SPECTRAGAIN: CARRIER-CONDITIONED CIRCULAR ADAPTATION VIA LOW-RANK SPECTRAL GAINS

Anonymous authors
Paper under double-blind review

ABSTRACT

Directly factorizing a weight correction ties its rank to the trainable-parameter budget. We introduce **Carrier-Conditioned Spectral Gain Adaptation (SpectraGain)**, which instead factorizes a real-valued spectral gain to control circular mixing of frozen weights. This parameterization allows the realized correction rank to exceed the generator rank while preserving the pretrained model at initialization. We derive its separable-kernel structure and evaluate it across language and vision tasks. SpectraGain achieves competitive performance, including a 1.01-percentage-point improvement over Hadamard High-Rank Adaptation (HiRA) in a near-parameter-matched Commonsense-15K comparison. Same-generator coupling replacements reveal task-dependent differences between circular and pointwise adaptation associated with locally accessible gradient directions. Fixed-checkpoint interventions show accuracy recovering as severe correction-rank truncations are relaxed. A rematerialized implementation supports long-context training, with memory and runtime costs evaluated against standard and fused low-rank adaptation. These findings characterize carrier-conditioned circular mixing as a competitive approach to compact weight adaptation.

1 INTRODUCTION

Adapting large pre-trained models (Vaswani et al., 2017; Devlin et al., 2019; Brown et al., 2020) increasingly relies on parameter-efficient fine-tuning (PEFT). A compact parameterization determines both the number of trained variables and the weight corrections they can realize. Standard Low-Rank Adaptation (LoRA) (Hu et al., 2022) computes $h = (W + BA)x$, coupling parameter cost and correction rank through $\mathrm{rank}(BA) \leq r$. This constraint applies to the learned residual: $W + BA$ may be full rank and BA may be entrywise dense. LoRA variants (Kopiczko et al., 2024; Zhang et al., 2023; Meng et al., 2024; Valipour et al., 2023; Zhang et al., 2025b) retain this correction form, while FourierFT (Gao et al., 2024) and the Layerwise Fourier Masked Adapter (LFMA) (Park et al., 2026) restrict the residual to selected spectral coordinates. Under a limited trainable-parameter budget, the choice of correction geometry is therefore consequential.

Higher-rank LoRA improves commonsense reasoning in the experiments of Huang et al. (2025), motivating high-rank adaptation with compact generators. Their Hadamard High-Rank Adaptation (HiRA) uses a low-rank spatial mask to modulate frozen weights pointwise. Circular Convolution Adaptation (C3A) (Chen et al., 2025) constructs a compact additive circular correction. High-rank realizability leaves open how a compact generator couples the frozen weights to form a correction. Pointwise modulation rescales individual weight entries, whereas circular mixing lets each correction entry combine shifted weight entries through a shared kernel.

Carrier-Conditioned Spectral Gain Adaptation (SpectraGain) implements this circular mixing by learning a compact real multiplier, or *spectral gain*, on the Fourier coefficients of the frozen weight W, which we call the *carrier*. For an admissible full-spectrum gain G, the Convolution Theorem gives $\Delta W = W \circledast \mathcal{F}^{-1}(G)$: each correction entry mixes shifted carrier entries with wraparound at the matrix boundaries. This parameterization lets us study how carrier-conditioned circular coupling organizes high-rank corrections under a compact trainable budget (Figure 1); Table 1 compares the closest update families.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

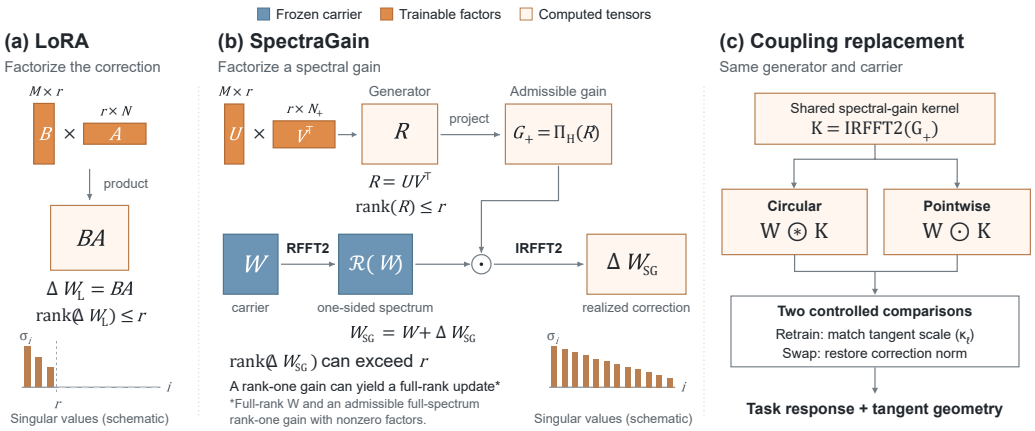


Figure 1: **SpectraGain and same-generator coupling replacement.** **(a)** Low-Rank Adaptation (LoRA) factorizes the weight correction. **(b)** Low-rank factors generate an admissible spectral gain that modulates the frozen carrier. **(c)** The same kernel and frozen carrier enter circular and pointwise branches, with tangent-scale matching for retraining or correction-norm restoration for checkpoint swaps. Singular-value bars are schematic.

Across language and vision tasks, SpectraGain achieves competitive performance, including a 1.01-percentage-point improvement over HiRA in the near-parameter-matched Commonsense-15K comparison. Same-generator replacements reveal task-dependent differences between circular and pointwise coupling associated with initialization-time accessible gradient directions. With a shared parameterization and budget, retraining optimizes each branch separately, whereas checkpoint swaps retain the learned factors (Figure 1(c)). Fixed-checkpoint rank interventions show accuracy recovering as severe truncations are relaxed, probing the contribution of retained correction directions (Figure 3).

The learned correction merges into the base weight for ordinary dense-layer inference. During training, a fused operator discards intermediate fast Fourier transform (FFT) states after the forward pass and rematerializes them during backward. Both fused LoRA and SpectraGain complete the profiled 120K-token setting, where standard LoRA runs out of memory. The matched fused-LoRA control quantifies the additional memory and compute required by the spectral path.

Our main contributions are summarized as follows:

(1) We introduce compact real spectral gains for carrier-conditioned circular adaptation, derive their separable-kernel structure and high-rank realizability, and preserve the frozen model exactly at initialization.

(2) We characterize task-dependent coupling responses through same-generator replacements and gradient-alignment diagnostics, and examine the role of retained correction directions through fixed-checkpoint rank interventions.

(3) We establish competitive language and vision performance and profile a rematerialized implementation through 120K tokens, quantifying its memory and runtime costs against standard and fused LoRA.

2 Related Work

Spatial Parameter-Efficient Fine-Tuning. Early alternatives to full fine-tuning (FT) injected auxiliary adapters (Houlsby et al., 2019) or prefix vectors (Li & Liang, 2021). Following LoRA (Hu et al., 2022), work has improved rank and parameter allocation (Zhang et al., 2023; he et al., 2025; Ye et al., 2026), initialization (Meng et al., 2024; Wang et al., 2024), optimization (Hayou et al., 2024; Wang et al., 2025; Tastan et al., 2026; Wu et al., 2026b), and parameter sharing (Kopiczko et al., 2024). Other methods change the update form: DoRA decomposes magnitude and direction (Liu et al., 2024), qGOFT uses quasi-Givens transformations (Ma et al., 2024), and Efficient Orthogonal Fine-Tuning with Principal Subspace Adaptation (PSOFT) combines principal-subspace orthogonal

Table 1: Carrier use and coordinate coupling in the closest compact weight-update operators.

Method	Compact object	Realized correction	Weight-coordinate coupling
LoRA (Hu et al., 2022)	A, B	BA	Additive low-rank
FourierFT (Gao et al., 2024)	Selected coefficients	$\alpha \operatorname{Re} \mathcal{F}^{-1}(S_\Omega)$	Additive Fourier modes
AdaDHP (Liu et al., 2025)	$\ell_{\mathrm{L}}, \ell_{\mathrm{R}}$	$W \odot (\ell_{\mathrm{L}} \ell_{\mathrm{R}}^\top - \mathbf{1})$	Carrier-conditioned pointwise
HiRA (Huang et al., 2025)	A, B	$W \odot (AB)$	Carrier-conditioned pointwise
C3A (Chen et al., 2025)	Circular kernel(s)	$C_{\mathrm{blk}}(w)$	Additive circular
SpectraGain	U, V	$\mathcal{R}^{-1}(\mathcal{R}(W) \odot \Pi_{\mathrm{H}}(UV^\top))$	Carrier-conditioned circular

S_Ω contains trainable coefficients on fixed sampled spectral entries Ω and zeros elsewhere; α is a fixed scale.

transformations with scaling vectors (Wu et al., 2026a). GaLore instead projects optimization gradients (Zhao et al., 2024).

AdaDHP (Liu et al., 2025) combines adaptive selection with dual-Hadamard row/column scaling of frozen weights; one vector pair yields the carrier-pointwise form in Table 1. HiRA (Huang et al., 2025) uses the low-rank spatial mask AB to form a potentially high-rank correction $W \odot (AB)$. We use it as the primary pointwise control because it is compared under the same targets and near-matched budgets. C3A (Chen et al., 2025) learns an additive circular correction.

Spectral Feature Modulation. Fourier coordinates provide cross-coordinate mixing: changing one coefficient affects the complete array after the inverse transform. FNet (Lee-Thorp et al., 2022) and GFNet (Rao et al., 2021) exploit this property on intermediate activations. SpectraGain instead learns a gain on the frozen weight spectrum, inducing circular mixing in weight coordinates.

Spectral Weight Adaptation. FourierFT (Gao et al., 2024) learns selected Fourier residual coefficients: its inverse transform directly supplies an additive correction. SpectraGain instead multiplies the frozen weight spectrum by a generated gain before inversion (Table 1). MaCP (Shen et al., 2025) selects a projected low-rank change within hierarchical cosine partitions; LFMA (Park et al., 2026) chooses Top-K components per layer by magnitude. Outside direct weight adaptation, F-Adapter (Zhang et al., 2025a) allocates adapter capacity across frequency bands in Fourier-layer large operator models.

3 Methodology

3.1 Low-Rank Spectral Gains

SpectraGain learns a real gain on the frozen weight spectrum. It multiplies that spectrum by $\mathbf{1} + \Pi_{\mathrm{H}}(UV^\top)$ and applies the inverse transform to obtain the adapted weight; the projection Π_{H} enforces the boundary symmetry required for a real inverse. For $W \in \mathbb{R}^{M \times N}$ and $N_+ = \lfloor N/2 \rfloor + 1$, let $\mathcal{R} : \mathbb{R}^{M \times N} \to \mathbb{C}^{M \times N_+}$ denote the one-sided real FFT and $\mathcal{R}_{M,N}^{-1}$ its length-aware inverse. Then

$$W_{\mathrm{SG}} = \mathcal{R}_{M,N}^{-1}(\mathcal{R}(W) \odot H_+), \qquad H_+ = \mathbf{1} + \Pi_{\mathrm{H}}(UV^\top), \tag{1}$$

where $U \in \mathbb{R}^{M \times r}$ and $V \in \mathbb{R}^{N_+ \times r}$. On the self-conjugate zero-frequency, or direct-current (DC), column and, for even N, the Nyquist column, Π_{H} averages entries paired by $u \leftrightarrow (-u) \bmod M$; it leaves interior columns unchanged and has no trainable parameters. Thus H_+ has a Hermitian full-spectrum completion and Eq. 1 is real by construction. Its signed real entries rescale components, with negative values also inducing a π phase flip. The resulting circular kernel is jointly centrosymmetric, $K_{i,j} = K_{(-i) \bmod M, (-j) \bmod N}$. This zero-or-$\pi$-phase bias defines SpectraGain's real-gain hypothesis class; Appendix A gives the full expressivity conditions.

