Gated GRIT: Geometry-Aware Parameter-Efficient Fine-Tuning for Multimodal Vision-Language Models

Anonymous ACL submission

Abstract

Parameter-efficient fine-tuning (PEFT) has become a practical approach for adapting large vision-language models while maintaining a small trainable footprint. Among PEFT methods, Low-Rank Adaptation (LoRA) reduces the cost of adaptation through low-rank updates. In this work, we propose **Gated GRIT**, a multimodal extension of GRIT that introduces curvature-derived rank allocation for vision-language adaptation. Instead of treating all rank directions uniformly, our method derives rank-wise importance weights directly from the Fisher spectrum obtained during K-FAC approximation and uses them to selectively modulate low-rank updates. We evaluate the proposed approach on visual question answering using Gemma-4B and Llama-3.2-11B-Vision backbones under identical parameter budgets. Experimental results show that Gated GRIT achieves competitiveness relative to LoRA and standard GRIT while preserving parameter efficiency. To foster future research in this area, we make our work publicly available.

1 Introduction

Vision-language models (VLMs) have achieved strong performance across multimodal tasks such as visual question answering, image captioning, and visual reasoning. Recent models including Gemma (Team et al., 2024) and Llama-Vision (Grattafiori et al., 2024) leverage large-scale multimodal pretraining to learn rich visual and textual representations. However, adapting these models to downstream tasks remains computationally expensive due to their large parameter counts.

Parameter-efficient fine-tuning (PEFT) addresses this challenge by updating only a small subset of model parameters while keeping the pretrained backbone frozen. Among PEFT methods, Low-Rank Adaptation (LoRA) (Hu et al., 2021) introduces trainable low-rank matrices into transformer layers, substantially reducing memory and computational requirements. That said, LoRA operates purely through first-order optimization; it has no built-in way to account for the local curvature of the loss landscape.

Geometry-Aware Reprojection Instruction Tuning (GRIT) builds upon LoRA by incorporating Fisher-guided preconditioning, allowing updates to follow directions that are more informative under the local loss geometry. GRIT improves optimization efficiency but stops short of controlling how adaptation capacity is spread across individual rank directions within the low-rank subspace.

Not all rank directions contribute equally during adaptation. Based on this observation, we introduce **Gated GRIT**, a curvature-aware extension of GRIT that allocates adaptation capacity across individual low-rank directions. K-FAC already provides a Fisher spectrum for each low-rank module. We use this spectrum to estimate the relative importance of different rank directions and scale their contributions during adaptation. Unlike conventional gating approaches, Gated GRIT introduces no additional trainable parameters and operates entirely within the curvature information already computed by GRIT.

We study Gated GRIT on a visual question answering benchmark using Gemma-4B (Team et al., 2024) and Llama-3.2-11B-Vision (Grattafiori et al., 2024), with all methods trained under identical parameter budgets.Despite introducing no additional trainable parameters, Gated GRIT performs competitively against both LoRA and standard GRIT. Effective-rank analysis suggests that the adaptation concentrates in a small subset of directions.

We mention our contributions as:

- We introduce Gated GRIT, a curvature-aware rank allocation mechanism for multimodal low-rank adaptation.
- We show how Fisher-spectrum information can be used to assign different levels of empha-

sis to different rank directions without adding new trainable parameters.

- We demonstrate competitive performance on Gemma-4B and Llama-3.2-11B-Vision models under identical parameter budgets.

2 Related Work

Parameter-Efficient Fine-Tuning. Parameter-efficient fine-tuning (PEFT) adapts large foundation models while updating only a small subset of parameters. LoRA (Hu et al., 2021) introduces trainable low-rank matrices into transformer layers, substantially reducing memory and computational requirements while maintaining competitive downstream performance. Subsequent methods such as QLoRA (Dettmers et al., 2023) further improve efficiency through quantization-aware adaptation.

Geometry-Aware Adaptation. The recent work has explored incorporating geometric information into low-rank adaptation. GRIT (Saha et al., 2026) extends LoRA through Fisher-guided preconditioning, enabling updates that better align with the local curvature of the loss landscape. GRIT exploits second-order information in the low-rank subspace to achieve optimization efficiency while keeping parameter efficiency.

Multimodal PEFT. PEFT methods have been widely applied to vision-language models for multimodal tasks such as visual question answering and image understanding. While existing approaches primarily focus on reducing computational cost, our work investigates selective allocation of adaptation capacity within the low-rank subspace through curvature-derived rank weighting. The Recent work on superweights also claims that a small subset of parameters might dominate model behavior (Yu et al., 2025). We have explored a similar idea for the low-rank directions.

3 Methodology

3.1 LoRA and GRIT

Low-Rank Adaptation (LoRA) (Hu et al., 2021) performs parameter-efficient fine-tuning by constraining task-specific updates to a low-rank subspace. Given a pretrained weight matrix $W_0 \in \mathbb{R}^{d\times k}$, LoRA parameterizes the adaptation as

$$W = W_0 + \Delta W,$$

where the update matrix is expressed as

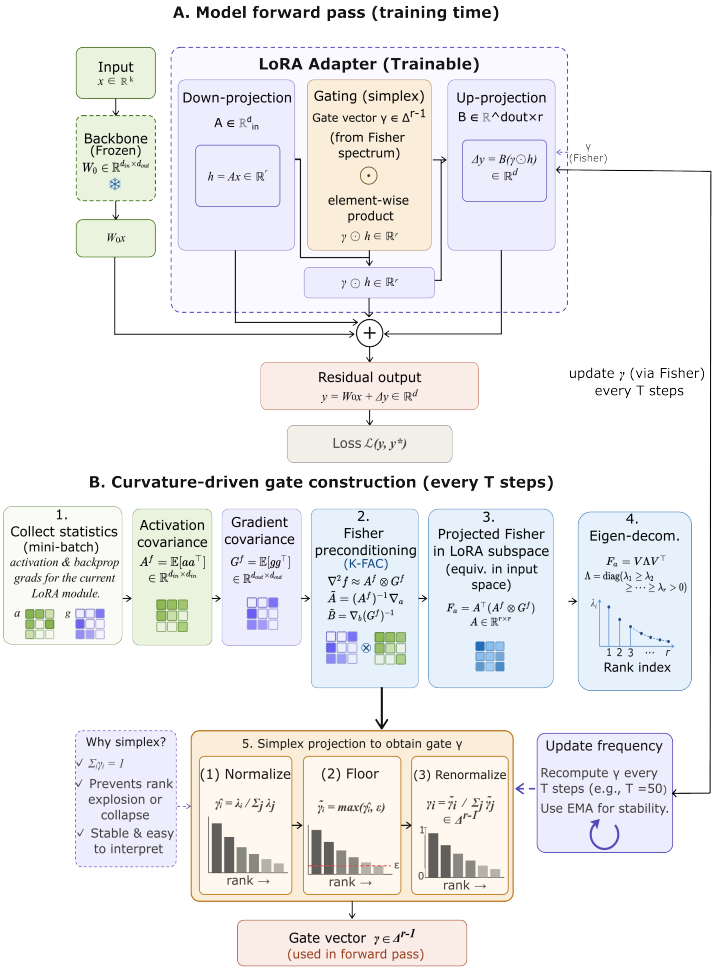

Figure 1: The Gate Module computes importance weights for each rank by drawing on the Fisher eigenspectrum, then uses these weights to scale the LoRA update prior to Fisher preconditioning (shown in blue). Also, this entire mechanism adds no new trainable parameters to the model.

