NONLINEARITY-AWARE LORA: STRUCTURED GATE ADAPTATION UNDER LOW-RANK CONSTRAINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-rank adaptation (LoRA) constrains weight updates, yet their matrix-space fit does not by itself determine which features a nonlinear feed-forward network (FFN) transmits. We identify *gate-selection mismatch*: target and adapter scores can induce different transmission coefficients even when their weight-space discrepancy is small. To explain this mechanism, we introduce a multi-criteria decision-making interpretation of FFN evaluation and selection, then develop a homogeneity-based response framework for low-rank adaptation. It links relative gate response, channel aggregation, and the coupled FFN output, motivating persistent channel prioritization and regime-conditioned coordination. We instantiate these principles in **NA-LoRA**, using a temporally smoothed response-sensitive channel mask and gradient modulation coordinated across the gate, value, and output projections. The mask generalizes across FFN architectures, while the current macro rule specializes to SwiGLU's characteristic geometry. Language and vision-language experiments show improvements over vanilla LoRA and competitive results against strong PEFT variants; component ablations and nonlinear-state diagnostics support the proposed framework and design. Both training-only controls preserve LoRA's rank, forward parameterization, and inference cost.

1 INTRODUCTION

Low-rank adaptation (LoRA) (Hu et al., 2022) restricts trainable updates to low-rank factors. Figure 1 illustrates why matrix fit alone is insufficient: a target update and its low-rank approximation can induce different feature-transmission coefficients inside a feed-forward network (FFN). We call this discrepancy *gate-selection mismatch*, distinct from mismatch in the transmitted values.

This gap arises from nonlinear FFN behavior: equal-sized score perturbations can affect transmission differently across gate states, especially near regime boundaries. Weight- and gradient-alignment methods (Wang et al., 2024; 2025) improve matrix-space approximation, but this criterion alone does not resolve the nonlinear consequences of the remaining error (§ 2).

To explain this gap, we interpret score-dependent suppression and transmission as an evaluation–selection process, drawing on the existing grey multi-criteria decision-making (MCDM) interpretation of activation (Cai, 2023; 2024; Cai et al., 2026b):

Intuition 1.1 (MCDM interpretation of the projection–activation pipeline)**.** *In MCDM, candidate solutions, commonly termed* alternatives*, are evaluated according to multiple* criteria*, often relative to a reference or ideal solutions (Deng, 1982; Liu, 2025; Rezaei, 2016; Qin et al., 2017; Xu et al., 2020; Joshi & Kumar, 2016). We interpret the FFN projection–activation pipeline through this lens. Specifically, an input feature vector is treated as an alternative, its channels serve as evaluation criteria, and each projection filter acts as an updatable reference solution (*i.e.*, pattern). The projected score is then interpreted as an indication of the feature's importance relative to that reference.*

The subsequent activation (or gating) recalibrates this indication into state-dependent transmission, interpreted as directed feature selection. Under restricted adaptation, a score discrepancy can thus become a nonlinear discrepancy in the selection/transmission coefficient.

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.

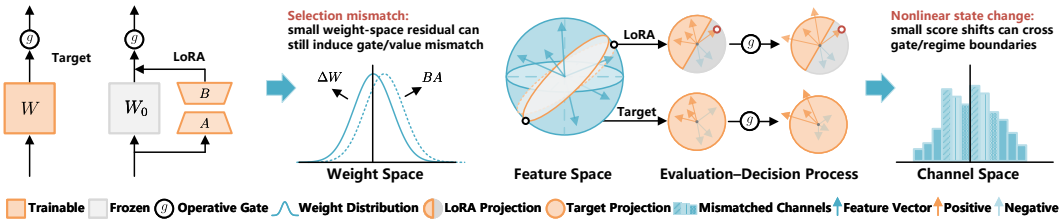


Figure 1: **Selection mismatch beyond weight-space fit.** Left: a small matrix residual can still induce gate/value mismatch. Middle: target and adapter projections lead to different feature selection. Right: score shifts near gate/regime boundaries alter channel transmission. This motivates interpreting evaluation and selection (Intuition 1.1) before quantifying the nonlinear response. The target provides an analytical reference for characterizing selection mismatch.

Building on this account of *gate-selection mismatch*, we develop a *homogeneity-based response description* for low-rank adaptation. Starting with SwiGLU, effective homogeneity characterizes relative gate changes, while the value branch and output projection determine their contribution to the coupled FFN response. This characterization motivates two complementary design principles:

(1) **Adaptation should prioritize population-relative responsiveness with temporal memory.** Score changes act through each channel's nonlinear state: a strongly suppressed channel can have extreme relative sensitivity but weak absolute response. Since the FFN combines many channels, we assess responsiveness in the context of the channel population. As responsiveness varies across inputs and training steps, this preference should retain history beyond one batch. This motivates *history-aware channel prioritization*, guided by task gradients.

(2) **Selection-focused modulation should be coordinated at the FFN level.** In SwiGLU's positive-asymptotic regime, nonlinear reweighting accounts for a decreasing share of gate response as transmission becomes more direct. This suggests tempering selection-focused modulation as the regime becomes prevalent. Yet transmitted values and their output aggregation can remain mismatched. We therefore seek *regime-conditioned FFN coordination*, using evolving gate populations to modulate coupled-path gradients.

These principles lead to **Nonlinearity-Aware LoRA (NA-LoRA)**. Its channel-level (*micro*) and FFN-level (*macro*) controls specify where to emphasize selection reconfiguration and how to coordinate the coupled response as gate states evolve. Both act only during training, preserving LoRA's rank, forward parameterization, and inference cost (§ 4).

Contributions. (1) We identify gate-selection mismatch and develop an MCDM-inspired, homogeneity-based FFN response framework explaining why matrix-space fit alone need not preserve feature transmission under low-rank adaptation. (2) Building on this analysis, we motivate population-relative channel prioritization with temporal memory and regime-aware FFN coordination, and instantiate them in NA-LoRA. (3) Language generation, language understanding, and image-classification transfer experiments demonstrate NA-LoRA's practical value across SwiGLU, ReLU, and QuickGELU models. Ablations support the complementary controls, while diagnostics track the nonlinear-state changes targeted by the framework.