Optimizing an unrestricted $M \times N_+$ gain is prohibitive. SpectraGain instead parameterizes its residual gain with two low-rank factors:

$$R = UV^\top, \qquad G_+ = \Pi_{\mathrm{H}}(R), \qquad H_+ = \mathbf{1} + G_+. \tag{2}$$

This uses exactly $r(M + N_+)$ trainable parameters. Because the real fast Fourier transform stores only the N_+ nonredundant frequency columns, this is smaller than LoRA's $r(M + N)$ count at the same rank; for approximately square weights, the reduction approaches 25%. Substituting the residual gain into Eq. 1 gives

$$\Delta W = \mathcal{R}_{M,N}^{-1}\big(\mathcal{R}(W) \odot \Pi_{\mathrm{H}}(UV^\top)\big), \qquad W_{\mathrm{SG}} = W + \Delta W. \tag{3}$$

The trainable-parameter count is $\mathcal{O}(dr)$ for dimensions of order d, but the factor rank does not bound the rank of ΔW. To see the distinction, consider the full-spectrum form $\widehat{W} = F_M W F_N^*$ and an admissible rank-one residual gain uv^* with nonzero entries. Then

$$\widehat{\Delta W} = \widehat{W} \odot (uv^*) = \mathrm{diag}(u)\widehat{W}\,\mathrm{diag}(\overline{v}), \qquad \mathrm{rank}(\Delta W) = \mathrm{rank}(W). \tag{4}$$

The equality follows because the diagonal factors and discrete Fourier transform (DFT) matrices are invertible. A rank-one *gain generator* can therefore induce a full-rank correction when W is full rank. More generally, an admissible rank-s full-spectrum gain induces a sum of s separable circular kernels acting on the frozen carrier. Appendix A distinguishes the raw factor rank, which controls parameter cost, from completed-gain rank and realized-correction rank, and derives this weight-domain form.

Coupling Topology. Let $K = \mathcal{R}_{M,N}^{-1}(G_+)$. The SpectraGain residual is $\Delta W_{\mathrm{circ}} = W \circledast K$, which lets each output coordinate aggregate circularly shifted carrier entries. A same-generator pointwise replacement is

$$\Delta W_{\mathrm{point}} = \kappa_\ell (W \odot K), \tag{5}$$

where κ_ℓ is a fixed layer scale matching active-factor Jacobian Frobenius norms at the shared identity start. Both branches use the same U, V factorization, initialization, gain construction, carrier, and parameter count. This pointwise replacement retains K, unlike HiRA's spatial mask AB. Section 4.5 evaluates independent retraining and evaluation-only swaps of selected SpectraGain checkpoints.

The fused implementation rematerializes weight-sized spectral tensors in backward and removes a separate adapter branch; the base linear layer retains its sequence-dependent activation cache. Appendix B states the saved-tensor and workspace costs, and Section 5 reports the resulting end-to-end crossover.

3.2 Identity-Preserving Initialization

We initialize U randomly and set $V = 0$, making the residual gain zero while allowing V to learn on the first step. Specifically, $U_{ij} \sim \mathcal{N}(0, 2/M)$ gives $UV^\top = 0$ and $H_+ = \mathbf{1}$ for every rank, so Eq. 1 preserves $W_{\mathrm{SG}} = W$ exactly. In weight coordinates, the all-ones full spectrum corresponds to the discrete Dirac delta (Oppenheim et al., 1999).

With random U and $V = 0$, $\nabla_V \mathcal{L}$ is generally nonzero at the first step, whereas $\nabla_U \mathcal{L} = 0$. Once V moves away from zero, $\nabla_U \mathcal{L}$ can also become nonzero. The variance scale follows the fan-in principle of He et al. (2015), while the zero second factor is the SpectraGain-specific identity constraint:

$$U_{ij} \sim \mathcal{N}(0, 2/M), \qquad V_{kj} = 0. \tag{6}$$

Appendix C compares initialization variants and their first-step gradients.

4 Experiments

We evaluate language and vision adaptation with Text-to-Text Transfer Transformer (T5)-base (Raffel et al., 2020), LLaMA-3.1-8B, and Contrastive Language–Image Pre-training (CLIP) ViT-B/16 (Radford et al., 2021), followed by targeted mechanism interventions. Source-reported baselines retain their original protocols. The added comparisons use near-matched trainable-parameter budgets, with ranks selected separately for each parameterization. For T5 and CLIP, a *controller-dev* holdout within the training split selects configurations; official validation splits provide the final reported results. Appendices D and E give the selection rules, complete configurations, and dataset definitions.

4.1 Natural Language Understanding Tasks

SpectraGain attains the highest observed parameter-efficient average, 87.62, in the controlled T5-base comparison (Table 2), with leading MNLI, QNLI, and MRPC results. We follow RaLoRA's five-task General Language Understanding Evaluation (GLUE) protocol (Ye et al., 2026; Wang et al., 2019): Multi-Genre Natural Language Inference (MNLI), Stanford Sentiment Treebank 2 (SST-2), the Corpus of Linguistic Acceptability (CoLA), Question Natural Language Inference (QNLI), and the

Microsoft Research Paraphrase Corpus (MRPC). Every column reports validation accuracy, including CoLA. Source-reported LoRA-family results accompany the parameter-matched FourierFT, HiRA, C3A-style, and SpectraGain rows. Under the one-sided parameter count above, uniformly rank-11 SpectraGain is approximately parameter matched to rank-8 LoRA across the adapted layers; this controlled setting also supports the mechanism interventions in Section 4.5.

Table 2: T5-base validation accuracy under the RaLoRA protocol. Rows from Full fine-tuning through RaLoRA-Pro are reported by Ye et al. (2026); FourierFT, HiRA, C3A-style, and SpectraGain are the added comparisons. Values are mean ± std over three seeds; averages are unweighted across tasks. Bold/underline mark best/second-best parameter-efficient fine-tuning (PEFT) results.

Method	MNLI	SST-2	CoLA	QNLI	MRPC	Average
Full fine-tuning	86.33±0.00	94.75±0.21	80.70±0.24	93.19±0.22	84.56±0.73	87.91
LoRA	85.30±0.04	94.04±0.11	69.35±0.05	92.96±0.09	68.38±0.01	82.01
MELoRA	85.57±0.19	93.62±0.30	77.28±0.36	89.59±5.24	84.11±1.45	86.03
MoRA	83.46±0.11	94.00±0.06	77.57±0.14	92.19±0.14	84.52±0.36	86.35
AdaLoRA	85.45±0.11	93.69±0.20	69.16±0.24	91.66±0.05	68.14±0.28	81.62
DoRA	85.67±0.09	94.04±0.53	72.04±0.94	93.04±0.06	68.08±0.51	82.57
RSLoRA	85.73±0.10	94.19±0.23	72.32±1.12	93.12±0.09	52.86±2.27	79.64
LoRA+	85.81±0.09	93.85±0.24	77.53±0.20	93.14±0.03	74.43±1.39	84.95
PiSSA	85.75±0.07	94.07±0.06	74.27±0.39	93.15±0.14	76.31±0.51	84.71
OLoRA	85.39±0.17	92.85±0.05	76.89±0.54	92.73±0.23	83.75±0.92	86.32
RaLoRA	85.76±0.03	94.22±0.29	78.11±0.45	93.36±0.14	84.74±0.27	87.24
RaLoRA-Pro	85.82±0.13	94.00±0.17	77.85±0.35	93.38±0.04	85.09±0.39	87.23
FourierFT (Gao et al., 2024)	86.14±0.09	<u>94.40</u>±0.24	**78.30**±0.46	93.29±0.10	85.04±0.35	87.43±0.05
HiRA (Huang et al., 2025)	86.15±0.09	**94.46**±0.26	<u>78.28</u>±0.39	<u>93.74</u>±0.08	85.32±0.32	<u>87.59</u>±0.07
C3A-style (Chen et al., 2025)	<u>86.16</u>±0.09	94.23±0.22	78.07±0.46	93.54±0.08	<u>85.34</u>±0.35	87.47±0.05
SpectraGain (Ours, $r = 11$**)**	**86.17**±0.08	94.32±0.22	78.21±0.41	**93.97**±0.10	**85.42**±0.33	**87.62**±0.08

MNLI: Multi-Genre Natural Language Inference; SST-2: Stanford Sentiment Treebank 2; CoLA: Corpus of Linguistic Acceptability; QNLI: Question Natural Language Inference; MRPC: Microsoft Research Paraphrase Corpus.

4.2 Reasoning and Nearest-Neighbor Comparison

SpectraGain improves over HiRA by 1.01 percentage points on Commonsense-15K, with a paired-seed 95% confidence interval of $[0.850, 1.171]$ (Table 3). This near-matched-budget comparison tests the closest spatial-multiplicative neighbor: rank-8 HiRA and uniformly rank-11 SpectraGain use 14.16M and 14.06M trainable parameters, respectively, over the same $q/k/v/up/down$ targets. Both follow PSOFT's LLaMA-3.1-8B protocol (Wu et al., 2026a), whose source-reported rows span low-rank, subspace, and orthogonal adapters on eight benchmarks. SpectraGain has the highest average, 82.69, and leads five tasks; LoRA-XS remains strongest on HellaSwag, while PSOFT leads ARC-Easy and OpenBookQA.

4.3 Instruction Following, Reasoning, and Code Generation

Table 4 covers multi-turn instruction following (MT-Bench), mathematical reasoning (GSM8K), and code generation (HumanEval). SpectraGain leads the eleven adapter baselines on MT-Bench and HumanEval and scores 0.21 points below RaLoRA-Pro on GSM8K (Cobbe et al., 2021). We follow RaLoRA's LLaMA-3.1-8B-Base training and evaluation protocols (Ye et al., 2026). SpectraGain uses rank 12 uniformly over W_q, W_k, W_v, W_o, with 7.079M parameters, within 3.87% of rank-8 LoRA's 6.816M. Appendix D.3 specifies the task-specific training corpora and evaluation procedure.

Table 3: LLaMA-3.1-8B accuracy under PSOFT's Commonsense-15K protocol. OFTv2 through PSOFT are reported by Wu et al. (2026a); Hadamard High-Rank Adaptation (HiRA) and SpectraGain are added at the rank-8 LoRA budget with the same $Q/K/V/U/D$ targets. Their averages are mean $\pm$ std over three seeds. Bold/underline mark best/second-best values.