$$\Delta W = BA. \tag{1}$$

Here, $A \in \mathbb{R}^{r\times k}$ and $B \in \mathbb{R}^{d\times r}$ are trainable low-rank matrices with rank $r \ll \min(d, k)$. This methodology significantly reduces the number of trainable parameters, and also preserves the adaptation performance.

Despite its efficiency, LoRA (Hu et al., 2021) relies on standard gradient updates and does not directly account for the curvature of the optimization landscape. Unlike LoRA, GRIT (Saha et al., 2026) incorporates information about the local curvature of the loss landscape. It does so through a K-FAC (Martens and Grosse, 2020) approximation of the Fisher Information Matrix.

Let ∇_A and ∇_B denote the gradients with respect to the LoRA parameters. GRIT computes Fisher-preconditioned gradients as

$$\tilde{\nabla}_A = A_f^{-1}\nabla_A, \qquad \tilde{\nabla}_B = \nabla_B G_f^{-1}, \tag{2}$$

where A_f and G_f represent the Kronecker factors of the Fisher approximation. This allows curvature information from the Fisher approximation to influence the update.

3.2 Motivation for Rank Allocation

Despite using curvature-aware optimization, GRIT (Saha et al., 2026) does not provide a mechanism for assigning different levels of importance to different rank directions. Hence, all the rank components will remain active after preconditioning, irrespective of their relative importance during adaptation.

We hypothesize that adaptation capacity can be utilized more efficiently if rank directions are weighted according to their contribution to the local curvature structure. This is precisely what motivates a rank-wise allocation mechanism, one that puts more weight on dominant directions and dials back the weaker ones. Instead of introducing additional trainable parameters, we derive these importance weights directly from the Fisher spectrum already computed during GRIT optimization.

3.3 Curvature-Derived Gate Computation

The activation-side covariance matrix calculated during K-FAC (Martens and Grosse, 2020) captures the distribution of the information within the low-rank adaptation subspace. Its eigenspectrum helps us in providing a compact characterization of the relative importance of different rank directions. Intuitively, directions associated with larger eigenvalues contribute more strongly to the local curvature structure, and directions associated with smaller eigenvalues contribute comparatively less.

Let $A_f \in \mathbb{R}^{r\times r}$ denote the activation-side Fisher covariance matrix. We compute its eigendecomposition as

$$A_f = V\Lambda V^\top, \tag{3}$$

where:

$$\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_r) \tag{4}$$

contains the eigenvalues of A_f. We use these eigenvalues as curvature-derived importance scores for the corresponding rank directions, which helps us in assigning numerical priority of rank directions instead of treating all the directions equally.

To obtain a normalized allocation over rank components, the eigenvalues are transformed into a gate vector:

$$\hat{g}_i = \frac{\lambda_i}{\sum_{j=1}^{r}\lambda_j}. \tag{5}$$

We prevent the complete suppression of any direction; by keeping a minimum floor value ϵ and the gates are renormalized:

$$g_i = \frac{\max(\hat{g}_i, \epsilon)}{\sum_{j=1}^{r}\max(\hat{g}_j, \epsilon)}. \tag{6}$$

The resulting vector we get is,

$$g = [g_1, g_2, \ldots, g_r] \tag{7}$$

The resulting gate vector assigns different levels of importance to different rank directions instead of treating them uniformly.

3.4 Gated Low-Rank Updates

Given the gate vector g, the LoRA update is modified as

$$\Delta W = B \cdot \mathrm{diag}(g) \cdot A, \tag{8}$$

where the $\mathrm{diag}(g)$ denotes a diagonal matrix whose entries are corresponding to the gate values.

So, expanding the update will yield us

$$\Delta W = \sum_{i=1}^{r} g_i\, b_i a_i^\top, \tag{9}$$

where b_i and a_i represent the i-th rank component of the low-rank decomposition. This allows different rank directions to contribute unequally during adaptation. Directions with larger gate values play a larger role in the update, while directions with smaller values remain active but contribute less. In Gated GRIT, Fisher information is used in two different ways. GRIT(Saha et al., 2026) relies on it during optimization, whereas the gate values use the same information to assign different levels of emphasis to different rank directions. These two mechanisms operate together but influence different parts of the adaptation process.

4 Our Experimental Setup

Dataset. Experiments are conducted on the VQA Online dataset (chongyenchen/VQAonline) using a 655/83/83 train-validation-test split.

Models. Experiments are conducted using Gemma-4B and Llama-3.2-11B-Vision-Instruct as multimodal vision-language backbones. Here, we apply the LoRA, GRIT and Gated GRIT (our work) to identical target modules to ensure an even comparison.

Table 1: Main results on the Chongyenchen VQA Online dataset (Chen et al., 2024). Best results within each rank are shown in bold. Here, GRIT and Gated GRIT (our work) generally outperforms the LoRA under the same rank budget. Also, the rank-4 curvature-aware variants often match or exceed our rank-8 LoRA results, which shows more efficient usage of the low-rank adaptation capacity.