2 Related Work

Low-rank adaptation. Within PEFT (Houlsby et al., 2019; Lester et al., 2021; Ding et al., 2023), LoRA variants follow four nonexclusive routes (He et al., 2026). *Rank adjustment* expands, shares, or budgets capacity, including AdaLoRA's importance-based and GoRA's gradient-driven allocation (Zhang et al., 2023; He et al., 2025). *Optimization-process adjustment* improves stability or alignment: rsLoRA and LoRA+ adjust scaling and factor learning rates (Kalajdzievski, 2023; Hayou et al., 2024); DoRA and LoRA-Pro address magnitude–direction adaptation and equivalent-gradient approximation (Liu et al., 2024; Wang et al., 2025). *Initialization adjustment* selects starting subspaces from weights, gradients, or activations; PiSSA and LoRA-GA exemplify the first two (Meng et al., 2024; Wang et al., 2024). *MoE integration* coordinates multiple low-rank experts through

losses, routing, or expert structure (Feng et al., 2024). Activation-informed designs combine mechanisms within these categories: EVA (Paischer et al., 2025) couples activation-based initialization with explained-variance rank budgeting, while AIRA (Li et al., 2025) combines outlier-weighted initialization, outlier-driven rank budgeting, and activation-informed optimization. These statistics indirectly guide update configuration or optimization; by themselves, they do not characterize how the FFN's coupled nonlinearity converts residual score error into selection mismatch. We instead study the FFN as a whole.

Nonlinear selection and our position. IIEU, AdaShift, CSaN, and FleS (Cai, 2023; 2024; Cai et al., 2026a;b) interpret activation as directed feature selection, providing the conceptual basis for examining the projection–gating pipeline. Within the alignment-enhancing branch of optimization-process adjustment (He et al., 2026), LoRAN (Li et al., 2024) applies Sinter, a nonlinear transformation, to the low-rank factor product, while AuroRA (Dong et al., 2025) inserts an adaptive nonlinear layer between low-rank projections. RanLoRA (Luo et al., 2026) combines SVD-based residual initialization with nonlinear projection and vector modulation, spanning initialization and optimization-process adjustment. In contrast to prior approaches that introduce additional nonlinearity into the adapter, NA-LoRA exploits the nonlinear selection performed by the FFN to guide adaptation while preserving the standard LoRA forward parameterization.

3 Rethinking Low-Rank Adaptation through FFN Selection

We analyze SwiGLU before discussing other FFN architectures.

3.1 Selection mismatch and its nonlinear cause

In SwiGLU (Shazeer, 2020) (Figure 1), z, u, h, y are the score, value, hidden state, and output:

$$z = W_g x, \qquad u = W_u x, \qquad h = g(z) \odot u, \qquad y = W_d h, \quad g(z) = \mathrm{SiLU}(z) = z\sigma(z). \tag{1}$$

Biases are included in the appendix. The full SiLU response is the *operative gate* multiplying u, yielding a signed transmission coefficient.

For analytical target state T and adapter state R, the hidden discrepancy is exactly

$$h_T - h_R = \underbrace{[g(z_T) - g(z_R)] \odot u_R}_{\text{gate-selection mismatch}} + \underbrace{g(z_T) \odot (u_T - u_R)}_{\text{gated-value mismatch}}. \tag{2}$$

Even with matched values, score discrepancies can alter nonlinear transmission via the first term. Corollary D.1 allows down-projection changes.

Why matrix fit can select the wrong direction. Holding value and down paths fixed, a small score-projection residual ξE around (z_0, u_0, W_d) gives leading squared output discrepancy $\xi^2 \mathcal{Q}(E)$ under the assumptions in § A.4, where

$$\mathcal{Q}(E) = \mathbb{E}_x \|J_g(x) E x\|_2^2, \qquad J_g(x) = W_d \mathrm{diag}\big(u_0(x) \odot g'(z_0(x))\big). \tag{3}$$

The metric weights residuals by nonlinear sensitivity, values, and output mixing. Under a constant-Gram condition, weighted approximation (Eckart & Young, 1936) yields a same-rank counterexample: the matrix-optimal choice has functional residual 9, versus 4 for another rank-one update. This motivates characterizing response geometry under rank-constrained FFN adaptation.

3.2 Homogeneity-based characterization of the FFN response

Intuition 3.1 (Channel response in the context of the FFN)**.** *A relative score change must be interpreted through both gate response and output aggregation. Channel-wise homogeneities can vary, and their contributions to the block response depend jointly on hidden values and output mixing.*

Define the gate's *effective homogeneity* by

$$\kappa_g(z) = \frac{z g'(z)}{g(z)}. \tag{4}$$

Removable values are extended by continuity. For $z, g(z) \neq 0$ and locally C^2 g, $\Delta g/g = \kappa_g(z)\Delta z/z + \mathcal{O}(\Delta z^2)$ connects homogeneity to relative transmission changes. The following result extends this description to the FFN output.

Proposition 3.1 (Homogeneity-based FFN response)**.** *Fix z, u, W_d and a gate differentiable at the channel scores. Along the joint score–value ray $h(\alpha) = g(\alpha z) \odot (\alpha u)$, $y(\alpha) = W_d h(\alpha)$, let $\chi_{\mathrm{FFN}} = 1 + \kappa_g$ wherever defined. Then*

$$\left.\frac{\mathrm{d}\log|h_c(\alpha)|}{\mathrm{d}\log\alpha}\right|_{\alpha=1} = \chi_{\mathrm{FFN}}(z_c), \qquad y'(1) = W_d\big[h \odot \chi_{\mathrm{FFN}}(z)\big]. \tag{5}$$

The logarithmic identity requires $h_c \neq 0$; the vector identity also holds through removable gate singularities by continuity. For bias-free branches this is the input-scaling ray; with biases it is a local score–value perturbation.

The product rule and output linearity establish this channel-to-output aggregation (§ F.2).