Method	Params.	BoolQ	PIQA	SIQA	HellaS
OFTv2 ($b = 32$)	14.3M	70.83	84.44	73.34	90.63
LoRA ($r = 8$)	14.2M	<u>73.18</u>	85.31	74.36	86.57
PiSSA ($r = 8$)	14.2M	71.22	86.02	75.38	90.27
DoRA ($r = 8$)	14.9M	73.09	85.96	75.08	90.48
LoRA-XS ($r = 298$)	14.2M	72.35	<u>86.51</u>	75.18	**91.73**
PSOFT ($r = 424$)	14.5M	72.17	<u>86.51</u>	<u>75.79</u>	91.28
HiRA ($r = 8$)	14.16M	72.17	86.24	75.29	90.14
SpectraGain (Ours, $r = 11$)	14.06M	**73.28**	**86.61**	**75.88**	<u>91.55</u>

Method	WinoG	ARC-e	ARC-c	OBQA	Average
OFTv2 ($b = 32$)	74.11	90.87	80.12	81.80	80.77
LoRA ($r = 8$)	74.19	90.95	80.29	84.00	81.11
PiSSA ($r = 8$)	74.19	89.90	79.44	84.00	81.30
DoRA ($r = 8$)	75.53	90.74	81.40	84.40	82.09
LoRA-XS ($r = 298$)	74.98	90.74	79.52	84.00	81.88
PSOFT ($r = 424$)	<u>75.61</u>	**91.46**	<u>81.48</u>	**86.00**	<u>82.54</u>
HiRA ($r = 8$)	74.84	90.70	79.90	84.17	81.68$\pm$0.04
SpectraGain (Ours, $r = 11$)	**75.68**	<u>91.38</u>	**81.57**	<u>85.59</u>	**82.69**$\pm$0.04

PIQA: Physical Interaction Question Answering; SIQA: Social Interaction Question Answering; HellaS: HellaSwag; WinoG: WinoGrande; ARC-e/c: AI2 Reasoning Challenge easy/challenge sets; OBQA: OpenBookQA. M denotes millions of trainable parameters.

Table 4: LLaMA-3.1-8B-Base results under the RaLoRA protocol. All rows except SpectraGain are reported by Ye et al. (2026). Values are mean $\pm$ std over three seeds. Bold/underline mark best/second-best adapter results.

Method	MT-Bench ↑	GSM8K ↑	HumanEval ↑
Full fine-tuning	5.88 ± 0.23	73.69 ± 0.28	51.63 ± 1.27
LoRA	6.15 ± 0.02	67.78 ± 1.25	43.09 ± 0.35
MELoRA	6.11 ± 0.12	71.27 ± 1.1	45.73 ± 2.80
MoRA	6.38 ± 0.12	68.12 ± 0.91	29.68 ± 5.73
AdaLoRA	6.19 ± 0.16	70.63 ± 0.77	41.46 ± 3.66
DoRA	6.24 ± 0.12	69.17 ± 1.00	43.70 ± 1.54
RSLoRA	6.18 ± 0.09	68.36 ± 0.74	45.78 ± 2.80
LoRA+	6.35 ± 0.10	71.29 ± 0.93	44.51 ± 2.11
PiSSA	6.08 ± 0.09	68.56 ± 1.03	44.10 ± 1.54
OLoRA	6.13 ± 0.04	68.54 ± 0.42	43.29 ± 2.44
RaLoRA	6.38 ± 0.07	72.25 ± 0.59	<u>48.78 ± 1.61</u>
RaLoRA-Pro	<u>6.72 ± 0.04</u>	$\mathbf{73.01 \pm 0.53}$	48.37 ± 2.54
SpectraGain (Ours, $r = 12$)	$\mathbf{6.75 \pm 0.05}$	<u>72.80 ± 0.56</u>	$\mathbf{49.26 \pm 1.97}$

4.4 Image Classification Tasks

SpectraGain has the highest observed seven-dataset average, 90.88, while HiRA reaches 90.76 and remains best on EuroSAT and RESISC45 (Table 5). Under RaLoRA's CLIP-ViT-B/16 protocol (Ye et al., 2026), SpectraGain and HiRA share all-linear-layer adaptation and use 1.37M and 1.33M trainable parameters, respectively; across paired seeds, the SpectraGain-minus-HiRA average difference, computed from unrounded task averages, is $+0.11$ points (95% confidence interval $[0.057, 0.169]$).

Across the seven CLIP datasets, the SpectraGain–HiRA accuracy differences correlate with gradient alignment measured before adapter training (Spearman $\rho = 0.964$). This diagnostic measures weight-gradient energy accessible to SpectraGain and HiRA at initialization; the pointwise space here is HiRA's, distinct from the same-generator replacement in Section 4.5. The energy difference associates the observed performance differences with the operators' local geometry. Appendix D.4 defines the Jacobian spaces, dimension differences, and energy normalization.

Table 5: Contrastive Language–Image Pre-training (CLIP) ViT-B/16 validation accuracy under the RaLoRA protocol. Zero-shot through RaLoRA-Pro are reported by Ye et al. (2026); Hadamard High-Rank Adaptation (HiRA) and SpectraGain are added under the same protocol. Values are mean ± std over three seeds. Bold/underline mark best/second-best results.

Method	Rank	Cars	DTD	EuroSAT	GTSRB	RESISC45	SUN397	SVHN	Average
Zero-shot	–	63.75	44.39	42.22	35.22	56.46	62.56	15.53	45.73
LoRA	8	82.31±0.08	76.97±0.51	98.38±0.20	97.10±0.06	94.99±0.11	77.19±0.19	96.62±0.06	89.08±0.10
AdaLoRA	8	73.58±0.09	73.79±0.48	96.96±0.12	58.87±0.38	89.07±0.60	72.00±0.10	94.26±0.13	79.79±0.27
DoRA	8	82.44±0.26	76.86±0.84	98.43±0.17	97.25±0.12	95.10±0.16	77.30±0.17	96.63±0.04	89.14±0.07
LoRA+	8	86.61±0.36	73.33±1.30	98.54±0.14	98.99±0.20	96.06±0.38	76.80±0.34	96.98±0.08	89.62±0.19
PiSSA	8	83.36±0.38	77.38±0.57	98.54±0.09	98.32±0.09	95.92±0.40	77.46±0.13	97.00±0.09	89.71±0.25
RaLoRA	8	86.63±0.30	77.75±0.20	98.66±0.27	98.98±0.11	96.62±0.28	77.86±0.05	97.24±0.11	90.53±0.03
RaLoRA-Pro	8	86.40±0.51	**78.42**±0.67	98.88±0.30	99.23±0.17	96.66±0.09	<u>77.87</u>±0.27	97.15±0.08	90.66±0.11
HiRA	8	<u>86.87</u>±0.45	78.06±0.51	**99.07**±0.24	<u>99.30</u>±0.12	**97.03**±0.18	77.67±0.16	<u>97.35</u>±0.08	<u>90.76</u>±0.05
SpectraGain (Ours)	11	**86.92**±0.44	<u>78.07</u>±0.51	<u>98.97</u>±0.27	**99.62**±0.10	<u>96.89</u>±0.15	**78.05**±0.13	**97.62**±0.12	**90.88**±0.07

DTD: Describable Textures Dataset; GTSRB: German Traffic Sign Recognition Benchmark; RESISC45: Remote Sensing Image Scene Classification 45; SUN397: Scene Understanding 397; SVHN: Street View House Numbers.

4.5 Mechanism Interventions

We evaluate the adaptation value of compact high-rank corrections at three levels: complete-operator performance, the effect of coupling under a shared generator parameterization, and the contribution of directions retained in trained checkpoints.

Nearest-Operator Performance. Under the five-task T5-base protocol, SpectraGain reaches 87.62, versus 87.59 for carrier-pointwise HiRA-style (Huang et al., 2025) and 87.47 for carrier-independent, additive-circular C3A-style (Chen et al., 2025). All three use 192 targets, three paired seeds, and 3.244–3.348M trainable parameters. This compares complete operator families; the following replacement holds the generator parameterization fixed to examine coupling.

Same-Generator Coupling Replacement. Both branches share the U, V factorization, initialization, gain construction, carrier, targets, budget, paired seeds, and controller-dev selection budget; they compare $W \circledast K$ with Eq. 5. Norm-restored swaps test selected circular checkpoints: circular minus pointwise is $+0.143$ points on T5-base (paired 95% interval $[0.071, 0.215]$) and -0.094 on CLIP ($[-0.172, -0.015]$). Independent retraining lets each operator adapt its factors, giving $+0.107$ ($[-0.051, 0.266]$) and -0.058 ($[-0.211, 0.094]$). Its mean directions agree with the swaps, with both intervals spanning zero.

At initialization, small changes to the trainable factors span a set of possible weight-correction directions for each coupling operator. The diagnostic measures how much of the task gradient lies in each set; a positive contrast indicates greater gradient coverage by circular coupling. Figure 2 relates observed effects to this contrast a_t in gradient energy accessible to the two initialization-time tangent spaces. Within T5-base, Spearman correlations are $0.500/0.700$ for retraining/swaps; within CLIP they are $0.714/0.821$. The corresponding pooled values are $0.818/0.916$. Deleting one task gives pooled ranges $[0.773, 0.855]/[0.891, 0.945]$, while the smaller T5 retraining sample ranges from 0 to 0.8. Thus the diagnostic summarizes an observed geometric association; Appendix F gives its definition, stratified sensitivity, and per-setting results.

Fixed-Checkpoint Rank Intervention. Relaxing strict correction-rank truncation recovers task accuracy in the trained SpectraGain updates. After training and selecting each untruncated model once, we apply norm-restored best rank-k corrections only at evaluation (Figure 3). Mean accuracy rises from 87.36 at $2r$ to 87.54 at $4r$ and 87.70 at $8r$, compared with 87.62 for the full correction. This recovery supports the contribution of directions removed by stronger truncation. The paired $8r$-minus-full difference is $+0.082$ points (95% interval $[-0.453, 0.617]$). Appendix F specifies the intervention; Table F.2 reports all tasks and effective-rank statistics.

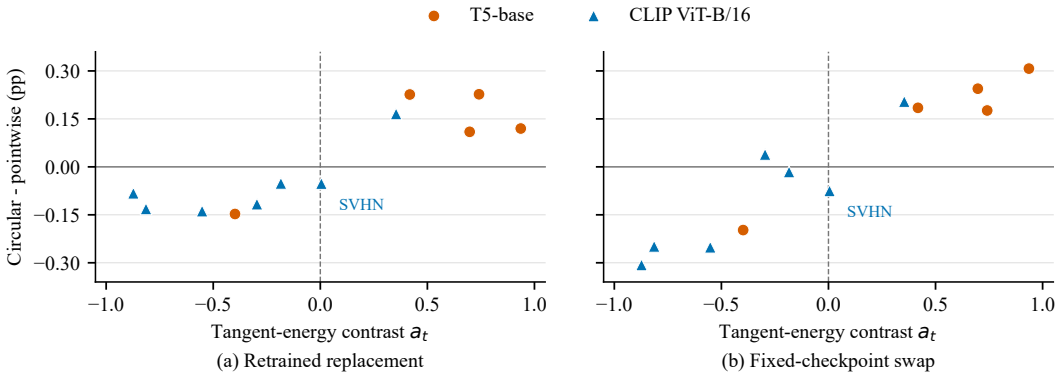


Figure 2: Same-generator coupling diagnostic across five T5-base and seven CLIP tasks. Each point pairs a_t with the three-seed mean circular-minus-pointwise accuracy difference (percentage points); both branches use the same spectral-kernel parameterization. **(a)** Independent retraining; **(b)** norm-restored checkpoint swaps. Zero lines separate the signs of the diagnostic and response.