Method	Loss↓	Rouge-L↑	METEOR↑	MoverScore↑	BERTScore↑
Gemma-3-4B, $r = 4$, LoRA	15.081	0.1577	0.1595	0.5355	0.5485
Gemma-3-4B, $r = 4$, GRIT	15.636	**0.1828**	**0.1634**	**0.5357**	**0.5922**
Gemma-3-4B, $r = 4$, Gated GRIT	**9.438**	0.1746	0.1304	0.5329	0.5789
Gemma-3-4B, $r = 8$, LoRA	14.232	0.1570	0.1617	0.5345	0.5463
Gemma-3-4B, $r = 8$, GRIT	15.069	**0.1804**	**0.1634**	**0.5359**	**0.5892**
Gemma-3-4B, $r = 8$, Gated GRIT	**9.793**	0.1761	0.1632	0.5330	0.5872
Llama-3.2-11B-Vision, $r = 4$, LoRA	**12.889**	0.1497	**0.1504**	0.5231	**0.5042**
Llama-3.2-11B-Vision, $r = 4$, GRIT	13.057	0.1906	0.1489	0.5274	0.4941
Llama-3.2-11B-Vision, $r = 4$, Gated GRIT	13.977	**0.1947**	0.1470	**0.5279**	0.5020
Llama-3.2-11B-Vision, $r = 8$, LoRA	**11.964**	0.1444	**0.1524**	0.5217	0.4964
Llama-3.2-11B-Vision, $r = 8$, GRIT	12.433	**0.1951**	0.1469	**0.5276**	0.4917
Llama-3.2-11B-Vision, $r = 8$, Gated GRIT	13.706	0.1918	0.1456	0.5271	**0.4970**

Training Configuration. We have evaluated adaptation ranks of $r = 4$ and $r = 8$. All methods by us are trained under the same parameter budget and experimental protocol. For the Gated GRIT, gate values are computed from the Fisher spectrum after our K-FAC (Martens and Grosse, 2020) estimation and applied during low-rank adaptation.

Evaluation Metrics. The performance we have shown is evaluated using ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and BERTScore (Zhang et al., 2020). We also report test loss as an addition to our work to assess optimization quality under the identical training configurations.

5 Results

Gemma-4B. Across both the rank configurations, Gated GRIT achieves significantly lower test loss than both LoRA and standard GRIT while maintaining the competitive generation-quality metrics. These results suggest that curvature-derived rank allocation improves optimization efficiency and enables more effective utilization of the low-rank adaptation subspace.

Llama-3.2-11B-Vision. For the larger multimodal backbone, Gated GRIT achieves the strongest ROUGE-L score at rank 4 and competitive BERTScore at rank 8. Although LoRA attains slightly lower loss values, GRIT and Gated GRIT consistently improve generation-oriented metrics, indicating that curvature-aware adaptation can produce higher-quality textual responses without increasing parameter count.

Table 2: Average effective rank of Gated GRIT. The effective rank remains below the nominal rank across all the configurations.

Model	Rank	Effective Rank
Gemma-4B	4	2.50
Gemma-4B	8	3.59
Llama-3.2-11B	4	2.89
Llama-3.2-11B	8	4.05

Short discussion on importance of Effective Rank. Even when operating at nominal ranks of 4 and 8, Gated GRIT uses only a fraction of the available rank capacity in practice. In practice, only a subset of rank directions receives most of the adaptation capacity, and the remaining directions play a comparatively smaller role. The trend becomes more pronounced at higher nominal ranks, where adaptation capacity remains concentrated despite the availability of additional rank directions.

6 Conclusion

We presented Gated GRIT, a curvature-aware extension of GRIT for multimodal vision-language models. So, gate values are derived directly from the Fisher spectrum without introducing any additional trainable parameters. Across the evaluated models, Gated GRIT delivers performance comparable to existing low-rank adaptation methods while keeping the trainable parameter count unchanged. Hence, the findings shown suggest that curvature-derived rank allocation provides an effective mechanism for improving the usage of low-rank adaptation capacity in multimodal fine-tuning.

Limitations

Dependence on the Curvature Estimation. Here, our proposed gating mechanism derives the rank-wise importance from the Fisher spectrum, which is estimated by the K-FAC. Also, the quality of gate allocation depends on the accuracy of the underlying curvature approximation.

Limited Evaluation Scope. The Experiments we conducted are on a single visual question answering benchmark and two multimodal vision-language backbones. A broader evaluation across the additional multimodal tasks and the model scales is required by us to assess generalization.

Rank Allocation being dynamic. As training progresses, the Fisher spectrum and the resulting gate values continue to change. Exploring ways to adapt rank allocation to these longer-term shifts remains an open direction for future work.

Our Theoretical Understanding. Although our experiments consistently show the benefits of curvature-derived rank allocation, the precise relationship between Fisher spectra, effective rank, and downstream task performance is not yet fully understood.

Ethical Considerations

This work studies a parameter-efficient approach for adapting multimodal vision-language models using publicly available data. We do not collect new data or conduct experiments involving human participants. Like other vision-language models, the behavior of the resulting models may reflect biases present in the pretrained models and evaluation datasets. This work is intended for research purposes. Like other vision-language models, it may reflect biases present in the data used during training and evaluation, and these limitations should be considered when using the method in practice.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Chongyan Chen, Mengchen Liu, Noel Codella, Yunsheng Li, Lu Yuan, and Danna Gurari. 2024. Fully authentic visual question answering dataset from online communities. *Preprint*, arXiv:2311.15562.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

James Martens and Roger Grosse. 2020. Optimizing neural networks with kronecker-factored approximate curvature. *Preprint*, arXiv:1503.05671.

Pritish Saha, Chandrav Rajbangshi, Rudra Goyal, Mohit Goyal, Anurag Deo, Biswajit Roy, Ningthoujam Dhanachandra Singh, Raxit Goswami, and Amitava Das. 2026. Grit – geometry-aware peft with k-facpreconditioning, fisher-guided reprojection, anddynamic rank adaptation. *Preprint*, arXiv:2601.00231.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Mengxia Yu, De Wang, Qi Shan, Colorado J Reed, and Alvin Wan. 2025. The super weight in large language models. *Preprint*, arXiv:2411.07191.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *Preprint*, arXiv:1909.02622.

A Details of Additional Experiments

A.1 Dataset and the Evaluation Protocol

The experiments were conducted on the VQA Online benchmark using a train-validation-test split of 655/83/83 samples respectively. All methods were evaluated under identical experimental settings to ensure fair comparison between LoRA, GRIT, and the proposed Gated GRIT framework.

Evaluation was performed using ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), MoverScore (Zhao et al., 2019) (Zhao et al., 2019), and BERTScore (Zhang et al., 2020). In addition, test loss was monitored to evaluate optimization stability and convergence behavior across different low-rank adaptation strategies.

A.2 Backbone Models

We evaluate the proposed method using following multimodal vision-language backbones:

- Gemma-4B
- Llama-3.2-11B-Vision-Instruct

Here, we adapted both the model using parameter-efficient fine-tuning by keeping the pre-trained backbone frozen.

A.3 LoRA Configuration

LoRA, GRIT, and Gated GRIT were applied to identical target modules to ensure consistent parameter budgets across all methods. Experiments were conducted using adaptation ranks of $r = 4$ and $r = 8$.

Here, we parameterize the low rank as:

$$\Delta W = BA$$

where $A \in \mathbb{R}^{r \times d}$ and $B \in \mathbb{R}^{d \times r}$ are trainable low-rank matrices and r denotes the adaptation rank.