Relative gate response and absolute sensitivity. For SwiGLU, $g' = \sigma\kappa_g$ by continuity. In the negative tail, $\kappa_g \to -\infty$ while $|g'| \to 0$: extreme relative sensitivity can accompany weak absolute response. $|g'|$ measures local responsiveness, but task relevance also requires values and downstream mixing (Eq. (3)). A population-relative response signal therefore complements task gradients, which retain the value, downstream, and loss factors.

3.3 Regime structure and implications for adaptation

For SwiGLU, $\kappa_g = 1 + z(1 - \sigma(z))$, with $\kappa_g(0) = 1$. Its negative zero and positive maximum are $-\tau_z$ and $+\tau_z$, where $\tau_z\sigma(\tau_z) = 1$ and $\tau_z \simeq 1.27846$ (§ F.3). These define *suppressive* ($z < -\tau_z$), *responsive* ($|z| \leq \tau_z$), and *positive-asymptotic* ($z > \tau_z$) regions.

To understand the positive region, separate $g' = \sigma + z\sigma'$. The relative share of nonlinear reweighting is

$$\frac{z\sigma'(z)}{g'(z)} = \frac{\kappa_g(z) - 1}{\kappa_g(z)}, \qquad z > \tau_z. \tag{6}$$

Beyond τ_z, the nonlinear-reweighting share decreases to zero while $g'(z) \to 1$, so transmission becomes increasingly direct and remains locally responsive (Corollary F.3).

Intuition 3.2 (Gate-regime evolution and coupled FFN response)**.** *Gate regimes characterize the current transmission mechanism, while agreement with a target depends jointly on the gate, value, and output paths.*

The coupling is explicit in the complete local differential:

$$\mathrm{d}y = W_d[u \odot g'(z) \odot \mathrm{d}z] + W_d[g(z) \odot \mathrm{d}u] + (\mathrm{d}W_d)h. \tag{7}$$

Together, Eqs. (6) and (7) motivate full-FFN coordination as gate states evolve (§ 4.2).

3.4 Scope across FFN architectures

Architectural extension starts from the multiplicative operative gate. SwiGLU gates an independent value with SiLU; conventional FFNs have $h = g(z) \odot z$: QuickGELU uses $g(z) = \sigma(1.702z)$ (Radford et al., 2021), exact GELU uses $g(z) = \Phi(z)$ (Hendrycks & Gimpel, 2016), and ReLU uses $g(z) = \mathbb{I}(z > 0)$ (Nair & Hinton, 2010). QuickGELU/GELU lack the negative characteristic point required by the symmetric SwiGLU partition; ReLU has only the hard boundary at zero. Regime definitions therefore follow each architecture's operative-gate geometry; § C gives complete definitions and boundary conditions.

4 From FFN Response Analysis to NA-LoRA

NA-LoRA translates the analysis in § 3 into complementary micro and macro controls (Figure 2): identify persistently responsive channels and coordinate the FFN as selection evolves. The functional metric characterizes the response geometry; NA-LoRA uses derivative and regime statistics to guide online gradient modulation.

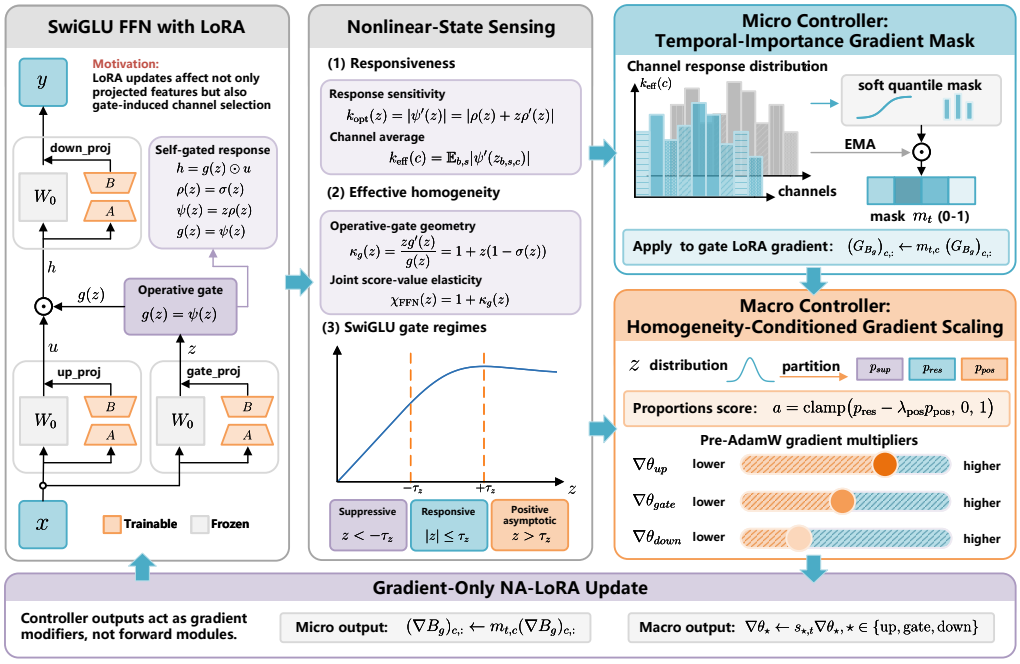


Figure 2: **NA-LoRA: sensing nonlinear state, modulating training.** Micro maps channel responsiveness to a history-smoothed, population-relative mask; macro maps regime occupancy to projection-specific FFN scales. Both modify gradients before AdamW, preserving the forward pass.

4.1 Population-Relative Channel Prioritization

Intuition 4.1 (Persistent, population-relative channel responsiveness)**.** *§ 3.2 distinguishes extreme relative response from local responsiveness. We therefore assess channels relative to their population. Because this assessment varies across inputs and training steps, its history provides a more persistent reference than the current batch alone.*

Response statistic. The self-gated response (Cai et al., 2026b) and derivative magnitude are:

$$\psi(z) = z\rho(z), \qquad h = g(z) \odot v, \qquad k_{\text{opt}}(z) = |\psi'(z)| = |\rho(z) + z\rho'(z)|. \tag{8}$$

For SwiGLU, $\psi = g = \text{SiLU}$ and $v = u$, so $k_{\text{opt}} = |g'|$. For conventional FFNs, $g = \rho$ and $v = z$, giving $k_{\text{opt}} = |g + zg'|$, which includes both gate and tied-value contributions. ReLU uses its almost-everywhere derivative, without differentiating across the hard switch.