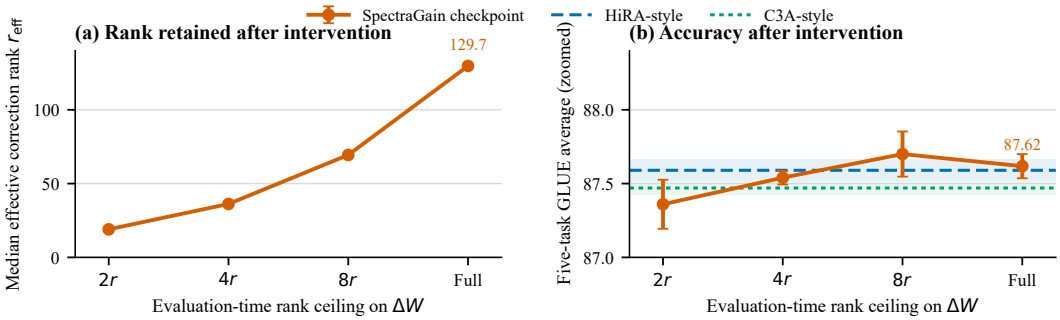


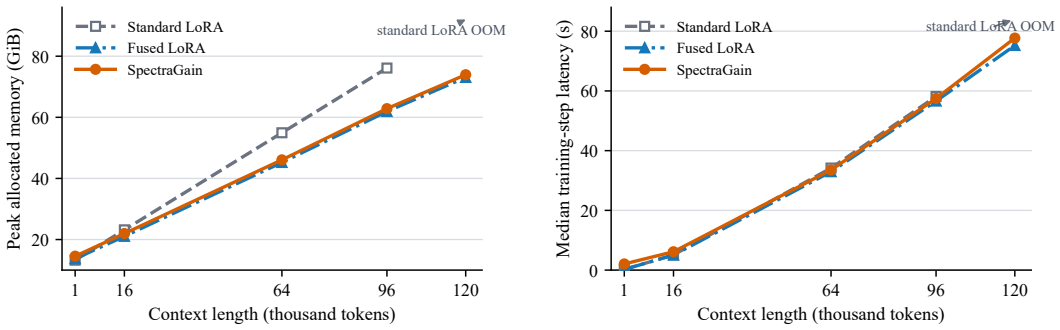
Figure 3: Evaluation-time rank intervention on fixed, uniformly rank-11 SpectraGain checkpoints. **(a)** Effective correction rank after norm-restored rank-k truncation. **(b)** Five-task General Language Understanding Evaluation (GLUE) average under the same intervention; horizontal lines mark separately trained HiRA-style and C3A-style controls. Error bars and bands show std over three paired seeds.

5 Long-Context Training Cost and Memory Profile

The long-context profile separates shared rematerialization savings from spectral-operator cost. Standard LoRA retains sequence-length-dependent adapter activations; fused LoRA and SpectraGain rematerialize that branch state. SpectraGain additionally requires fast Fourier transform (FFT), inverse fast Fourier transform (IFFT), and spectral workspace. Appendix B gives the implementation and complexity analysis.

We profile rank-one standard LoRA, fused LoRA, and SpectraGain on one 96-GiB RTX PRO 6000 Blackwell using bfloat16 (BF16), batch size one, a 512-token target segment, gradient checkpointing, and FlashAttention-2. LLaMA2-7B uses linear rotary position embedding (RoPE) scaling with factor 32 and a 131,072-token maximum position. Table B.2 specifies timing, allocator measurement, and out-of-memory (OOM) handling.

At 96K tokens, SpectraGain uses 0.94 GiB more allocated memory and takes 0.864 seconds longer per step than fused LoRA. Fusing LoRA saves 14.22 GiB relative to the standard branch, and both fused methods complete 120K, where standard LoRA runs out of memory. At 1K, the step times are 2.064 and 0.241 seconds for SpectraGain and fused LoRA: the weight-sized spectral work is amortized over more tokens as length increases. Figure 4 shows this length dependence, with exact values in Appendix Table B.3. This rank-one LLaMA2 profile characterizes the shared rematerialized setup; the accuracy comparisons use their own backbones and ranks.



(a) Peak allocated memory.

(b) Per-batch latency.

Figure 4: Long-context implementation attribution on LLaMA2-7B. The standard Low-Rank Adaptation curve is out of memory at 120K tokens.

6 Conclusion

SpectraGain enables high-rank adaptation by factorizing spectral gains and mixing circularly shifted frozen weights. Its compact generator yields a structured correction family with identity-preserving initialization and competitive language and vision performance. Fixed-checkpoint interventions show accuracy recovering as stricter correction-rank truncations are relaxed. Same-generator replacements expose task-dependent coupling effects, associated with locally accessible gradient directions. These structural and empirical findings, together with merged dense-layer inference and an explicit training-cost profile, establish spectral-gain modulation as a practical design for compact high-rank adaptation.

AI use statement

The authors led the research, including its conception, methodological design, and scientific interpretation. Generative AI tools assisted with literature research, language polishing, figure preparation, mathematical derivations, theoretical proofs, and preliminary manuscript feedback. The authors reviewed and verified the mathematical arguments, evaluated all AI-assisted suggestions, and made the final scientific and editorial decisions. The authors take responsibility for the final manuscript.

Reproducibility Statement

Section 3 and Appendices A–B.2 specify the operator, mathematical conditions, initialization, custom backward, and tensor-level complexity. Appendix D details target inventories, parameter budgets, optimization, datasets, controller-dev selection, and reporting splits; Appendix E lists the five T5 tasks and their metrics. Appendix F specifies the same-generator coupling replacement, tangent-space interpretation, nearest-operator controls, fixed-checkpoint rank intervention, and effective-rank aggregation.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Aochuan Chen, Jiashun Cheng, Zijing Liu, Ziqi Gao, Fugee Tsung, Yu Li, and Jia Li. Parameter-efficient fine-tuning via circular convolution. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 2004–2019, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.102. URL https://aclanthology.org/2025.findings-acl.102/.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Ziqi Gao, Qichao Wang, Aochuan Chen, Zijing Liu, Bingzhe Wu, Liang Chen, and Jia Li. Parameter-efficient fine-tuning with discrete Fourier transform. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 14884–14901. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/gao24o.html.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. LoRA+: Efficient low rank adaptation of large models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 17783–17806. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/hayou24a.html.

haonan he, Peng Ye, Yuchen Ren, yuan yuan, LuyangZhou, ShucunJu, and lei chen. Gora: Gradient-driven adaptive low rank adaptation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 114074–114102. Curran Associates, Inc., 2025. doi: 10.52202/085713-3805. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a5e4907a40c0dcb8433a35c714ba9d79-Paper-Conference.pdf.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/houlsby19a.html.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Qiushi Huang, Tom Ko, Zhan Zhuang, Lilian Tang, and Yu Zhang. HiRA: Parameter-efficient hadamard high-rank adaptation for large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=TwJrTz9cRS.

Dawid Jan Kopiczko, Tijmen Blankevoort, and Yuki M Asano. VeRA: Vector-based random matrix adaptation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NjNfLdxr3A.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontañón. FNet: Mixing tokens with Fourier transforms. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4296–4313, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.319. URL https://aclanthology.org/2022.naacl-main.319/.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353/.

Han Liu, Changya Li, Xiaotong Zhang, Feng Zhang, Fenglong Ma, Wei Wang, and Hong Yu. AdaDHP: Fine-grained fine-tuning via dual Hadamard product and adaptive parameter selection. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9492–9504, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.467. URL https://aclanthology.org/2025.acl-long.467/.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32100–32121. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/liu24bn.html.

Xinyu Ma, Xu Chu, Zhibang Yang, Yang Lin, Xin Gao, and Junfeng Zhao. Parameter efficient quasi-orthogonal fine-tuning via givens rotation. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 33686–33729. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/ma24a.html.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. Pissa: Principal singular values and singular vectors adaptation of large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 121038–121072. Curran Associates, Inc., 2024. doi: 10.52202/079017-3846. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/db36f4d603cc9e3a2a5e10b93e6428f2-Paper-Conference.pdf.

Alan V Oppenheim, Ronald W Schafer, and John R Buck. *Discrete-time signal processing*. Pearson Education, 1999.

Junyoung Park, Soo Yong Kim, Sang Heon Lee, and Jeonghwan Lee. Lfma: Parameter-efficient fine-tuning via layerwise fourier masked adapter with top-k frequency selection. In Francisco Acosta, Simone Azeglio, Bahareh Tolooshams, Chase van de Geijn, Christian Shewmake, Sophia Sanborn, and Nina Miolane (eds.), *Proceedings of the 4th (2025) and 3rd (2024) NeurIPS Workshops on Symmetry and Geometry in Neural Representations*, volume 282 of *Proceedings of Machine Learning Research*, pp. 415–427. PMLR, 14 Dec 2024–07 Dec 2025 2026. URL https://proceedings.mlr.press/v282/park26a.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume

139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 980–993. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/07e87c2f4fc7f7c96116d8e2a92790f5-Paper.pdf.

Yixian Shen, Qi Bi, Jia-hong Huang, Hongyi Zhu, Andy D. Pimentel, and Anuj Pathania. MaCP: Minimal yet mighty adaptation via hierarchical cosine projection. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20602–20618, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1006. URL https://aclanthology.org/2025.acl-long.1006/.

Nurbek Tastan, Stefanos Laskaridis, Martin Takáč, Karthik Nandakumar, and Samuel Horváth. LoFT: Low-rank adaptation that behaves like full fine-tuning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=86P3sb1dpr.

Mojtaba Valipour, Mehdi Rezagholizadeh, Ivan Kobyzev, and Ali Ghodsi. DyLoRA: Parameter-efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. In Andreas Vlachos and Isabelle Augenstein (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3274–3287, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.239. URL https://aclanthology.org/2023.eacl-main.239/.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJ4km2R5t7.

Shaowen Wang, Linxi Yu, and Jian Li. Lora-ga: Low-rank adaptation with gradient approximation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 54905–54931. Curran Associates, Inc., 2024. doi: 10.52202/079017-1741. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/62c4718cc334f6a0a62fb81c4a2095a1-Paper-Conference.pdf.

Zhengbo Wang, Jian Liang, Ran He, Zilei Wang, and Tieniu Tan. Lora-pro: Are low-rank adapters properly optimized? In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 93787–93808, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/ea184f920a0f0f8d8030aa1bd7ac9fd4-Paper-Conference.pdf.

Fei Wu, Jia Hu, Geyong Min, and Shiqiang Wang. Efficient orthogonal fine-tuning with principal subspace adaptation. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=FSHrinMArK.

Yize Wu, KE GAO, Ling Li, and Yanjun Wu. Stable-loRA: Stabilizing feature learning of low-rank adaptation. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=xSa19DAieH.

Jingqi Ye, Haonan He, Minglei Li, Fujun Han, Tao Chen, and Peng Ye. Gradient intrinsic dimensionality alignment:narrowing the gap between low-rank adaptation and full fine-tuning. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 55765–55790, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/5b4b967d4222d87fa5b28b6ec7144058-Paper-Conference.pdf.

Hangwei Zhang, Chun Kang, Yan Wang, and Difan Zou. F-adapter: Frequency-adaptive parameter-efficient fine-tuning in scientific machine learning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 111120–111162. Curran Associates, Inc., 2025a. doi: 10.52202/085713-3710. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a12e362d89d4e0b40760f839f91550ee-Paper-Conference.pdf.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=lq62uWRJjiY.

Yuanhe Zhang, Fanghui Liu, and Yudong Chen. LoRA-one: One-step full gradient could suffice for fine-tuning large language models, provably and efficiently. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 75513–75574. PMLR, 13–19 Jul 2025b. URL https://proceedings.mlr.press/v267/zhang25ax.html.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. GaLore: Memory-efficient LLM training by gradient low-rank projection. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 61121–61143. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/zhao24s.html.