Here, for the Gated GRIT, the low-rank update becomes:

$$\Delta W = B\Gamma A$$

where Γ represents the diagonal supergate matrix derived from the Fisher spectrum. LoRA adaptation was applied to the query, key, and value projection matrices (q_{proj}, k_{proj}, and v_{proj}) in both the language and vision towers. The same target modules were used for LoRA, GRIT, and Gated GRIT.

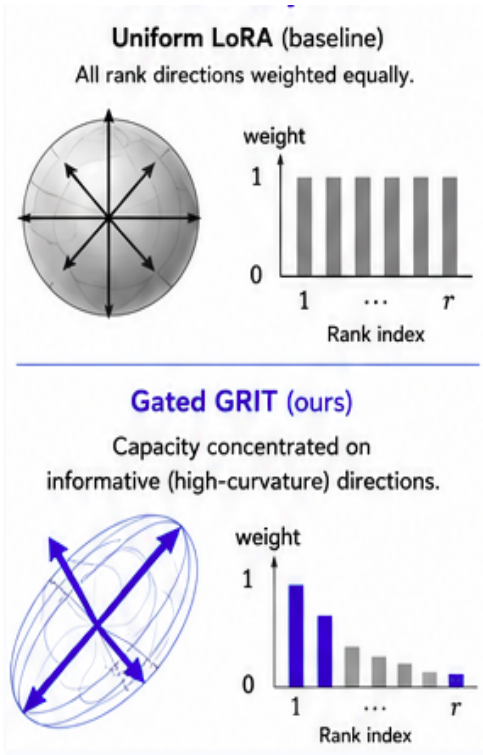

Figure 2: Diagramatic representation of our work's direction

A.4 GRIT Configuration

GRIT uses Fisher-guided natural-gradient preconditioning through a Kronecker-Factored Approximate Curvature (K-FAC) approximation of the Fisher Information Matrix (Martens and Grosse, 2020). The corresponding curvature statistics are computed directly in the low-rank adaptation subspace.

The GRIT framework includes the following:

- Fisher-guided low-rank preconditioning
- Estimation of K-FAC covariance
- Geometric reprojection of the low-rank directions
- Curvature-aware optimization within the LoRA subspace

Natural-gradient updates are computed using the approximated Fisher inverse obtained from the Kronecker-factored covariance matrices.

A.5 Configuration of supergate

Our proposed Gated GRIT method extends the GRIT through the rank-wise modulation of low-rank updates using our present curvature-derived gate values. The supergate values we found are

from the Fisher spectrum, and that is computed during our K-FAC estimation.

For each rank direction, a scalar gate value γ_i is assigned based on the relative importance of the corresponding Fisher eigendirection:

$$\Gamma = \mathrm{diag}(\gamma_1, \gamma_2, \ldots, \gamma_r)$$

The gating mechanism does not introduce additional trainable parameters and it operates entirely using curvature statistics already computed by GRIT. In order to prevent inactive rank directions, a minimum gate floor of 0.04 is applied before renormalization. This makes sure that all rank components remain trainable while still allowing dominant Fisher directions to receive higher allocation. As the Fisher spectrum changes during the training, a direction that appears to be unimportant can become important as the training progress. Pruning to zero will destroy the module. Hence, we give a low floor value as model might mistakenly prune the module to zero though potentially it may have been important as training would have progressed. Giving a floor values keeps the module alive.

Our proposed supergate mechanism includes:

- Rank allocation which is derived from Fisher spectrum.
- Rank-wise modulation of the low-rank updates
- Method to suppress the low-important and low-utility directions
- More priority given to dominant curvature-aligned directions
- Optimally using the adaptation capacity

A.6 Overview on the optimizations we implemented

All methods shared the same training configuration and parameter budget. Training was performed in mixed precision, with low-rank adaptation applied only to the selected transformer projection layers.

B Effective Rank Analysis

The effective rank provides a simple way to examine how the available rank directions are being used. While the nominal rank is fixed before training, the effective rank reflects the adaptation pattern that emerges during training.

Table 3: Hyperparameters we have used to train the GRIT and Gated GRIT.

Category	Hyperparameter	Value
LoRA	Rank (r)	4, 8
	LoRA Alpha	8
	LoRA Dropout	0.05
	Target Modules	$q_{proj}, k_{proj}, v_{proj}$
	Vision LoRA	Enabled
GRIT	K-FAC Damping	0.005
	K-FAC Update Frequency	150
	K-FAC Minimum Samples	256
	Reprojection Warmup	200
	Rank Adaptation Start	500
	Reprojection Frequency	300
	Two-Sided Reprojection	True
	Curvature Estimation	K-FAC
Supergate	Gate Source	Fisher Spectrum
	Supergate Floor	0.04
	Additional Parameters	None
Training	Batch Size	1
	Gradient Accumulation	4
	Epochs	3
	Learning Rate(all runs)	2×10^{-6}
	Precision	bf16
	Learning Rate(Gemma Gated GRIT)	8×10^{-6}
	Maximum Sequence Length	1024
	Optimizer	AdamW
	Warmup Steps	100

The effective rank is used here as a summary measure of rank utilization. It is computed from the entropy of the normalized Fisher eigenvalue spectrum:

$$r_{\text{eff}} = \exp\left(-\sum_i p_i \log p_i\right)$$

and

$$p_i = \frac{\lambda_i}{\sum_j \lambda_j}$$

and λ_i represents the Fisher eigenvalues, which are related to the given low rank module.

Table 4: Average effective rank observed for Gated GRIT.

Model	Nominal Rank	Effective Rank
Gemma-4B	4	2.50
Gemma-4B	8	3.59
Llama-11B	4	2.89
Llama-11B	8	4.05

For every configuration we evaluated, the effective rank remains below the nominal rank. By this, we can understand that the adaptation capacity is not distributed uniformly across the all rank directions. A small number of directions stand out with

noticeably larger gate values, whereas most remain comparatively small.

For Gemma-4B, the average effective rank increases from 2.50 to 3.59 as the nominal rank increases from 4 to 8. A similar trend is observed for Llama-3.2-11B-Vision, where the effective rank rises from 2.89 to 4.05. These observations indicate that the Gated GRIT concentrates the optimization capacity into a smaller number of informative directions while maintaining the flexibility of higher-rank adaptation.

Across all experiments, the effective rank stays below the nominal rank. The available rank directions are therefore not used equally during adaptation.

C Gate Statistics

In order to better understand the behavior of the proposed supergate mechanism, we do an analysis of the distribution of gate values across vision-language, vision-only, and language-only modules. For each of our configuration, we have reported the descriptive statistics, which includes stats like mean, standard deviation, minimum value, maximum value, and entropy.