We combine sample/token averaging, channel quantiles, and mask history (Intuition 4.1):

$$\begin{aligned} k_{\text{eff}}(c) &= \mathbb{E}_{b,s}|\psi'(z_{b,s,c})|, \qquad \tau_k = \text{Quantile}(k_{\text{eff}}, 1 - \rho_{\text{keep}}), \\ m_c^{\text{new}} &= \sigma\big(\beta_m(k_{\text{eff}}(c) - \tau_k)\big), \qquad m_t = \eta_m m_{t-1} + (1 - \eta_m)m^{\text{new}}. \end{aligned} \tag{9}$$

Back-propagation combines local response derivatives with sample-dependent task signals. Micro adds a persistent channel preference through sample/token averaging of absolute response derivatives, population-relative quantile comparison, and EMA of mask assignments.

Where the preference acts. Rows of $B \in \mathbb{R}^{d_h \times r}$ correspond to score channels, so m_t multiplies its gradient rowwise; A remains unmasked. Larger mask entries retain a larger fraction of the current B-gradient row. The task gradient retains value, downstream, and loss information.

4.2 Regime-Conditioned FFN Coordination

Intuition 4.2 (Regime-conditioned coordination of the coupled FFN)**.** *Eq.* (6) *suggests tempering selection-focused modulation as positive-tail transmission becomes established. Eq.* (7) *shows why the value and output paths must remain involved. We therefore use the gate population to coordinate the strength of FFN gradient inputs, while task supervision provides the learning signal.*

Table 1: **Natural language generation (NLG) tasks.** MT-Bench first-turn score, GSM8K exact-answer accuracy (%), and HumanEval pass@1 (%). All-linear PEFT uses rank/reference rank 8 by default; bold/underline mark the best/second-best default-setting means, including ties. Protocols and provenance: § I.

	Llama-3.1-8B			Llama-2-7B		
Method	**MT-Bench**	**GSM8K**	**HumanEval**	**MT-Bench**	**GSM8K**	**HumanEval**
Full	6.02±0.12	73.69±0.28	51.63±1.27	5.30±0.11	59.36±0.85	35.31±2.13
LoRA	6.08±0.02	71.52±0.73	41.05±2.85	5.61±0.10	42.08±0.04	14.76±0.17
rsLoRA	6.16±0.02	73.85±0.23	42.27±3.46	5.25±0.03	45.62±0.10	16.01±0.79
AdaLoRA	6.18±0.16	73.74±0.71	43.49±1.63	5.57±0.05	50.72±1.39	17.80±0.44
DoRA	6.32±0.10	74.65±1.19	45.12±2.44	5.97±0.02	53.07±0.75	19.75±0.41
LoRA+	6.27±0.02	74.42±0.88	43.09±2.85	5.71±0.08	52.11±0.62	18.17±0.52
PiSSA	6.10±0.01	73.59±1.11	43.70±4.47	5.30±0.02	44.54±0.27	16.02±0.17
LoRA-GA	6.00±0.05	74.30±1.06	44.10±2.85	5.95±0.16	53.60±0.30	19.81±1.46
LoRA-Pro	6.21±0.06	74.93±1.01	43.63±1.49	5.72±0.03	57.57±0.50	22.97±0.35
GoRA	6.30±0.02	74.55±1.04	45.32±1.62	5.61±0.12	54.04±0.22	24.80±1.04
RaLoRA	6.25±0.05	73.92±0.23	44.31±1.41	5.83±0.05	53.21±1.36	23.37±0.81
RaLoRA-Pro	6.34±0.09	73.49±0.81	45.32±1.96	**5.98±0.09**	53.14±0.45	24.18±0.70
NA-LoRA$_{r=8}$	**6.39±0.04**	**75.16±1.34**	**46.34±1.22**	5.92±0.03	**57.97±3.08**	**26.02±0.81**
NA-LoRA$_{r=32}$	6.20±0.02	75.33±1.42	47.13±1.01	5.72±0.04	59.46±0.78	27.64±0.41
NA-LoRA$_{r=128}$	6.12±0.04	76.06±0.48	47.76±1.02	5.78±0.02	59.74±1.13	28.66±2.44

From regime population to coordinated scales. Pooling samples, tokens, and channels gives $p_{\text{res}} = \mathbb{E}[\mathbb{I}(|z| \le \tau_z)]$ and $p_{\text{pos}} = \mathbb{E}[\mathbb{I}(z > \tau_z)]$. Occupancy summarizes the block state through the prevalence of each regime. We use a shared signal with projection-specific maps:

$$\begin{aligned} a &= \text{clamp}(p_{\text{res}} - \lambda_{\text{pos}} p_{\text{pos}}, 0, 1), \\ s_\star^{\text{new}} &= \text{clip}(1 + \alpha_\star(2a - 1), s_{\min}^\star, s_{\max}^\star), \\ s_{\star,t} &= \beta_s s_{\star,t-1} + (1 - \beta_s) s_\star^{\text{new}}, \qquad \star \in \{\text{up}, \text{gate}, \text{down}\}. \end{aligned} \tag{10}$$

Responsive occupancy increases relative modulation, while positive-asymptotic occupancy tempers it. A shared state coordinates the three paths, while distinct $\alpha_\star$ and clipping ranges respect their different roles. Positive floors avoid shutting off a projection, and EMA smooths the feedback. The scales multiply both factors' gradients in each FFN adapter; the score-producing B also receives the micro mask.

4.3 Gradient-only realization and applicability

Both controls act before AdamW's moment update (Kingma & Ba, 2015; Loshchilov & Hutter, 2019); attention adapters retain standard LoRA. Micro applies across the studied architectures, whereas ReLU and QuickGELU/GELU use $s_\star = 1$ because they lack the SwiGLU partition (§ 3.4).