A Update Expressivity and Dependency Analysis

A.1 A Low-Rank Gain Can Induce a Full-Rank Update

Carrier-Conditioned Spectral Gain Adaptation (SpectraGain) applies an admissible gain to a frozen spectral carrier. Let F_M and F_N be the invertible discrete Fourier transform (DFT) matrices and $\widehat{W} = F_M W F_N^*$. An admissible real gain has a Hermitian-symmetric full-spectrum completion. This set contains rank-one gains uv^* whose factors are nonzero and satisfy the required conjugate symmetry.

Proposition 1 (factor-rank separation). *If W has rank q and uv^* is an admissible rank-one residual gain with no zero factor entries, then the SpectraGain update induced by this gain has rank q.*

Proof. Using the Hadamard–diagonal identity,

$$\widehat{W} \odot (uv^*) = \mathrm{diag}(u)\widehat{W}\,\mathrm{diag}(\overline{v}). \tag{A.1}$$

Both diagonal matrices are invertible, so this product has rank $\mathrm{rank}(\widehat{W}) = \mathrm{rank}(W) = q$. Applying the inverse DFT on the left and right preserves rank. Thus the realized update may be full rank even though its gain generator has rank one. □

A concrete example is $\Delta W = cW$ for real $c \neq 0$: rank-one factors $U = c\mathbf{1}_M$ and $V = \mathbf{1}_{N_+}$ give an admissible constant gain. Representing this correction exactly with an additive BA requires rank at least $\mathrm{rank}(W)$. The construction demonstrates how carrier conditioning realizes high-rank

corrections with a compact generator; Section 4.5 evaluates learned corrections and their retained directions.

For the implemented one-sided generator, let $R = UV^\top$, let $G_+ = \Pi_{\mathrm{H}}(R)$ be the admissible residual gain, and let $\widetilde{G} = C(G_+)$ denote its Hermitian full-spectrum completion. The raw factorization guarantees $\mathrm{rank}(R) \le r$, but the projection and completion need not preserve that rank. Writing $s = \mathrm{rank}(\widetilde{G})$, the completed columns lie in the spans of U and its conjugate row-reversal, so $s \le 2r$. If $\widetilde{G} = \sum_{\ell=1}^{s} a_\ell b_\ell^*$, then

$$\widehat{\Delta W} = \sum_{\ell=1}^{s} \mathrm{diag}(a_\ell)\widehat{W}\,\mathrm{diag}(\overline{b}_\ell), \qquad \mathrm{rank}(\Delta W) \le \min\{M, N, sq\}. \tag{A.2}$$

The raw factor rank r, completed-gain rank s, and realized-correction rank are therefore distinct quantities. The special admissible construction in Proposition 1 has $s = 1$ and attains rank q; a generic one-sided rank-r generator guarantees $s \le 2r$. These statements characterize representational separation within SpectraGain's $\mathcal{O}(dr)$ family. Figure 3 tests whether SpectraGain uses directions removed by a strict correction-rank ceiling; Appendix F separately isolates circular from pointwise coupling under the same gain generator.

A.2 Spatial Pointwise and Spectral Circular Coupling

Hadamard High-Rank Adaptation (HiRA) (Huang et al., 2025) provides the closest structural counterpoint. With a low-rank spatial mask $M = AB$, its correction is

$$\Delta W_{\mathrm{HiRA}} = W \odot M, \qquad (\Delta W_{\mathrm{HiRA}})_{i,j} = W_{i,j} M_{i,j}. \tag{A.3}$$

For a fixed mask, each correction entry uses the co-located carrier entry. The Hadamard rank inequality nevertheless allows this correction to have rank greater than $\mathrm{rank}(M)$, so pointwise spatial multiplication already supplies the generator–correction rank separation.

SpectraGain applies the same multiplicative idea before the inverse Fourier transform. Let $K = \mathcal{F}^{-1}(\widetilde{G})$. The Convolution Theorem (Oppenheim et al., 1999) gives

$$\Delta W_{\mathrm{SG}} = W \circledast K, \qquad (\Delta W_{\mathrm{SG}})_{i,j} = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} W_{m,n} K_{(i-m) \bmod M,(j-n) \bmod N}. \tag{A.4}$$

Because the two-dimensional inverse DFT is an invertible left–right basis change, $\mathrm{rank}(K) = \mathrm{rank}(\widetilde{G}) = s \le 2r$. Writing $K = \sum_{\ell=1}^{s} p_\ell q_\ell^\top$ shows that SpectraGain learns a low-rank *sum* of separable circular kernels, not a single separable kernel unless $s = 1$. Moreover, the implemented gain is real. Its completed spectrum is therefore real and Hermitian, which implies the joint centrosymmetry

$$K_{i,j} = K_{(-i) \bmod M,(-j) \bmod N}. \tag{A.5}$$

This zero-or-π-phase restriction reduces parameters and enforces a symmetric circular mixing bias; it also excludes general phase-shifting kernels. Appendix F compares the resulting operator with pointwise and additive-circular controls, then isolates coupling topology with a same-generator replacement. Figure 3 measures the separate effect of truncating realized corrections under the controlled Text-to-Text Transfer Transformer (T5)-base setting.

B Implementation and Memory-Efficient Operator Design

Both a frozen linear layer and its adapter require sequence-dependent information to form parameter gradients. A conventional Low-Rank Adaptation (LoRA) branch additionally retains or rematerializes the low-dimensional adapter path, while a naive spectral implementation retains weight-sized complex fast Fourier transform (FFT) intermediates. The fused implementation is evaluated through end-to-end allocated and reserved memory, which incorporates framework-level storage reuse.

SpectraGain discards its spectral intermediates after forward, rematerializes them in backward, and uses the base linear path without a separate adapter activation. It still saves X (or relies on the enclosing checkpoint to rematerialize it) and allocates weight-sized FFT workspace during recomputation. Algorithm B.1 makes both facts explicit.

Algorithm B.1: Memory-efficient SpectraGain training with a custom backward pass.

Input: Input X, Frozen Weights W, Low-rank Factors U, V
Output: Output Y, Gradients $\nabla_X \mathcal{L}, \nabla_U \mathcal{L}, \nabla_V \mathcal{L}$

; // Two-dimensional real-input fast Fourier transform (RFFT2)
; // Two-dimensional inverse real-input fast Fourier transform (IRFFT2)
/* -- Forward Pass -- */
$W_{\mathcal{F}} \leftarrow \texttt{RFFT2}(W)$;
$R \leftarrow UV^\top$; $G_+ \leftarrow \Pi_{\mathrm{H}}(R)$; $H_+ \leftarrow \mathbf{1} + G_+$;
$W_{\mathrm{SG}} \leftarrow \texttt{IRFFT2}(W_{\mathcal{F}} \odot H_+;\, s = (M, N))$;
$Y \leftarrow X W_{\mathrm{SG}}^T$;
// Discard $W_{\mathcal{F}}$, R, G_+, H_+, and W_{SG}; cache $\{X, W, U, V, s\}$.
/* -- Backward Pass (Custom Autograd) -- */
Input: Upstream Gradient $\nabla_Y \mathcal{L}$
$G_W \leftarrow \mathrm{reduce}_{batch,token}((\nabla_Y \mathcal{L})^\top X)$;
$W_{\mathcal{F}}, R, G_+, H_+, W_{\mathrm{SG}} \leftarrow$ rematerialize forward weight path ;
$\nabla_X \mathcal{L} \leftarrow (\nabla_Y \mathcal{L}) W_{\mathrm{SG}}$;
$Z \leftarrow (\texttt{IRFFT2}_s)^*(G_W)$; // Exact real adjoint, including endpoint weights and normalization
$G_H \leftarrow \mathrm{Re}(\mathrm{Conj}(W_{\mathcal{F}}) \odot Z)$;
$G_R \leftarrow \Pi_{\mathrm{H}}^*(G_H)$; $\nabla_U \mathcal{L} \leftarrow G_R V$; $\nabla_V \mathcal{L} \leftarrow G_R^\top U$;
return $\nabla_X \mathcal{L}, \nabla_U \mathcal{L}, \nabla_V \mathcal{L}$

B.1 Computational and Memory Complexity Analysis

Graphics processing unit (GPU) memory during fine-tuning consists of *Model State Memory* (parameters and optimizer states) and *Activation Memory* (cached intermediate tensors).

Model State Memory. For one $M \times N$ target, LoRA stores $r(M + N)$ trainable factors and SpectraGain stores $r(M + N_+)$, where $N_+ = \lfloor N/2 \rfloor + 1$. Optimizer-state memory follows the same counts. The Hermitian admissibility projection adds no parameter.

For the 192 T5-base targets in Table D.1, rank-8 SpectraGain has 2,434,560 parameters versus 3,244,032 for rank-8 LoRA, a 24.95% reduction. The main comparison allocates this compactness to rank 11, giving 3,347,520 parameters. With four 32-bit arrays for factors, gradients, and Adam's two moments, the parameter-associated state is only 1.58 MiB larger than rank-8 LoRA. Activation storage and transient workspace follow the separate ledger below.

Saved Tensors, Workspace, and Compute. Let S be the number of input vectors seen by a target. Relative to a frozen linear layer that needs the input-gradient path, every row in Table B.1 retains $X \in \mathbb{R}^{S \times N}$ unless an enclosing checkpoint rematerializes it. Standard LoRA also leaves its parallel adapter branch to ordinary automatic differentiation, which may retain the $S \times r$ intermediate and framework-level matrix-multiplication operands. Fused LoRA discards that branch state and recomputes it in backward. SpectraGain likewise saves no spectral intermediate; it rematerializes the weight path and uses transient $\Theta(MN_+)$ complex workspace while forming the dense $G_W = (\nabla_Y \mathcal{L})^\top X$. The ledger identifies the adapter state that scales with S, while the end-to-end profile captures its effect together with checkpoint boundaries, saved operands, layout, and allocator reuse.

Table B.1 summarizes these distinct quantities. Table B.3 reports end-to-end allocated memory, and Appendix Table B.4 reports allocated and reserved differences from fused LoRA.

Table B.1: Per-target adapter state and rematerialization ledger relative to a frozen linear layer. S is the number of input vectors and $N_+ = \lfloor N/2 \rfloor + 1$.

Path	Persistent adapter state	Rematerialized / transient state	Adapter-path cost
Standard LoRA	X; branch autograd state up to $\Theta(Sr)$	none	$\Theta(Sr(M+N))$
Fused LoRA	X; no branch intermediate	$XA^\top \in \mathbb{R}^{S\times r}$	two adapter evaluations, each $\Theta(Sr(M+N))$
SpectraGain	X; no spectral intermediate	$W_{\mathcal{F}}, G_+, H_+, W_{\mathrm{SG}}$; $\Theta(MN_+)$ complex workspace	$\Theta(SMN + MN\log(MN) + MNr)$

B.2 Architectural Optimizations for Rematerialized Spectral State

Our `SpectraGainFusedOp` implementation bypasses standard automatic differentiation to control GPU memory through three mechanisms:

1. Operator Fusion and Gradient Rematerialization. We fuse the two-dimensional real-input fast Fourier transform (RFFT2), spectral modulation, and its inverse (IRFFT2) into one operator. In the forward pass, we discard fixed-size spectral intermediates after computing W_{SG}; in the backward pass, we rematerialize $W_{\mathcal{F}}$ on the fly. The layer then uses a single materialized weight rather than a sequence-dependent parallel adapter branch. The fused-LoRA control implements one custom branch operation that caches only X, A, B and rematerializes $XA^\top$ before evaluating the two factor gradients. SpectraGain and fused LoRA therefore share the same branch-rematerialization boundary; their remaining difference is the spectral weight path and its transform workspace.