Gate values are derived from the normalized Fisher spectrum and therefore, it reflects the relative importance assigned to individual low-rank directions during adaptation. Higher variance and lower entropy indicate quite a strong concentration of adaptation capacity into a subset of dominant directions. Unlike uniform rank utilization, our proposed gating mechanism here enables the selective emphasis of curvature-informed directions while preserving weak components through the gate floor. The gate statistics provide us a direct view of how the adaptation capacity is distributed across the low-rank subspace. Examining these distributions across different model architectures and adaptation ranks provides additional insight into how the supergate mechanism allocates adaptation capacity. The gate values are obtained from the normalized Fisher spectrum and reflect the relative importance of individual low-rank directions during adaptation. Higher variance and lower entropy here indicate a stronger concentration of adaptation capacity in a subset of dominant directions.

C.1 Vision-Language Modules

Here, the Table 5 gives us insights on supergate statistics which is present across both vision and language components of the model.

Table 5: We are representing here the supergate values for our vision and language modules when we consider them combined

Model	Mean	Std	Min	Max	Entropy
Gemma-R4	0.25	0.23857	0.04378	0.86864	6.17597
Gemma-R8	0.125	0.15796	0.03734	0.73855	6.73735
Llama-R4	0.25	0.22409	0.04589	0.86232	6.50457
Llama-R8	0.125	0.15878	0.04001	0.71519	7.02183

C.2 Vision Modules

The vision encoder gives us a distinct gate distribution as compared to the language tower. Table 6 reports the corresponding statistics.

Table 6: Supergate statistics for vision modules.

Model	Mean	Std	Min	Max	Entropy
Gemma-R4	0.25	0.18102	0.04916	0.71820	5.52421
Gemma-R8	0.125	0.12163	0.04536	0.61754	6.10561
Llama-R4	0.25	0.19768	0.04865	0.79988	5.87940
Llama-R8	0.125	0.14018	0.04311	0.65056	6.41737

C.3 Language Modules

Table 7 tells us on gate distributions, which is observed within the language tower. Since the answer generation is ultimately performed through the language model, these statistics provide us with insight into how adaptation capacity is being allocated during the multimodal reasoning.

Table 7: Supergate statistics for Gemma and Llama

Model	Mean	Std	Min	Max	Entropy
Gemma-R4	0.25	0.27608	0.04378	0.86864	5.46180
Gemma-R8	0.125	0.18178	0.03734	0.73855	6.00729
Llama-R4	0.25	0.24791	0.04589	0.86232	5.74345
Llama-R8	0.125	0.17550	0.04001	0.71519	6.24002

From the table, we observe that the gate distributions are non-uniform across all configurations, which indicates that the adaptation capacity is allocated to a subset of dominant rank directions on priority. This behavior is also consistent with the effective-rank analysis as we have discussed in the Appendix B.

In addition, the presence of low minimum gate values together with larger maximum values suggests to us that the proposed mechanism is suppressing weaker directions and preserving their contribution through the gate floor. So, we can say that the resulting distribution reflects the more structured allocation of adaptation capacity instead of uniform usage of all the rank components.

C.4 Discussion of the results from Appendix C

The gate distributions follow a similar pattern across all models and rank settings. Most directions receive small gate values, whereas only a handful stand out with larger ones.

Another consistent trend appears in the comparison between the vision and language towers. For both Gemma and Llama, the language tower contains higher peak gate values and larger variation than the vision tower. This means that a few language directions receive much stronger emphasis, while the vision tower distributes allocation more evenly across directions. Given that final answer generation takes place in the language model, this behavior is not surprising.

Third, the entropy also consistently increases when moving from rank 4 to rank 8. This increase in entropy indicates that higher-rank configurations are involved in activating a broader set of directions within the low-rank subspace. However, the simultaneous presence of large maximum gate values also demonstrates that allocation remains structured rather than uniform, even for the larger nominal ranks.

Finally, the combined vision-language statistics shows higher entropy than either modality considered separately. This suggests to us that multimodal adaptation requires the utilization of a broader set of directions rather than vision-only or language-only adaptation. Lastly, the non-uniform gate distributions indicate that the adaptation capacity remains concentrated more around a subset of Fisher-informed directions, which supports the central design principle of the proposed supergate mechanism.

D Supergate Distributions

Looking across the different models and rank settings, the gate distributions appear broadly similar. A small number of directions stand out with larger gate values, whereas most remain much smaller.

D.1 Llama-11B Rank 4

The distributions presented over here show us a pronounced long-tail structure, with the majority of the gate values concentrated near the gate floor and a relatively small subset receiving fairly larger allocations. This type of behavior shows that the adaptation capacity is more selectively concentrated into a limited number of Fisher informed directions.

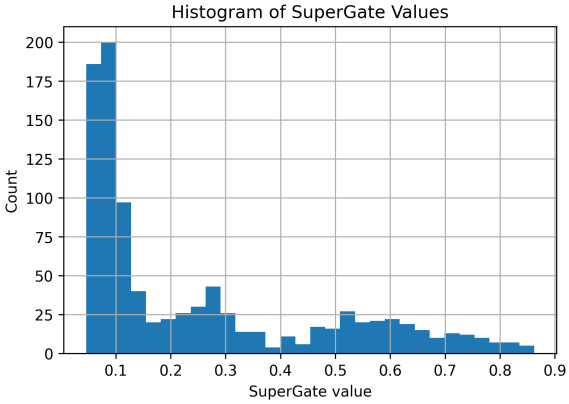

Figure 3: Overall distribution of the supergate values for Llama-4B at rank 4.

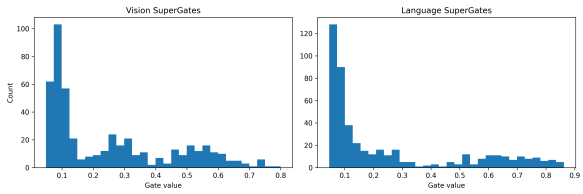

Figure 4: Plot of our supergate values across the vision and language modules for the Llama-11B at rank 4.

The comparison of vision and language modules also reveals that the language gates are generally attaining larger maximum values and showing a broader distribution. This phenomenon advises that the curvature-derived gating mechanism allocates greater adaptation capacity to language representations, which ultimately drives answer generation in multimodal visual question answering tasks.

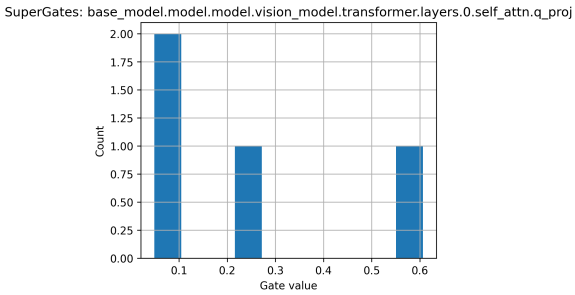

Figure 5: Plot of our supergate distribution for an attention projection module in Llama-11B at rank 4.