With $W_{\text{eff}} = W_0 + \gamma BA$ and actual factor directions $P^+ = P - \eta_t U_{P,t}$, the score change at fixed input and bias, with adapter dropout disabled, is exactly

$$\delta z = \gamma[-\eta_t(U_{B,t}A + BU_{A,t}) + \eta_t^2 U_{B,t}U_{A,t}]x. \tag{11}$$

The controls modify the gradients accumulated by AdamW; the resulting moment-dependent directions $U_{A,t}$ and $U_{B,t}$ determine the actual factor and score updates. We provide further optimizer analysis in § F.5 and implementation details in § H.

5 Experiments

We evaluate task performance, controller contributions, and nonlinear-state reconfiguration.

Baselines. Tables 1–3 compare LoRA (Hu et al., 2022) with rsLoRA's rank-stabilized scaling (Kalajdzievski, 2023), LoRA+'s asymmetric factor learning rates (Hayou et al., 2024), and DoRA's

Table 2: **Natural language understanding (NLU) tasks.** T5-Base development-set accuracy (%) on five GLUE tasks; Average is their arithmetic mean. Protocol and result sources: § I.

Method	MNLI	SST-2	CoLA	QNLI	MRPC	Average
Full	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	82.08
PiSSA	85.75±0.07	94.07±0.06	74.27±0.39	93.15±0.14	76.31±0.51	84.71
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
DoRA	85.67±0.09	94.04±0.53	72.04±0.94	93.04±0.06	68.08±0.51	82.57
AdaLoRA	85.45±0.11	93.69±0.20	69.16±0.24	91.66±0.05	68.14±0.28	81.62
LoRA-GA	85.70±0.09	94.11±0.18	80.57±0.20	93.18±0.06	85.29±0.24	87.77
LoRA-Pro	86.03±0.19	94.19±0.13	**81.94±0.24**	93.42±0.05	86.60±0.14	**88.44**
GoRA	85.91±0.02	**94.68±0.43**	79.86±0.35	93.27±0.08	86.10±0.20	87.96
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
NA-LoRA	**86.29±0.10**	93.96±0.07	81.46±0.13	**93.44±0.06**	**87.01±0.00**	88.43

Table 3: **CLIP-ViT-B/16 transfer.** Accuracy on seven image-classification datasets under the LoRA-Pro protocol; baseline sources are listed in § I.4.

Method	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
Full FT	84.23±0.06	77.44±0.19	98.09±0.03	94.31±0.28	93.95±0.00	75.35±0.10	93.04±0.18	88.06
LoRA	72.81±0.13	73.92±0.38	96.93±0.07	92.40±0.10	90.03±0.14	70.12±0.18	88.02±0.07	83.46
rsLoRA	82.38±0.20	78.03±0.76	98.06±0.08	95.04±0.11	93.96±0.18	75.38±0.24	92.74±0.18	87.94
LoRA+	72.87±0.18	74.07±0.45	97.01±0.02	92.42±0.18	89.96±0.11	70.17±0.15	88.08±0.05	83.51
DoRA	73.72±0.06	73.72±0.33	96.95±0.01	92.38±0.17	90.03±0.08	70.20±0.19	88.23±0.05	83.48
LoRA-GA	85.18±0.41	77.50±0.12	98.05±0.27	95.28±0.10	94.43±0.19	75.44±0.06	93.68±0.35	88.51
LoRA-Pro	**85.87±0.08**	**78.64±0.25**	98.46±0.03	95.66±0.05	94.75±0.21	76.42±0.14	94.63±0.20	89.20
GoRA	85.76±0.19	78.17±0.32	98.77±0.35	**96.66±0.36**	95.16±0.26	**76.46±0.08**	95.32±0.13	89.47
NA-LoRA	85.28±0.20	78.26±0.77	**98.96±0.04**	96.14±0.46	**95.57±0.11**	75.84±0.16	**97.14±0.04**	**89.60**

magnitude–direction decomposition (Liu et al., 2024). PiSSA and LoRA-GA initialize from principal weight components and gradients, respectively (Meng et al., 2024; Wang et al., 2024). LoRA-Pro adjusts factor gradients to approximate full fine-tuning (Wang et al., 2025). AdaLoRA allocates rank by importance (Zhang et al., 2023); GoRA combines gradient-based rank allocation and initialization (He et al., 2025). RaLoRA uses block-diagonal adapters to align effective rank with gradient intrinsic dimensionality; RaLoRA-Pro adds loss-sensitivity-based parameter reallocation across layers (Ye et al., 2026). Full fine-tuning and zero-shot CLIP provide references.

Setup. NLG follows LoRA-GA/LoRA-Pro-style protocols (Wang et al., 2024; 2025) with Llama-3.1-8B (Grattafiori et al., 2024) and Llama-2-7B (Touvron et al., 2023). We fine-tune on WizardLM, MetaMathQA, and Code-Feedback for MT-Bench, GSM8K, and HumanEval, respectively. NLU uses T5-Base (Raffel et al., 2020) on five GLUE tasks (Wang et al., 2019); image classification uses CLIP-ViT-B/16 (Radford et al., 2021) on seven datasets. PEFT defaults to non-head linear modules at rank/reference rank 8; NA-LoRA controls only FFNs. § I details evaluation protocols, ranks, and result sources.

5.1 Task performance across gate families

Natural language generation tasks. At rank/reference rank 8, NA-LoRA leads PEFT means on five of the six NLG evaluations (Table 1), including all Llama-3.1-8B tasks, where its gains over LoRA are 0.31, 3.64, and 5.29. RaLoRA-Pro has the highest Llama-2-7B MT-Bench mean (5.98 versus NA-LoRA's 5.92). Higher NA-LoRA ranks (32, 128) improve reasoning and code, while MT-Bench favors rank 8.