2. Real-Valued FFT Acceleration and Tensor Contraction. Since W is real, SpectraGain uses `rfft2`/`irfft2` and explicitly projects the direct-current (DC) and Nyquist boundary columns before multiplication. The custom vector–Jacobian product (VJP) applies the exact adjoint of `irfft2`, accounting for endpoint multiplicities and normalization. We compare the fused VJP with library autograd on rectangular odd/even shapes and ranks $\{1, 4, 8\}$, accepting a maximum relative gradient error of 10^{-5}. Factor gradients are contracted in frequency space, and the base linear path forms the dense G_W.

3. Zero-Overhead Deployment via Weight Merging. SpectraGain has zero adapter overhead after merging. The deployment weight is $W_{deploy} = \mathrm{IRFFT2}(\mathrm{RFFT2}(W) \odot [\mathbf{1} + \Pi_{\mathrm{H}}(UV^\top)]; s = (M, N))$. The adapter factors are then removed and inference uses a standard `nn.Linear`.

Memory Attribution Control. Table B.4 reports differences from the fused-LoRA control. The standard branch becomes progressively more expensive than fused LoRA, reaching 14.22 GiB at 96K and an out-of-memory (OOM) boundary at 120K. This empirical difference aggregates the branch graph, saved matrix-multiplication operands, checkpoint interaction, and allocator reuse; the tensor ledger predicts its sequence-length dependence rather than a byte-exact total. SpectraGain remains 0.84–0.94 GiB above fused LoRA in allocated memory and 0.96–1.06 GiB above it in reserved memory, assigning the residual fixed-size gap to the spectral weight path.

Table B.2: Long-context timing and memory measurement protocol.

Component	Setting
Warm-up	Ten training steps at each sequence length.
Timing	Five synchronized training-step repetitions; median latency.
Allocator reset	Reset peak statistics before each repetition.
Peak memory	Maximum allocated and maximum reserved memory, separately, over the five completed repetitions.
OOM handling	Retry in a fresh process; record OOM if the retry fails. No interpolation.

Table B.3: Long-context training profile on one 96-GiB graphics processing unit (GPU). Latency is median seconds per batch over five repetitions; memory is peak allocated GiB. Bold denotes the lowest non-out-of-memory value at each context length.

	Latency (seconds)			Allocated memory (GiB)		
Tokens (thousands)	Standard LoRA	Fused LoRA	SpectraGain	Standard LoRA	Fused LoRA	SpectraGain
1	**0.228**	0.241	2.064	**13.25**	13.72	14.56
16	5.091	**5.022**	6.151	23.17	**21.05**	21.96
64	34.170	**32.941**	33.532	54.94	**45.22**	46.08
96	58.144	**56.603**	57.467	76.12	**61.90**	62.84
120	OOM	**75.214**	77.698	OOM	**73.05**	73.94

Table B.4: Long-context memory attribution relative to fused Low-Rank Adaptation (LoRA). Positive values use more memory than fused LoRA.

Tokens (thousands)	Standard LoRA − fused allocated	SpectraGain − fused allocated	SpectraGain − fused reserved
1	−0.47	0.84	0.96
16	2.12	0.91	1.02
64	9.72	0.86	0.98
96	14.22	0.94	1.06
120	OOM	0.89	1.01

End-to-end profiling and the fused-LoRA attribution control are provided in Section 5.

C Initialization Diagnostics

We compare three precisely defined factor initializations on Recognizing Textual Entailment (RTE) with T5-base. Let $\epsilon_{\mathrm{id}} = \|W_{\mathrm{init}} - W\|_F / \|W\|_F$. The both-zero residual uses $H_+ = \mathbf{1} + \Pi_{\mathrm{H}}(0\,0^\top)$ and blocks both bilinear gradients. The expectation-only direct gain uses $H_+ = \Pi_{\mathrm{H}}(UV^\top)$ with independent $U_{ik} \sim \mathcal{N}(r^{-1/2}, 2/M)$ and $V_{jk} \sim \mathcal{N}(r^{-1/2}, 2/N_+)$; its expected gain is one, but a sampled gain is not the identity almost surely. The asymmetric residual of Eq. 6 is exactly identity preserving and activates the V gradient on the first step.

Table C.1: T5-base Recognizing Textual Entailment (RTE) initialization diagnostic. Accuracy is mean ± sample standard deviation over three seeds.

Strategy	Gain at initialization	ϵ_{id}	Active first-step gradient	Accuracy
Both-zero residual (frozen-equivalent)	$\mathbf{1} + \Pi_{\mathrm{H}}(0\,0^\top)$	0	none	0.00
Expectation-only direct gain	$\Pi_{\mathrm{H}}(UV^\top)$	> 0 almost surely	U, V	71.48 ± 1.57
Asymmetric residual (Ours)	$\mathbf{1} + \Pi_{\mathrm{H}}(U0^\top)$	0	V	71.12 ± 1.30

Writing $G_R = \partial\mathcal{L}/\partial R$, the bilinear chain gives $\nabla_U \mathcal{L} = G_R V$ and $\nabla_V \mathcal{L} = G_R^\top U$. Both gradients therefore vanish for the both-zero residual, while the asymmetric residual has $\nabla_U \mathcal{L} = 0$ and generally nonzero $\nabla_V \mathcal{L}$ at the first step. Of the three initializations, only the asymmetric residual preserves the initial weight exactly while allowing first-step learning.

D Experiment Details

We provide the detailed hyperparameters used for all experiments in this section. T5-base experiments on the General Language Understanding Evaluation (GLUE), mechanism interventions, and initialization ablation use a single NVIDIA RTX 4090 with 24 GB graphics processing unit (GPU) memory in 32-bit floating-point (FP32) precision. LLaMA2-7B long-context and Contrastive Language–Image Pre-training (CLIP) experiments use a single NVIDIA RTX PRO 6000 Blackwell with 96 GB GPU memory; the language-model profile uses bfloat16 (BF16) precision and CLIP uses FP32 precision. Unless otherwise specified, our experiments adopt the following default configuration:

Multi-Seed and Fixed-Run Reporting. Accuracy and generation tables with uncertainty use three paired seeds and report mean ± sample standard deviation. The long-context profile reports five-repeat median latency and peak allocator readings under the protocol in Table B.2.

1. **Optimizer.** We use the AdamW optimizer with $\beta = (0.9, 0.999)$ and weight decay set to 0.0.
2. **Scheduler.** We employ a cosine annealing scheduler with a linear warmup ratio of 0.03. This means the first 3% of total training steps are used for flat-top warmup.
3. **Adaptation Scope.** We apply adaptation to all linear modules in the architecture. This specifically excludes embedding layers, normalization layers, residual connections, and model prediction heads. Image classification tasks using CLIP-ViT-B/16 are an exception (see Appendix D.4).

Furthermore, we enable gradient clipping with a maximum gradient norm of 1 across all experimental runs.

Compute Resources and Training Time. For T5-base tasks on the RTX 4090, per-run training time ranges from minutes for the Microsoft Research Paraphrase Corpus and Corpus of Linguistic Acceptability to roughly 60 minutes for Multi-Genre Natural Language Inference under the unified five-task protocol.

On the RTX PRO 6000, CLIP image classification takes between 15 seconds for small datasets and roughly 7.5 minutes for Street View House Numbers (SVHN), with a combined total of under 20 minutes across all seven datasets per run.

Parameter-Budget Fairness. We count the actual trainable tensors rather than equating nominal ranks. The selected 192 T5-base transformer projections comprise 144 square attention projections and 48 feed-forward-network (FFN) projections. Table D.1 gives the integer allocation. FourierFT trains one real coefficient vector of length n per adapted matrix (Gao et al., 2024); because the targets have three shapes, we generate and freeze one non-trainable index map per shape. Within a shape, the map is shared across layers, while coefficient vectors remain layer-specific. The Circular Convolution Adaptation (C3A)-style control follows the corresponding block-circulant parameter rule.

Table D.1: T5-base trainable-parameter allocation for the controlled protocol. Gap is measured relative to Low-Rank Adaptation (LoRA) rank 8.

Method	Allocation	Per-target count	Total	Gap
LoRA	$r = 8$ for all targets	$r(M + N)$	3,244,032	0.00%
HiRA	$r = 8$ for all targets	$r(M + N)$	3,244,032	0.00%
C3A-style	$b_{attn} = 96,\ b_{ffn} = 48$	MN/b	3,244,032	0.00%
FourierFT	$n = 16{,}896$ per target	n	3,244,032	0.00%
SpectraGain	$r = 11$ for all targets	$r(M + \lfloor N/2 \rfloor + 1)$	3,347,520	+3.19%

For each T5 task, a class-stratified 10% controller-dev partition of the official training split, fixed with split seed 2027, selects FourierFT's $\alpha \in \{50, 100, 200, 300\}$, the learning rate, and an epoch horizon in $\{3, 5, 10\}$. FourierFT uses a fixed shape-group index-map seed of 2027 rather than tuning the map seed. The two-stage learning-rate search first evaluates $\{10^{-4}, 10^{-3}, 5 \times 10^{-3}\}$ and then refines the interval with $\{5 \times 10^{-4}, 2 \times 10^{-3}\}$. All four added operators receive the same five learning-rate candidates and trial budget; ties favor the shorter horizon and then the smaller learning rate. After freezing the configuration, we restart from the same initialization family, refit on the complete official training split for that horizon, and evaluate the official validation split once per seed. These three frozen evaluations supply the reported mean and sample standard deviation.

D.1 T5-Base Natural Language Understanding

We follow the RaLoRA T5-base configuration (Ye et al., 2026) on Multi-Genre Natural Language Inference (MNLI), Stanford Sentiment Treebank 2 (SST-2), the Corpus of Linguistic Acceptability

(CoLA), Question Natural Language Inference (QNLI), and the Microsoft Research Paraphrase Corpus (MRPC). Every task, including CoLA, is evaluated by validation accuracy; the table average is the unweighted mean of its five displayed task values. Training uses FP32, maximum sequence length 128, Adam with $(\beta_1, \beta_2) = (0.9, 0.999)$ and zero weight decay, batch size 32, and a cosine learning-rate schedule with 3% warmup. Adaptation covers all 192 linear layers except the language-modeling head. The source-reported LoRA-family rows retain their original configurations. The added FourierFT, HiRA, C3A-style, and SpectraGain rows share targets, schedule, three paired seeds, checkpoint-selection rule, and learning-rate trial budget; Table D.2 lists the selected settings. FourierFT uses index-map seed 2027. The C3A-style allocation uses the official block-circulant parameter rule MN/b (Chen et al., 2025); block sizes are chosen from common divisors of each target shape to match the rank-8 LoRA budget exactly, independently of task outcomes.

Table D.2: Selected T5-base optimization settings. Each method cell gives peak learning rate / epochs; α is FourierFT's scale.

