibiting a more stable rank ratio than gradients. For grouped updates, the sum of group ranks stays close to the full-matrix rank throughout the main training regime.

Figure 3 shows that both the full matrices and their grouped row partitions remain close to full row rank after the early training phase. This supports the empirical condition used in our analysis: for a row-wise partition with group row sizes m_i, the full rank satisfies $r \approx m$, each group rank satisfies $r_i \approx m_i$, and the aggregate grouped rank $\sum_i r_i$ remains close to the full-matrix rank.

D.2. Grouping-induced Norm Cost

Non-ideal Muon whitening. The rank-dependent expression above relies on the ideal polar-factor identity $\|P(G)\|_F^2 = \mathrm{rank}(G)$. For a practical Muon implementation, however, the whitening direction is produced by an approximate operator, such as a finite number of Newton–Schulz iterations. Let $O_{\text{full}} = \mathcal{A}(G)$ denote the approximate whitened update computed from the full gradient, and let $O_i = \mathcal{A}(G_i)$ denote the approximate whitened update computed from the i-th group gradient. The group-wise update is then

$$O_{\text{grp}} = \begin{bmatrix} O_1^\top & O_2^\top & \cdots & O_M^\top \end{bmatrix}^\top.$$

Applying the smoothness bound to the full update gives

$$\mathcal{L}(W - \eta O_{\text{full}}) \le \mathcal{L}(W) - \eta \langle G, O_{\text{full}} \rangle + \frac{\beta \eta^2}{2} \|O_{\text{full}}\|_F^2.$$

Similarly, applying it to the group-wise update gives

$$\mathcal{L}(W - \eta O_{\text{grp}}) \le \mathcal{L}(W) - \eta \sum_{i=1}^{M} \langle G_i, O_i \rangle + \frac{\beta \eta^2}{2} \sum_{i=1}^{M} \|O_i\|_F^2.$$

Therefore, the additional second-order cost induced by replacing full-matrix whitening with group-wise whitening is

$$\frac{\beta\eta^2}{2}\left(\sum_{i=1}^{M}\|O_i\|_F^2 - \|O_{\text{full}}\|_F^2\right).$$

Equivalently, after dividing the comparison criterion by $\eta > 0$, the corresponding grouping-induced norm cost is

$$\frac{\beta\eta}{2}\left(\sum_{i=1}^{M}\|O_i\|_F^2 - \|O_{\text{full}}\|_F^2\right).$$

Figure 4 shows that this Frobenius-norm gap is clearly nonzero in practice: head-wise whitening consistently yields larger $\sum_i \|O_i\|_F^2$ than the full-matrix counterpart $\|O_{\text{full}}\|_F^2$.

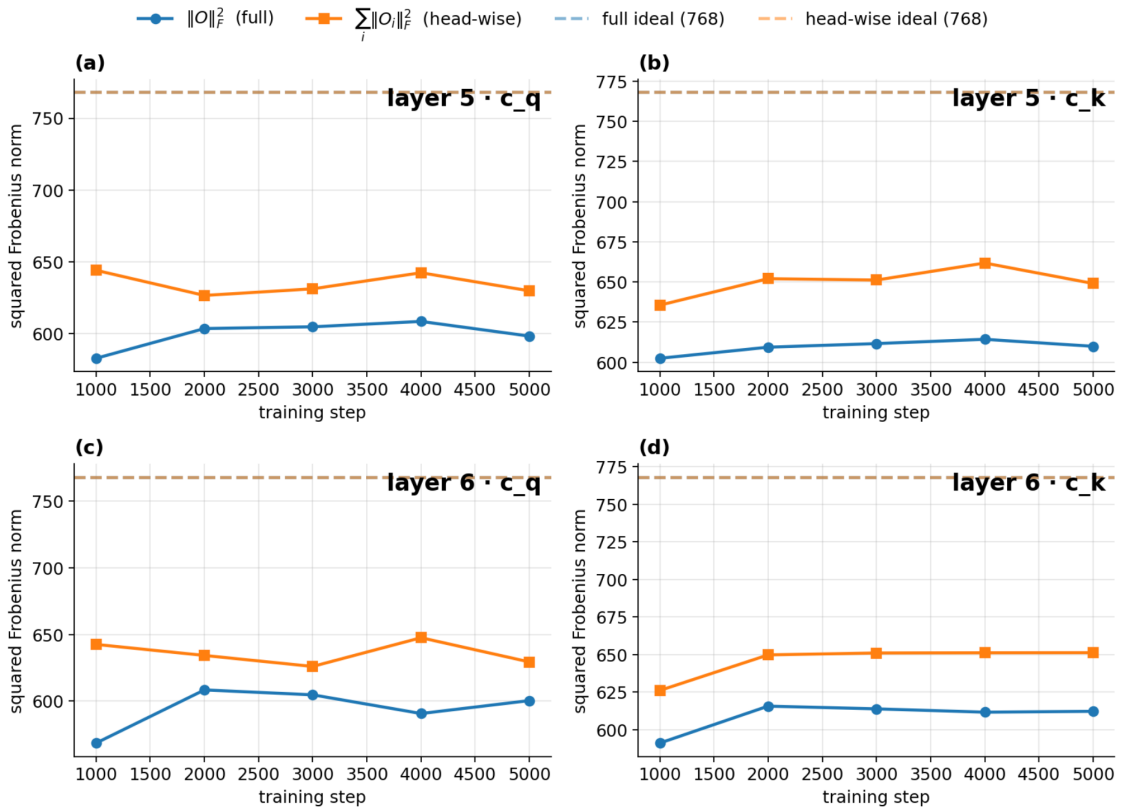

Figure 4: Squared Frobenius norm of the whitened Muon update under full-matrix whitening and head-wise whitening. Panels (a)–(d) report representative Q/K projections from layers 5 and 6; across all cases, head-wise whitening yields a larger $\sum_i \|O_i\|_F^2$ than the full-matrix counterpart $\|O\|_F^2$. This indicates that head-wise whitening introduces a larger grouping-induced norm cost, while both remain below the ideal full-rank value of 768.

Appendix E. Aligned Low-Rank Example

We provide a simple aligned low-rank example to illustrate that the gain–cost trade-off can already appear under ideal polar-factor Muon. Consider k nonzero group gradients that

share the same right singular direction:

$$G_i = a_i u_i v^\top, \qquad \|u_i\|_2 = \|v\|_2 = 1, \qquad i = 1, \ldots, k,$$

where $a_i \neq 0$. Then each nonzero group gradient has rank one. At the same time, since all groups share the same right direction v, the concatenated gradient also has rank one. Therefore,

$$\sum_{i=1}^{k} r_i - r = k - 1.$$

The group-wise nuclear-norm gain is

$$\sum_{i=1}^{k} \|G_i\|_* - \|G\|_* = \sum_{i=1}^{k} |a_i| - \left(\sum_{i=1}^{k} a_i^2 \right)^{1/2}.$$

If the group strengths are equal, i.e., $|a_i| = a$, then

$$\sum_{i=1}^{k} \|G_i\|_* - \|G\|_* = a(k - \sqrt{k}).$$

Thus the comparison criterion (4) becomes

$$a(k - \sqrt{k}) > \frac{\beta\eta}{2}(k - 1).$$

This example shows that finer splitting can increase both the group-wise whitening gain and the grouping-induced norm cost. Therefore, even under ideal polar-factor Muon, grouping is beneficial only when the former dominates the latter.