Table 4: **Component ablation.**

Method	GSM8K	HumanEval
FT	73.69±0.28	51.63±1.27
LoRA	71.52±0.73	41.05±2.85
rsLoRA	73.85±0.23	42.27±3.46
NA-LoRA	**75.16±1.34**	**46.34±1.22**
w/o Micro	74.53±1.82	45.53±1.42
w/o Macro	74.43±0.63	44.31±5.69

Table 5: **Target-module ablation.**

Method	Targets	GSM8K	HumanEval
FT	—	73.69±0.28	51.63±1.27
LoRA	All linear	71.52±0.73	41.05±2.85
LoRA	Attention	67.78±1.25	43.09±0.35
LoRA	FFN	71.97±1.23	42.07±0.61
NA-LoRA	FFN	73.34±0.63	44.31±2.64
NA-LoRA	All linear	**75.16±1.34**	**46.34±1.22**

Llama-3.1-8B; PEFT rank $r = 8$. FT is full fine-tuning (no adapter-target restriction). "w/o Micro/Macro" removes the corresponding control. Bold/underline mark the best/second-best PEFT means within each table. FT and all-linear reference results repeat Table 1.

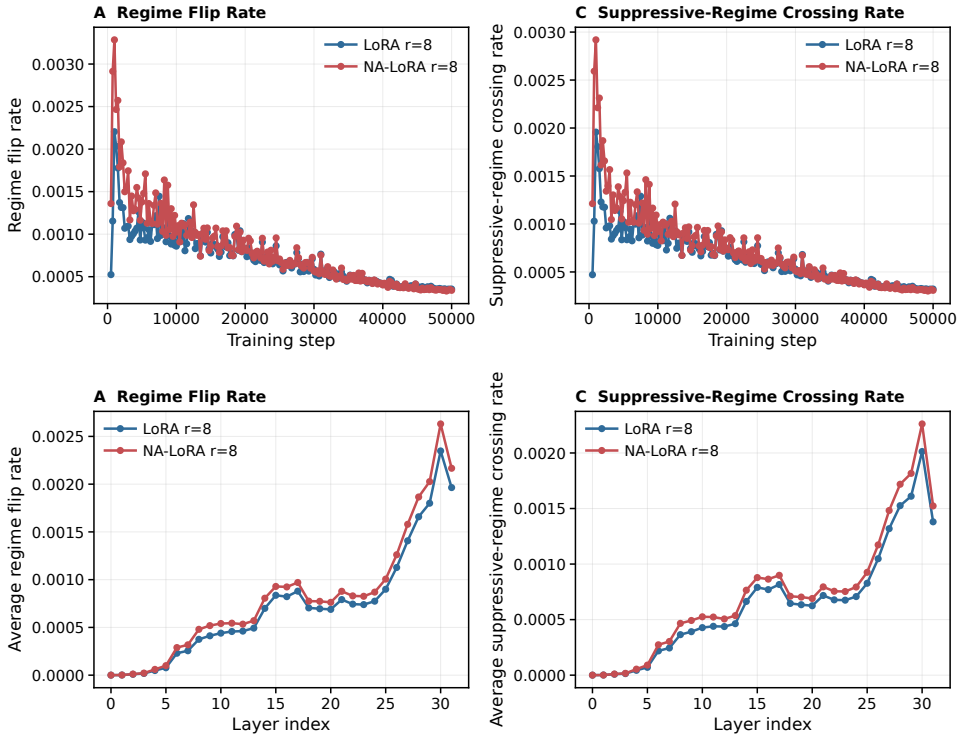


Figure 3: **Llama-3.1-8B regime trajectories.** Regime flips and suppressive-boundary crossings by step (top) and layer (bottom) on a fixed GSM8K-train probe. Red: NA-LoRA; blue: LoRA.

Natural language understanding tasks. On T5-Base (Table 2), NA-LoRA averages 88.43, above LoRA (82.08), RaLoRA (87.24), and RaLoRA-Pro (87.23) (Ye et al., 2026), and close to LoRA-Pro (88.44), supporting competitiveness beyond SwiGLU.

Image-classification transfer. On CLIP-ViT-B/16 (Table 3), NA-LoRA has the highest reported average (89.60), versus GoRA's 89.47 and LoRA-Pro's 89.20. T5 and CLIP test micro-only control ($s_\star = 1$) across ReLU and QuickGELU, while SwiGLU experiments and ablations test joint micro–macro control (§ B.3).

5.2 Testing the controller design

Complementarity of the control resolutions. Table 4 tests Intuitions 4.1 and 4.2. Removing micro lowers GSM8K and HumanEval by 0.63 and 0.81 points; removing macro lowers them by 0.73 and 2.03. Combining the two controls gives the highest means among the controller variants, supporting micro–macro complementarity.

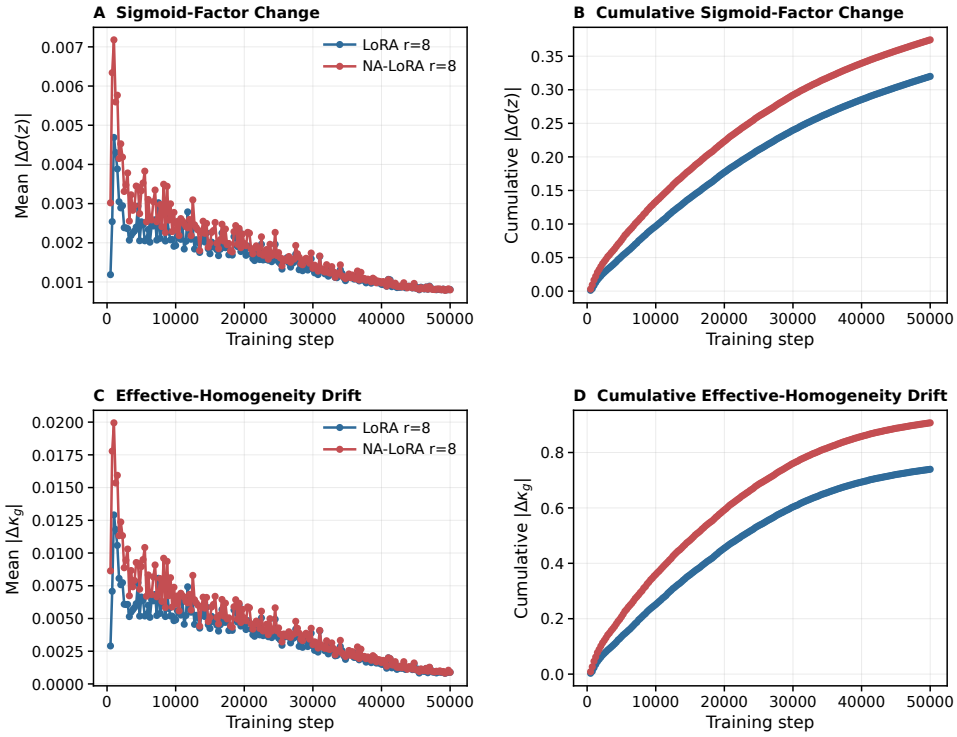


Figure 4: **Llama-3.1-8B drift.** Adjacent-checkpoint and cumulative drift for internal weight $\rho = \sigma$ (top) and operative-gate homogeneity κ_g, $g = \mathrm{SiLU}$ (bottom). Red: NA-LoRA; blue: LoRA.