The attention projection module shows the same pattern seen in the overall gate distributions. Within a single low-rank block, only a few directions receive large gate values, while most remain close to the lower end of the range. As a result, adaptation is driven primarily by a small number of directions

rather than being distributed evenly across all of them.

D.2 Llama-11B Rank 8

The distribution of the supergate for the Llama-3.2-11B at rank 8 present a strong non-uniform allocation pattern. Most directions receive relatively small gate values, whereas a small subset stands out with much larger allocations. The resulting adaptation pattern is therefore far from uniform, with capacity concentrated in a limited set of Fisher-informed directions.

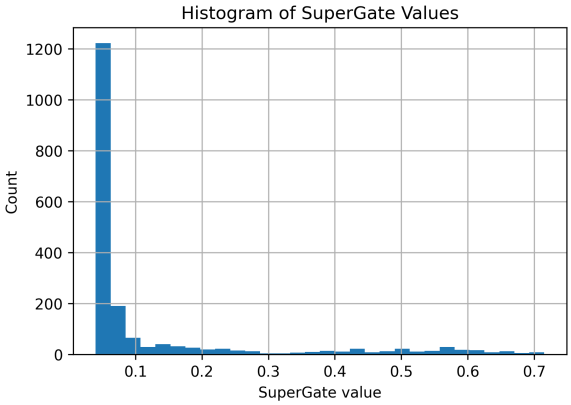

Figure 6: Overall distribution of supergate values for Llama-11B at rank 8.

The comparison of vision and language modules here shows that the language gates and the vision gates are mostly concentrated behind the 0.1 mark as we can we from the table. This phenomenon advises that the curvature-derived gating mechanism allocates greater adaptation capacity to language representations, which ultimately drives answer generation in multimodal visual question answering tasks. This suggests to us that most of the gates are close to the minimal allocation. This tells us that the adaptation capacity is concentrated in a small number of directions.

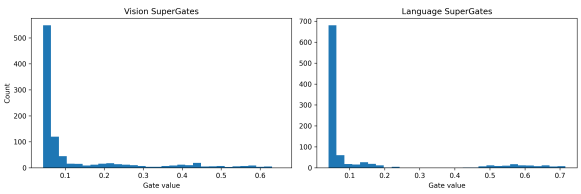

Figure 7: Distribution of supergate values across vision and language modules for Llama-11B at rank 8.

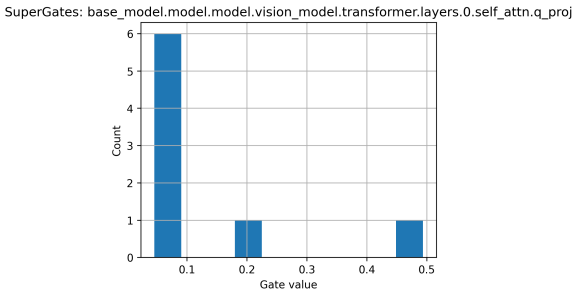

Figure 8: Supergate distribution for an attention projection module in Llama-11B at rank 8.

This type of the supergate distribution shows us the supergate values for a specific attention module. Here we can see that the most of the directions are present in large numbers near the minimum gate value and only a few of them receive legitimate larger values.

D.3 Gemma-4B Rank 4

For Gemma-4B at rank 4, most values are grouped near the lower end of the distribution. Only a few directions have larger values. This pattern appears throughout the figure, with a small number of directions standing out from the rest.

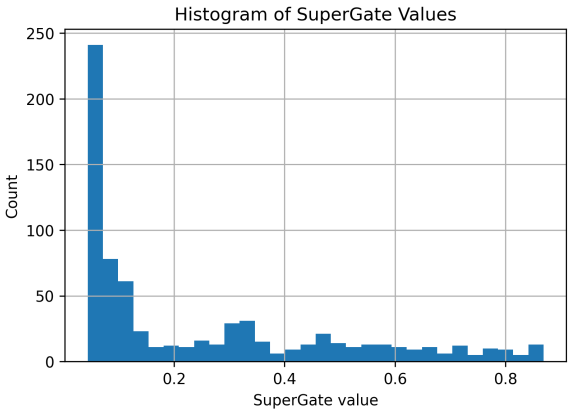

Figure 9: Total distribution of supergate values for gemma-4B at rank 4.

To better understand how adaptation capacity is allocated across modalities, we compare the gate distributions of the vision and language modules. The vision modules display a wider spread of gate values, suggesting that adaptation capacity is distributed across a larger number of directions. Language modules is displaying stronger concentration near the lower gate values, and together with a small number of highly emphasized directions. This recommends us that the allocation of adaptation capacity in gating mechanisms is different

across modalities, reflecting the differences in their curvature structure.

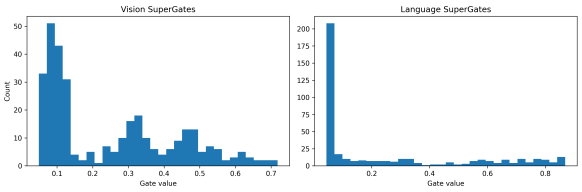

Figure 10: Distribution of supergate values across vision and language modules for gemma-4B at rank 4.

The representative attention projection module also shows us the rank specialization within the individual layers. Most rank directions remain almost near the minimum allocation level, with only a small number receiving significantly larger gate values. This type of behavior helps in support the central design principle of Gated GRIT, that only a subset of rank directions contribute meaningfully to adaptation and therefore warrant greater allocation of the available low-rank capacity.

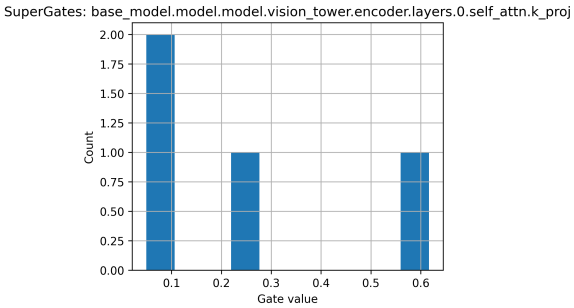

Figure 11: Plot for the supergate distribution for an attention projection module in Gemma-4B at the rank 4.

Overall, the gate distributions suggest that only a subset of directions receives strong emphasis, while the remaining directions are down-weighted rather than removed entirely.

D.4 Gemma-4B Rank 8

The supergate distributions for Gemma-4B at rank 8 show an even stronger concentration of adaptation capacity towards a small subset of dominant directions. The overall histogram shows that the majority of gate values remain clustered near the predefined minimum gate threshold, while only the limited number of directions receive a moderate or large allocations. Therefore, this type of behaviour shows us that even if we increase the nominal rank, it does not result in uniform usage of the additional rank components.