Task	FourierFT	HiRA	C3A-style	SpectraGain	α
MNLI	10^{-3} / 10	10^{-4} / 3	5×10^{-3} / 5	10^{-3} / 10	300
SST-2	5×10^{-3} / 3	10^{-4} / 5	10^{-4} / 5	2×10^{-3} / 3	300
CoLA	5×10^{-3} / 10	5×10^{-3} / 3	5×10^{-3} / 5	5×10^{-3} / 10	100
QNLI	10^{-4} / 5	10^{-4} / 10	10^{-3} / 5	5×10^{-3} / 5	200
MRPC	10^{-3} / 5	10^{-3} / 5	5×10^{-3} / 10	5×10^{-3} / 10	200

D.2 LLaMA-3.1-8B Commonsense-15K Comparison

Table 3 follows the main Commonsense-15K comparison of Wu et al. (2026a). The mixed training corpus covers BoolQ, Physical Interaction Question Answering (PIQA), Social Interaction Question Answering (SIQA), HellaSwag, WinoGrande, the AI2 Reasoning Challenge easy and challenge sets (ARC-e and ARC-c), and OpenBookQA. Training uses AdamW, 100 warm-up steps, a linear schedule, maximum sequence length 512, batch size 64, and three epochs. Every displayed method targets q_{proj}, k_{proj}, v_{proj}, up_{proj}, and $down_{\text{proj}}$.

The reported OFTv2, LoRA, PiSSA, DoRA, LoRA-XS, and PSOFT values follow Table 5 of Wu et al. (2026a). For the five target types in each of 32 layers, rank-8 HiRA has $14{,}155{,}776$ trainable parameters. SpectraGain uses rank 11 uniformly and has $14{,}059{,}232$ parameters under $r(M + \lfloor N/2 \rfloor + 1)$, closely matching the 14.2M rank-8 LoRA budget. HiRA and SpectraGain use the paired seeds 3407, 7229, and 9119. Both receive the peak-rate candidates $\{10^{-4}, 10^{-3}, 5 \times 10^{-3}\}$; the same task-stratified 10% holdout of the Commonsense-15K training mixture, fixed with split seed 2027, selects 10^{-3} for both methods. We select the epoch with the highest unweighted mean validation accuracy across the eight holdout partitions, breaking ties in favor of the earlier epoch, and evaluate the eight reported sets only after the rank, learning rate, and checkpoint rule are fixed. The table reports task means and the sample standard deviation of the eight-task average. Across paired seeds, the SpectraGain-minus-HiRA average differences are 1.06, 0.94, and 1.04 points (mean $+1.01$, sample standard deviation 0.06; $t(2)$ 95% confidence interval $[0.850, 1.171]$).

Table D.3 complements the main comparison with the extremely low parameter-count setting of PSOFT Table 15. It tests the same eight tasks and $Q/K/V/U/D$ scope across the source table's two budget clusters: rank-1 HiRA belongs to the 1.7M cluster and rank-2 SpectraGain to the 2.5M cluster. These integer ranks place the two multiplicative parameterizations in their respective budget clusters.

Table D.3: Ultra-low-budget LLaMA-3.1-8B accuracy on Commonsense-15K. Reported baselines follow Table 15 of Wu et al. (2026a); HiRA and SpectraGain use the same eight-task, $Q/K/V/U/D$ protocol. Their averages show sample standard deviation over three seeds; bold marks the best result within each parameter-budget cluster.

Method	Params.	BoolQ	PIQA	SIQA	HellaS	WinoG	ARC-e	ARC-c	OBQA	Average
1.7M-parameter cluster										
LoRA ($r = 1$)	1.77M	71.13	85.31	**74.67**	89.08	**72.61**	90.24	78.16	82.40	80.45
PiSSA ($r = 1$)	1.77M	**72.05**	84.60	74.21	**89.93**	70.88	90.15	79.01	82.00	80.35
LoRA-XS ($r = 104$)	1.73M	71.04	85.47	72.67	89.26	71.74	90.82	79.61	**83.20**	80.48
PSOFT ($r = 146$)	1.74M	71.31	**85.69**	73.18	89.38	72.38	**90.91**	**80.03**	83.00	**80.74**
HiRA ($r = 1$)	1.77M	71.25	85.22	73.82	89.21	72.38	90.46	79.03	82.76	80.52±0.03
2.5M-parameter cluster										
DoRA ($r = 1$)	2.56M	71.05	85.29	73.25	90.09	**73.32**	**90.74**	**79.75**	81.87	80.67
PSOFT ($r = 176$)	2.52M	**71.47**	**86.02**	**75.33**	**90.81**	72.69	90.45	78.75	**84.00**	**81.19**
SpectraGain ($r = 2$)	2.56M	71.27	85.96	75.19	90.63	72.57	90.50	78.63	83.98	81.09±0.02

HellaS: HellaSwag; WinoG: WinoGrande; ARC-e/c: AI2 Reasoning Challenge easy/challenge sets; OBQA: OpenBookQA. M denotes millions of trainable parameters.

D.3 LLaMA-3.1-8B Contemporary Comparison

We follow the camera-ready RaLoRA protocol (Ye et al., 2026). The backbone is LLaMA-3.1-8B-Base, and adaptation is restricted to the attention projections W_q, W_k, W_v, W_o. Mathematical reasoning uses a 100K subset of MetaMathQA and regex-extracted Grade School Math 8K (GSM8K) accuracy. Code generation uses a 100K code-only subset of Code-Feedback and HumanEval pass@1. Dialogue uses 52K WizardLM examples and the MT-Bench single-turn protocol, averaging GPT-4o, Gemini-1.5-Pro, and LLaMA-3.1-70B-Instruct judges under the standardized prompts. Every method is repeated with three random seeds.

The optimizer is AdamW with $(\beta_1, \beta_2) = (0.9, 0.999)$ and weight decay 5×10^{-4}, batch size 64, cosine decay factor 0.1, and warmup ratio 0.03. The peak learning rate is 5×10^{-5} for full fine-tuning, SpectraGain, and all reported baselines except AdaLoRA, which uses 5×10^{-4}. SpectraGain retains the exact identity initialization of Eq. 6.

For the 32-layer backbone with hidden width 4096 and key/value width 1024, rank-8 LoRA over the four projections has 6,815,744 trainable parameters. SpectraGain uses rank 12 uniformly over W_q, W_k, W_v, W_o, giving 7,079,424 parameters under the one-sided formula $r(M + \lfloor N/2 \rfloor + 1)$, a 3.87% difference. This actual-parameter match defines the comparison in Table 4.

D.4 Image Classification (CLIP)

We adopt the CLIP-ViT-B/16 protocol reported by Ye et al. (2026) across StanfordCars, the Describable Textures Dataset (DTD), EuroSAT, the German Traffic Sign Recognition Benchmark (GTSRB), Remote Sensing Image Scene Classification 45 (RESISC45), Scene Understanding 397 (SUN397), and Street View House Numbers (SVHN). Adaptation covers every linear layer in the vision backbone while the remaining parameters stay frozen. Training uses FP32 and the prompt template "a photo of a {class}." Optimization uses Adam with $(\beta_1, \beta_2) = (0.9, 0.999)$ and zero weight decay, batch size 64, and a cosine learning-rate schedule with 3% warmup. A fixed 10% class-stratified controller-dev

partition of each official training split, using split seed 2027, selects the peak rate and an epoch horizon in $\{5, 10, 20\}$; ties favor the shorter horizon and then the smaller learning rate. Table D.4 lists the selected configurations. After freezing both choices, we restart from the same initialization family, refit on the complete official training split for the selected horizon, and evaluate the official validation split once per seed. The reported average is the unweighted mean of the seven validation accuracies.

Table D.4: Selected CLIP ViT-B/16 optimization settings.

	HiRA		SpectraGain	
Dataset	Peak learning rate	Epochs	Peak learning rate	Epochs
Cars	10^{-3}	5	10^{-3}	20
DTD	10^{-4}	10	10^{-3}	5
EuroSAT	10^{-3}	5	5×10^{-3}	5
GTSRB	10^{-3}	10	5×10^{-3}	5
RESISC45	5×10^{-3}	20	5×10^{-3}	10
SUN397	10^{-4}	10	5×10^{-3}	10
SVHN	5×10^{-3}	10	5×10^{-3}	20

The reported baseline rows in Table 5 retain the values from RaLoRA Table 3. Its LoRA-family methods use rank 8 or reference rank 8. For the added operator comparison, rank-8 HiRA has 1,327,104 trainable parameters. SpectraGain uses rank 11 uniformly across every adapted linear layer, for 1,369,368 parameters—a 3.18% difference. This rank is fixed by the parameter budget, independently of controller-dev outcomes. Both methods use the same targets, optimizer, schedule, seed set, and checkpoint-selection rule.

At the frozen pretrained checkpoint, before adapter optimization, let $G_{d,\ell}$ be the controller-dev loss gradient averaged over the complete controller-dev partition for dataset d and adapted layer ℓ. At the common identity start, let $Q_{o,\ell}$ be Frobenius-orthonormal bases for the active-factor Jacobian ranges of the parameter-matched circular ($o = \mathrm{circ}$) and pointwise ($o = \mathrm{point}$) correction families. Thus $P_{o,\ell}(G) = \mathrm{unvec}(Q_{o,\ell} Q_{o,\ell}^{\top} \mathrm{vec}(G))$. Under the asymmetric starts, the active-factor coordinate counts are 456,984 for SpectraGain and 663,552 for HiRA; these bound the corresponding tangent dimensions. Appendix F gives the distinction between coordinate count and Jacobian rank. We aggregate projected energy over all adapted layers and standardize the difference across datasets:

$$e_{o,d} = \frac{\sum_{\ell} \|P_{o,\ell}(G_{d,\ell})\|_F^2}{\sum_{\ell} \|G_{d,\ell}\|_F^2}, \qquad a_d = e_{\mathrm{circ},d} - e_{\mathrm{point},d}, \qquad A_d = \frac{a_d - \bar{a}}{s_a}. \tag{D.1}$$

Table D.5 gives the raw energies, standardized scores, and held-out accuracy differences. The full seven-task Spearman correlation is $\rho = 0.964$; deleting one dataset at a time gives $\rho \in [0.943, 1.000]$. The diagnostic is fixed before official-validation evaluation and uses weight-gradient geometry rather than an input-image frequency label.

Table D.5: CLIP controller-dev gradient alignment and official-validation accuracy difference. Δ is SpectraGain minus Hadamard High-Rank Adaptation (HiRA), in percentage points.

Dataset	e_{point}	e_{circ}	a_d	A_d	Δ
Cars	0.1971	0.1994	0.0023	−0.193	0.05
DTD	0.2615	0.2462	−0.0153	−0.632	0.01
EuroSAT	0.2287	0.1937	−0.0350	−1.124	−0.10
GTSRB	0.2370	0.2783	0.0412	0.781	0.32
RESISC45	0.1800	0.1490	−0.0310	−1.025	−0.14
SUN397	0.2489	0.3165	0.0676	1.439	0.38
SVHN	0.2416	0.2818	0.0401	0.754	0.27

E Dataset Details

Table E.1 records the five GLUE subsets used by the unified T5-base comparison. The Commonsense-15K, generation, and CLIP dataset protocols are specified in Appendices D.2, D.3, and D.4, respectively.

Table E.1: T5-base natural-language-understanding tasks under the RaLoRA protocol.

Dataset	Task	Evaluation split	Metric
MNLI	Natural language inference	Validation	Accuracy
SST-2	Sentiment classification	Validation	Accuracy
CoLA	Linguistic acceptability	Validation	Accuracy
QNLI	Question–sentence inference	Validation	Accuracy
MRPC	Paraphrase classification	Validation	Accuracy

MNLI: Multi-Genre Natural Language Inference; SST-2: Stanford Sentiment Treebank 2; CoLA: Corpus of Linguistic Acceptability; QNLI: Question Natural Language Inference; MRPC: Microsoft Research Paraphrase Corpus.