Control beyond FFN target selection. Table 5 asks whether choosing FFNs alone explains the gain. With matched FFN-only targets, NA-LoRA improves GSM8K by 1.37 and HumanEval by 2.24 over LoRA. All-linear NA-LoRA reaches 75.16 and 46.34, retaining standard attention adapters. The matched gains support modulation beyond target selection.

5.3 Nonlinear-state reconfiguration during training

We probe Llama-3.1-8B `mlp.gate_proj` scores during MetaMathQA fine-tuning on fixed GSM8K-train samples every 250 steps. Regime crossings test qualitative state changes; homogeneity drift tests continuous response changes, using the operative SiLU gate and boundaries ± 1.27846.

Figure 3 shows more cumulative regime flips (0.1609 versus 0.1421) and suppressive-boundary crossings (0.1449 versus 0.1279), especially early and in middle/deeper layers. Distant-regime jumps stay near zero (3.36×10^{-7} versus 1.10×10^{-7}): changes mainly join adjacent regions.

Figure 4 measures $\mathbb{E}|\Delta\kappa_g|$ without regularization or clipping. Cumulative drift increases from 0.7394 to 0.9072 (+22.7%), with larger mean drift in all 32 layers. Auxiliary internal-weight motion $|\Delta\sigma(z)|$ rises from 0.3199 to 0.3744 (+17.0%), but is distinct from operative-gate homogeneity.

EMA smooths control signals while gate states evolve under task supervision. These diagnostics describe reconfiguration activity, without directly measuring mismatch reduction (§ K).

6 Conclusion

Our MCDM-inspired, homogeneity-based framework characterizes how state-dependent transmission and coupled output formation shape low-rank FFN adaptation. These response relations motivate NA-LoRA's channel prioritization and FFN coordination. Task results, ablations, and diagnostics support their practical value and associated state changes. Nonlinear-response analysis thus guides constrained adaptation within the existing LoRA parameterization.

AI USE STATEMENT

Generative AI assistance was used to draft and revise text, develop and check the local theoretical derivations, audit bibliographic records, and prepare the LaTeX manuscript. The derivations underwent algebraic and numerical consistency checks, and references were checked against primary publication records. The Llama-3.1-8B MT-Bench protocol uses GPT-4o; the Llama-2-7B reference protocol uses GPT-4. § I documents these judge conventions and the published instruction corpora. The authors are responsible for the final scientific content, including AI-assisted material.

REPRODUCIBILITY STATEMENT

The homogeneity-based response characterization, operative-gate definitions, theoretical assumptions, derivations, and optimizer identities are provided in §§ A–G. The theory-to-design bridge is detailed in § B; the controller procedure and reported hyperparameters appear in § H. Benchmark datasets, task protocols, target modules, rank configurations, evaluation metrics, result provenance, and seed conventions are documented in § I.

REFERENCES

Sudong Cai. IIEU: Rethinking Neural Feature Activation from Decision-Making. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5796–5806, 2023.

Sudong Cai. AdaShift: Learning Discriminative Self-Gated Neural Feature Activation With an Adaptive Shift Factor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5947–5956, 2024.

Sudong Cai, Shuai Yuan, Bingzhi Chen, Rui Mao, and Bing Wang. Selection-as-Nonlinearity: Bridging Attention and Activation via a Joint Game-Decision Lens for Interpretable, Discriminative Visual Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11621–11631, 2026a.

Sudong Cai, Shuyuan Zheng, Bingzhi Chen, Shuai Yuan, Chuan Xiao, Jianbin Qin, and Bing Wang. Toward Principled Flexible Scaling for Self-Gated Neural Activation. In *International Conference on Learning Representations (ICLR)*, pp. 150219–150255, 2026b.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating Large Language Models Trained on Code. *arXiv preprint arXiv:2107.03374*, 2021.

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. *arXiv preprint arXiv:2110.14168*, 2021.

Ju-Long Deng. Control Problems of Grey Systems. *Systems & Control Letters*, 1(5):288–294, 1982.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235, 2023.

Haonan Dong, Wenhao Zhu, Guojie Song, and Liang Wang. AuroRA: Breaking Low-Rank Bottleneck of LoRA with Nonlinear Mapping. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 36929–36961, 2025.

Carl Eckart and Gale Young. The Approximation of One Matrix by Another of Lower Rank. *Psychometrika*, 1(3):211–218, 1936.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-Weighted Linear Units for Neural Network Function Approximation in Reinforcement Learning. *Neural Networks*, 107:3–11, 2018.

Wenfeng Feng, Chuzhan Hao, Yuewei Zhang, Yu Han, and Hao Wang. Mixture-of-LoRAs: An Efficient Multitask Tuning Method for Large Language Models. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pp. 11371–11380, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. LoRA+: Efficient Low Rank Adaptation of Large Models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 17783–17806, 2024.

Haonan He, Peng Ye, Yuchen Ren, Yuan Yuan, Luyang Zhou, Shucun Ju, and Lei Chen. GoRA: Gradient-driven Adaptive Low Rank Adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 114074–114102, 2025.

Haonan He, Jingqi Ye, Minglei Li, Zhengbo Wang, Mengqi Li, Tao Chen, Lei Bai, and Peng Ye. A Unified Study of LoRA Variants: Taxonomy, Review, Codebase, and Empirical Evaluation. *arXiv preprint arXiv:2601.22708*, 2026.