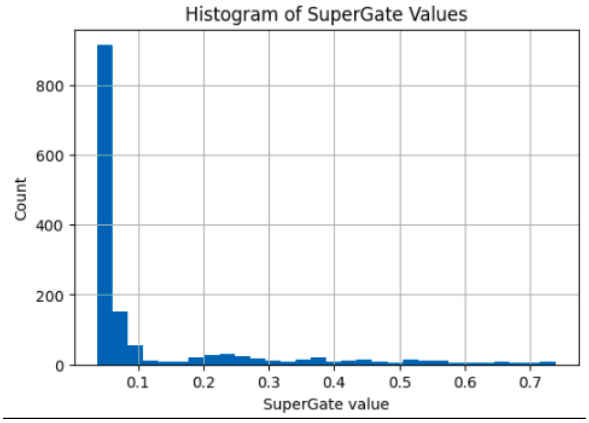

Figure 12: Overall distribution of supergate values for gemma-4B at rank 8.

The vision and language modules exhibit a similar trend: most directions remain near lower gate values, while only a few receive substantially larger allocations. This suggests that adaptation is driven primarily by a limited set of Fisher-informed directions rather than being distributed uniformly across all rank components.

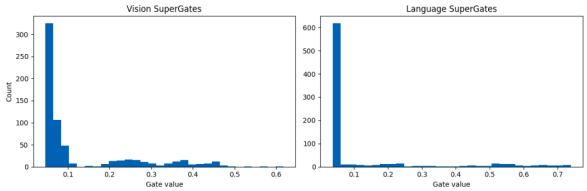

Figure 13: Plot on our distribution of supergate values across vision and language modules for Gemma-4B at rank 8.

The representative attention projection module also demonstrates a similar pattern at the layer level. In this case, most directions remain close to the gate floor, with only one or two receiving substantial emphasis. The resulting allocation pattern suggests that a relatively small number of directions accounts for most of the adaptation activity.

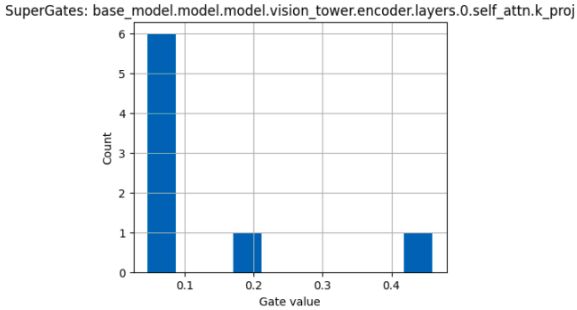

Figure 14: Plot for the supergate distribution for an attention projection module in Gemma-4B at rank 8.

So, increasing the nominal rank does not mean that all available directions are used equally. Instead, Gated GRIT continues to concentrate adaptation capacity in a relatively small set of informative curvature-aligned directions.

E Entropy Stats

Gate entropy helps us see whether the values are spread out or concentrated in only a few directions. Higher values correspond to a broader spread of activity, while lower values indicate that a smaller number of directions dominates the adaptation process.

In order to better understand the behavior of the proposed supergate mechanism, we analyze gate entropy as a function of layer depth for different backbone architectures taken into consideration for our work, along with the adaptation ranks. This analysis provides insight into whether the curvature-derived rank allocation remains consistent across layers or it varies meaningfully throughout the model.

E.1 Gemma-4B Rank 4

The entropy curve for Gemma-4B at rank 4 varies across the network rather than remaining at a constant level. A few layers stand out with noticeably lower entropy, whereas several intermediate layers exhibit higher values. This contrast suggests that some parts of the network rely on a smaller set of directions, while others make use of a broader portion of the available rank space. Overall, the degree of specialization changes across network depth.

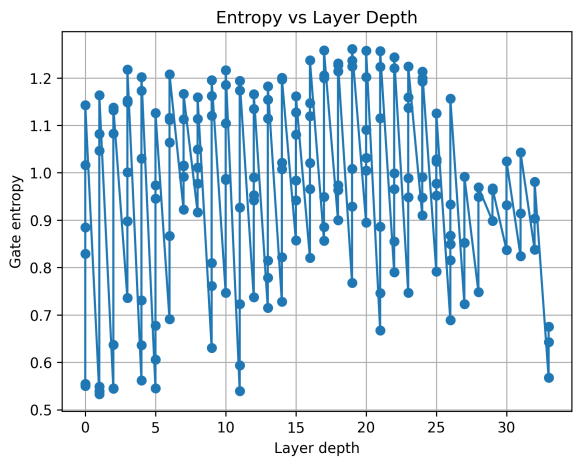

Figure 15: Plot of the entropy with layer depth in Gemma-4B at rank 4.

E.2 Gemma-4B Rank 8

Entropy values are generally higher for Gemma-4B at rank 8 than at rank 4. As a result, adaptation is distributed across a larger set of active directions rather than being concentrated in only a few. Local fluctuations are still visible, but their magnitude is noticeably smaller. This points to a more balanced use of the available low-rank subspace. Also, we claim that the entropy decreases in the deeper layers of the network, which indicates an increasingly selective allocation of adaptation capacity towards the end of the model. Overall, the results suggest that increasing the nominal rank also expands the diversity of active directions while preserving the structured allocation behavior induced by the proposed supergate mechanism.

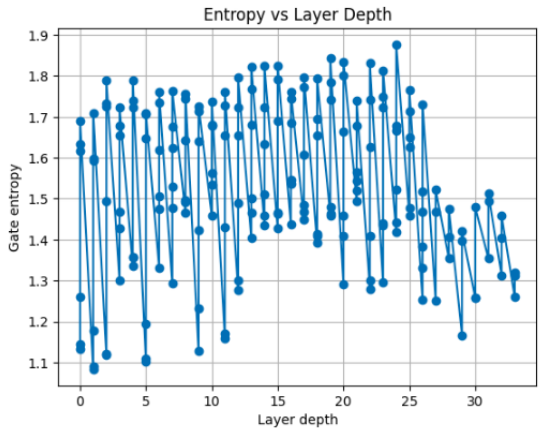

Figure 16: Plot of entropy against layer depth in Gemma-4B at rank 8.

E.3 Llama-11B Rank 4

For Llama-3.2-11B at rank 4, the curve remains relatively stable across the network. A few layers deviate from the surrounding trend, but these changes are brief and do not persist over many layers. Outside of these local fluctuations, the behavior remains largely unchanged across depth.