F Mechanism Replacements

Let $R = UV^\top$ denote the raw generator and $G_+ = \Pi_{\mathrm{H}}(R)$ the admissible one-sided residual gain. In this appendix, RFFT2 denotes the two-dimensional real-input fast Fourier transform and IRFFT2 its inverse, so the default operator is

$$W_{\mathrm{SG}} = \mathrm{IRFFT2}(\mathrm{RFFT2}(W) \odot (\mathbf{1} + G_+)), \qquad \Delta W_{\mathrm{SG}} = W_{\mathrm{SG}} - W. \tag{F.1}$$

The study compares the nearest operator families, tests circular versus pointwise coupling with a shared generator, and examines directions retained beyond a strict correction-rank ceiling after training.

Matched Operator Controls. The Hadamard High-Rank Adaptation (HiRA)-style branch uses the spatial pointwise operator (Huang et al., 2025),

$$\Delta W_{\mathrm{HiRA}} = W \odot (AB), \tag{F.2}$$

and therefore retains frozen-carrier multiplication without circular mixing. The Circular Convolution Adaptation (C3A)-style branch follows the additive block-circulant operator (Chen et al., 2025),

$$\Delta W_{\mathrm{C3A}} = C_{\mathrm{blk}}(w), \tag{F.3}$$

and therefore retains circular mixing without frozen-carrier multiplication. Rank-8 HiRA uses 3.244M parameters. For C3A-style, $b_{attn} = 96$ and $b_{ffn} = 48$ give exactly 3.244M learned kernel coefficients under MN/b. Uniformly rank-11 SpectraGain uses 3.348M parameters.

Same-Generator Coupling Replacement. To isolate coupling from generator design, both branches use the same U, V factorization and dimensions, projected-gain construction, kernel $K = \mathrm{IRFFT2}(G_+)$, frozen carrier W, targets, initialization seed, and trainable parameter count. The retrained replacement applies $\Delta W_{\mathrm{point}} = \kappa_\ell (W \odot K)$, where $\kappa_\ell = \|J_{\mathrm{circ},\ell}\|_F / \|J_{\mathrm{point},\ell}\|_F$ is fixed at the identity start using the Jacobians of the unscaled corrections $W \circledast K$ and $W \odot K$ with respect to the active factor V. It receives the same controller-dev candidates and selection rule as SpectraGain before refitting on the complete training split. The fixed-checkpoint arm instead reuses the selected SpectraGain factors, substitutes $W \odot K$ only at evaluation, and restores each layer's replacement correction to the circular correction's Frobenius norm.

For task t and operator o, let $Q_{o,t}$ be an orthonormal basis for the active-factor Jacobian range at initialization and g_t the controller-dev weight gradient. We measure $E_{o,t} = \|Q_{o,t}^\top g_t\|_2^2 / \|g_t\|_2^2$ and report

$$a_t = \frac{E_{\mathrm{circ},t} - E_{\mathrm{point},t}}{E_{\mathrm{circ},t} + E_{\mathrm{point},t}}. \tag{F.4}$$

Positive a_t indicates that the circular tangent captures more gradient energy. Each task contributes one initialization-time diagnostic score; the three paired accuracy seeds quantify downstream variation conditional on that score. The two operator spaces need not have the same dimension solely because their parameter counts agree.

Tangent Geometry and Optimization. Let B_ℓ map the active factor perturbation to $\mathrm{vec}(K)$, and let $L_{o,\ell}$ map the kernel to the correction. Then $J_{o,\ell} = L_{o,\ell} B_\ell$ and

$$\mathrm{rank}(J_{o,\ell}) = \mathrm{rank}(B_\ell) - \dim(\mathrm{range}(B_\ell) \cap \ker L_{o,\ell}) \,. \tag{F.5}$$

Thus both branches preserve the common generator tangent dimension if their coupling maps are injective on its range. Sufficient conditions are nonzero entries of W for pointwise multiplication and a nonzero full Fourier spectrum of W for circular convolution, with nonzero κ_ℓ. Shared factor dimensions alone do not establish these conditions. For numerical evaluation of the projector, a rank-revealing singular value decomposition of $J_{o,\ell}$ supplies orthonormal left singular vectors; retaining $\sigma_i > \max(m,n)\epsilon_{\mathrm{mach}}\sigma_1$ for an $m \times n$ Jacobian defines the numerical range. Layerwise projected energies are summed before normalization. Numerical dimension equality requires checking the retained ranks for both branches.

Projection energy measures local direction accessibility. Under ordinary gradient descent, the first-order weight change is $-\eta J J^\top g$ and the predicted loss decrease is $\eta \|J^\top g\|^2$. In contrast, $\|Q^\top g\|^2$ removes the singular-value weighting of J. An adaptive optimizer additionally rescales parameter directions. Matching $\|J\|_F$ fixes aggregate tangent scale, while a_t characterizes the accessible subspaces rather than the optimizer's step.

Numerical Operator Check. We explicitly construct the active-factor Jacobians for 16×16 and 24×16 matrices at rank 3, using 12 Gaussian initializations of U and $V = 0$ in 64-bit arithmetic. Each shape uses one fixed Gaussian carrier and the rank threshold above. Across all 24 initializations, the generator, circular, and pointwise tangent ranks are all 27. Replacing the carrier by an all-ones matrix leaves the pointwise rank at 27 but reduces the circular rank to 1: its Fourier spectrum supports only the constant mode. This small-matrix check illustrates how the carrier controls accessible directions under an unchanged generator and parameter count.

Table F.1: Same-generator circular and pointwise replacements using the same spectral-kernel parameterization. Values are validation accuracy mean ± sample standard deviation over three paired seeds; swap is evaluation-only with layerwise norm restoration.

Model	Task	Circular	Pointwise, retrained	Pointwise, swap	a_t
T5-base	MNLI	86.17±0.08	85.94±0.16	85.99±0.11	+0.741
	SST-2	94.32±0.22	94.21±0.20	94.08±0.19	+0.698
	CoLA	78.21±0.41	78.36±0.48	78.41±0.51	−0.398
	QNLI	93.97±0.10	93.74±0.10	93.79±0.09	+0.418
	MRPC	85.42±0.33	85.30±0.29	85.11±0.29	+0.936
	Average	87.62±0.08	87.51±0.10	87.48±0.11	–
CLIP ViT-B/16	Cars	86.92±0.44	87.06±0.48	87.17±0.33	−0.552
	DTD	78.07±0.51	78.15±0.53	78.37±0.58	−0.873
	EuroSAT	98.97±0.27	99.02±0.39	98.98±0.32	−0.184
	GTSRB	99.62±0.10	99.73±0.07	99.58±0.06	−0.296
	RESISC45	96.89±0.15	96.72±0.24	96.69±0.11	+0.354
	SUN397	78.05±0.13	78.18±0.25	78.30±0.19	−0.814
	SVHN	97.62±0.12	97.67±0.04	97.70±0.15	+0.005
	Average	90.88±0.07	90.93±0.10	90.97±0.04	–

For circular minus pointwise, the paired aggregate differences are $+0.107$ points on T5-base (95% confidence interval (CI) $[-0.051, 0.266]$) and -0.058 on CLIP ViT-B/16 (CI $[-0.211, 0.094]$) after retraining. The fixed-checkpoint differences are $+0.143$ (CI $[0.071, 0.215]$) and -0.094 (CI $[-0.172, -0.015]$), respectively. Intervals use the three paired seed-level task averages and Student's t distribution with two degrees of freedom. The fixed-checkpoint effects concern the selected circular solutions; the independent retraining means have the same directions with intervals spanning zero.

All correlations use unrounded seed-derived task means and the task-level alignment scores. For retraining/swaps, T5-base has $\rho = 0.500/0.700$ with leave-one-task-out ranges $[0, 0.800]/[0.400, 1.000]$; CLIP has $0.714/0.821$ with ranges $[0.543, 0.886]/[0.714, 0.886]$. Pooled correlations are $0.818/0.916$ and leave-one-out ranges are $[0.773, 0.855]/[0.891, 0.945]$. These ranges measure sensitivity to the observed task set, not confidence intervals or variability across initializations. Diagnostic and mean-response signs agree on 11 of 12 tasks after retraining and 10 of 12 under swaps. SVHN has a near-zero positive diagnostic (0.005) and negative mean responses; the swap also reverses the diagnostic sign on GTSRB. The scatter plots retain both cases. The taskwise association and paired effects support a task-conditioned role for coupling geometry in the evaluated T5 and CLIP settings.

Post-Hoc Realized-Rank Intervention. For each trained SpectraGain correction $D = U_D \Sigma_D V_D^\top$, let $P_k(D) = U_k \Sigma_k V_k^\top$ be its 32-bit floating-point (FP32) best rank-k approximation computed by singular value decomposition (SVD). At evaluation only, we apply

$$\widetilde{D}_k = \frac{\|D\|_F}{\|P_k(D)\|_F} P_k(D), \qquad k \in \{2r, 4r, 8r\}, \tag{F.6}$$

with the unmodified D as the full condition. The model is trained and selected once; no truncated branch is retrained. Frobenius-norm restoration removes update magnitude as a competing explanation, while the post-hoc design leaves the optimization path unchanged.

Intermediate-Rank Diagnostic. For singular values $\{\sigma_i\}$ of the applied correction, we report

$$r_{\text{eff}}(\Delta W) = \exp\left(-\sum_i p_i \log p_i \right), \qquad p_i = \frac{\sigma_i}{\sum_j \sigma_j}, \tag{F.7}$$

using zero for an identically zero correction. Effective rank summarizes the concentration of singular values, whereas Proposition 1 concerns algebraic rank, the number of nonzero singular values. For each task–seed checkpoint, we first take the median across all 192 adapted matrices. We then form each seed's five-task aggregate and report the mean and sample standard deviation across the three seeds. Effective rank is invariant to norm restoration. The operator controls compare coupling choices at matched budgets; the $2r$, $4r$, and $8r$ rows measure the task response to progressively retained correction directions on fixed trained checkpoints.

Table F.2: T5-base operator controls and post-hoc rank intervention (Figure 3). Means over three seeds; ± denotes sample standard deviation. Bold and underline mark best and second-best task results.

Operator / intervention	r_{eff}	MNLI	SST-2	CoLA	QNLI	MRPC	Average
HiRA-style	124.5±1.8	86.15	**94.46**	<u>78.28</u>	93.74	85.32	87.59±0.07
C3A-style	–	86.16	94.23	78.07	93.54	85.34	87.47±0.05
SpectraGain, post-hoc $2r$	19.0±0.7	85.82	94.24	77.60	93.80	85.34	87.36±0.17
SpectraGain, post-hoc $4r$	36.2±1.0	86.02	94.35	77.94	93.88	**85.51**	87.54±0.05
SpectraGain, post-hoc $8r$	69.4±1.4	**86.23**	<u>94.38</u>	**78.38**	**94.10**	85.41	**87.70**±0.15
SpectraGain, full	129.7±1.7	<u>86.17</u>	94.32	78.21	<u>93.97</u>	<u>85.42</u>	<u>87.62</u>±0.08

G Scope and Deployment Considerations

Operating Regime. The fused implementation trades fixed FFT workspace for avoiding sequence-dependent adapter state. Standard LoRA crosses both fused methods in memory at 16K; fused LoRA remains leaner, while SpectraGain stays within 0.94 GiB through 120K and approaches its latency profile from 64K. At 1K, transform cost dominates, motivating dispatch between standard and fused SpectraGain paths.

Evaluation Frontier. Evidence spans exact-answer and semantic generation, long-context memory, preference-judged dialogue, and unified CLIP. Matched-budget studies on additional backbones and judge-free instruction-following metrics would test transfer beyond this protocol family.