E.4 Llama-11B Rank 8

For Llama-3.2-11B at rank 8, the curve follows a shape similar to that observed at rank 4. Many of the peaks and dips occur in comparable regions of the network, although the values remain higher overall. Toward the final layers, a small decline becomes visible before the curve levels off again.

The entropy statistics that we have computed tells us the consistent trend across the both backbone architectures. In Rank-8 configurations, they

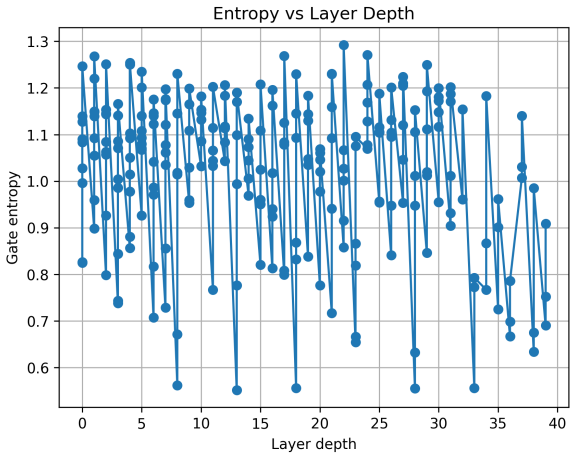

Figure 17: Plot of entropy against layer depth in Llama-11B at rank 4.

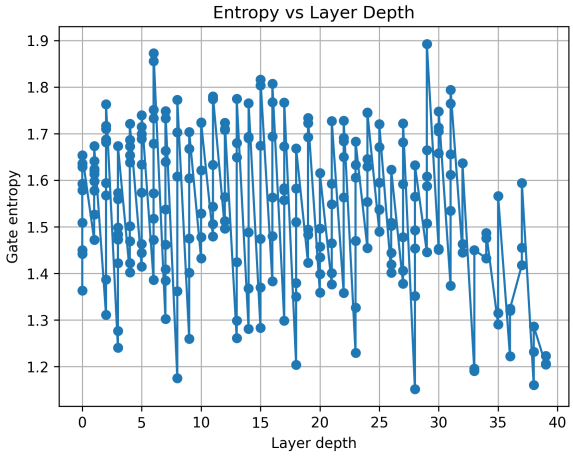

Figure 18: Plot of entropy against layer depth in Llama-11B at rank 8.

operate within a higher entropy regime than their rank-4 counterparts, showing broader usage of available low-rank directions. Even with this increase in the entropy magnitude, the overall entropy ranges remain comparable across models, which suggests that an increasing nominal rank primarily affects the diversity of active directions instead of fundamentally altering the layer-wise allocation behavior induced by the proposed supergate mechanism.

F Compression Ratio and Sparsity stats

To quantify more and show the impact of the proposed supergate mechanism, we are analyzing effective rank utilization, compression ratio, and sparsity statistics across backbone architectures considered and adaptation ranks. These metrics provide an insight into how efficiently the available low-rank capacity is used after curvature-guided rank allocation. In every configuration, the effective rank remains noticeably lower than the nominal rank, suggesting that adaptation relies primarily on a smaller set of active directions.

Table 8: Stats of gate entropy across model architectures and adaptation ranks.

Configuration	Min Entropy	Max Entropy	Range
Gemma R4	0.54	1.26	0.72
Gemma R8	1.08	1.89	0.81
Llama R4	0.55	1.29	0.74
Llama R8	1.16	1.89	0.73

Table 9: Effective-rank utilization statistics for Gated GRIT.

Metric	Gemma R4	Gemma R8	Llama R4	Llama R8
Avg Effective Rank	2.497	3.590	2.888	4.046
Compression Ratio	0.312	0.449	0.361	0.506
Sparsity (%)	0.41	0.48	0.31	0.31

The compression ratios we managed to compute also increase with the nominal rank, which we have defined earlier for both the Gemma and Llama backbones. This behavior suggests that additional rank capacity introduces more active directions, although utilization remains selective instead of uniform. Notably, even at rank 8, the effective rank remains considerably below the maximum available rank, which indicates continued suppression of low-utility directions.

Despite the relatively low sparsity values, the same trend remains visible: a few directions stand out, while many others play a smaller role.

When we take all the discussions together, we see the results suggest that only a subset of the available rank directions contributes strongly during adaptation, and the remaining directions play a comparatively smaller role.

G Superweight stats for Gemma-4B model

G.1 Gemma-4B Rank-4 and Rank-8

The vision and language projections follow different patterns. Entropy is generally higher in the vision modules, while the language modules contain a few larger gate values. Most directions in the language tower remain near lower gate values, with only a small number receiving substantially larger allocations.

Table 10: Summary statistics of curvature-derived superweights in GRIT-Gemma Rank-4. Here we claim that the language projections show lower entropy and higher peak gate values, and this indicates stronger concentration of adaptation capacity than vision projections.

Component	Entropy Range	Std Range	Max Gate Range
Vision Q-Proj	1.05–1.25	0.15–0.23	0.43–0.58
Vision K-Proj	1.01–1.25	0.15–0.26	0.41–0.62
Vision V-Proj	0.92–1.24	0.15–0.30	0.41–0.69
Language Q-Proj	0.55–1.01	0.23–0.41	0.48–0.86
Language K-Proj	0.54–1.07	0.23–0.41	0.50–0.87
Language V-Proj	0.66–1.02	0.23–0.38	0.46–0.82

Table 11: Summary statistics of curvature-derived superweights in Gated GRIT-Gemma Rank-8. Compared to Rank-4, entropy values are consistently higher, which is indicating broader utilization of available low-rank directions.

Component	Entropy Range	Std Range	Max Gate Range
Vision Q-Proj	1.38–1.84	0.10–0.20	0.30–0.62
Vision K-Proj	1.47–1.88	0.09–0.18	0.30–0.57
Vision V-Proj	1.59–1.83	0.10–0.15	0.30–0.49
Language Q-Proj	1.08–1.56	0.15–0.25	0.39–0.74
Language K-Proj	1.08–1.63	0.14–0.25	0.42–0.74
Language V-Proj	1.13–1.55	0.15–0.24	0.43–0.72

Compared to the rank-4 configuration, Gated GRIT at rank 8 operates within a substantially higher entropy regime across both vision and language modules. The increase in entropy indicates that adaptation capacity is distributed across a larger number of active directions rather than being concentrated into a small subset of dominant components.

Despite this broader utilization, language projections continue to exhibit lower entropy and larger peak gate values than vision projections. This suggests that the language backbone remains more selective and contains stronger dominant directions even when additional rank capacity is available.

Overall, the results indicate that increasing the nominal adaptation rank expands the diversity of active directions while preserving the structured allocation behavior induced by the proposed supergate mechanism.