Dan Hendrycks and Kevin Gimpel. Gaussian Error Linear Units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

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 *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799, 2019.

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 (ICLR)*, 2022.

Deepa Joshi and Sanjay Kumar. Interval-valued intuitionistic hesitant fuzzy Choquet integral based TOPSIS method for multi-criteria group decision making. *European Journal of Operational Research*, 248(1):183–191, 2016.

Damjan Kalajdzievski. A Rank Stabilization Scaling Factor for Fine-Tuning with LoRA. *arXiv preprint arXiv:2312.03732*, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Brian Lester, Rami Al-Rfou, and Noah Constant. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3045–3059, 2021.

Lujun Li, Dezhi Li, Cheng Lin, Wei Li, Wei Xue, Sirui Han, and Yike Guo. AIRA: Activation-Informed Low-Rank Adaptation for Large Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1729–1739, 2025.

Yinqiao Li, Linqi Song, and Hanxu Hou. LoRAN: Improved Low-Rank Adaptation by a Non-Linear Transformation. In *Findings of the Association for Computational Linguistics: Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3134–3143, 2024.

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 *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32100–32121, 2024.

Sifeng Liu. *Grey Systems Analysis: Methods, Models and Applications*. Springer, Singapore, 2nd edition, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Xu Luo, Yongbin Liu, Chunping Ouyang, and Ying Yu. RanLoRA: Residual-aware Nonlinear Low-Rank Adaptation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 17243–17258, 2026.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. PiSSA: Principal Singular Values and Singular Vectors Adaptation of Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 121038–121072, 2024.

Vinod Nair and Geoffrey E. Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pp. 807–814, 2010.

OpenAI. GPT-4o System Card. Technical report, OpenAI, 2024.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, et al. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023.

Fabian Paischer, Lukas Hauzenberger, Thomas Schmied, Benedikt Alkin, Marc Peter Deisenroth, and Sepp Hochreiter. Parameter Efficient Fine-tuning via Explained Variance Adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 46215–46267, 2025.

Jindong Qin, Xinwang Liu, and Witold Pedrycz. An extended TODIM multi-criteria group decision making method for green supplier selection in interval type-2 fuzzy environment. *European Journal of Operational Research*, 258(2):626–638, 2017.

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 *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

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.

Jafar Rezaei. Best-worst multi-criteria decision-making method: Some properties and a linear model. *Omega*, 64:126–130, 2016.

Noam Shazeer. GLU Variants Improve Transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*, 2023.

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 (ICLR)*, 2019.

Shaowen Wang, Linxi Yu, and Jian Li. LoRA-GA: Low-Rank Adaptation with Gradient Approximation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 54905–54931, 2024.

Zhengbo Wang, Jian Liang, Ran He, Zilei Wang, and Tieniu Tan. LoRA-Pro: Are Low-Rank Adapters Properly Optimized? In *International Conference on Learning Representations (ICLR)*, pp. 93787–93808, 2025.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. WizardLM: Empowering Large Pre-Trained Language Models to Follow Complex Instructions. In *International Conference on Learning Representations (ICLR)*, pp. 30745–30766, 2024.

Ye Xu, Ye Li, Lijun Zheng, Liang Cui, Sha Li, Wei Li, and Yanpeng Cai. Site selection of wind farms using GIS and multi-criteria decision making method in Wafangdian, China. *Energy*, 207: 118222, 2020.

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 *International Conference on Learning Representations (ICLR)*, pp. 55765–55790, 2026.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. MetaMath: Bootstrap Your Own Mathematical Questions for Large Language Models. In *International Conference on Learning Representations (ICLR)*, pp. 45040–45061, 2024.

Xin Yu, Yujia Wang, Jinghui Chen, and Lingzhou Xue. AltLoRA: Towards Better Gradient Approximation in Low-Rank Adaptation with Alternating Projections. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 37484–37522, 2025.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive Budget Allocation for Parameter-Efficient Fine-Tuning. In *International Conference on Learning Representations (ICLR)*, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 46595–46623, 2023.

Tianyu Zheng, Ge Zhang, Tianhao Shen, Xueling Liu, Bill Yuchen Lin, Jie Fu, Wenhu Chen, and Xiang Yue. OpenCodeInterpreter: Integrating Code Generation with Execution and Refinement. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12834–12859, 2024.

APPENDIX CONTENTS

A Detailed Framework: Selection, Homogeneity, and Functional Approximation **16**

A.1 From self-gated activation to architecture-aligned selection 16

A.2 Weight-space fit need not imply operative-gate alignment 17

A.3 Response sensitivity and homogeneity-based FFN characterization 17

A.4 A local functional metric under the rank constraint 19

A.5 Optimizer-consistent gate-state motion 20

A.6 Temporal control as a design principle 21

B Detailed Controller Construction **22**

B.1 From theoretical implications to two control resolutions 22

B.2 Temporal-importance channel mask 23

B.3 Homogeneity-conditioned gradient scaling 24

C Notation and Architecture-Level Gate Mapping **25**

C.1 Internal weight, scalar response, and operative gate 25

C.2 What is inherited from the decision-making interpretation 26

C.3 Algebraic regrouping and value-path dependencies 26

D Proofs: Operative-Gate Selection Mismatch **26**

D.1 Proof of Proposition A.2 26

D.2 Proof of Corollary A.3 26

D.3 Proof of Proposition A.4 27

D.4 Why the bound is data-aware 27

E Local Functional Metrics and Rank-Constrained Approximation **27**

E.1 Proof of Proposition A.10 27

E.2 General quadratic form and the separable special case 27

E.3 Proof of Proposition A.11 27

E.4 Counterexample and architectural scope 28

F Gate Homogeneity, Characteristic Points, and Practical Sensitivity **28**

F.1 Proof of Lemma A.6 28

F.2 Proof of Proposition A.7 28

F.3 Activation-specific gate geometry 29

F.4 Proof of Proposition A.13 30

F.5 AdamW response to a gradient multiplier 31

F.6 Response sensitivity for temporal channel control 31

G Margin-Based Temporal Bounds **32**
G.1 Proof of Proposition A.14 32
G.2 Proof of Corollary A.15 32
G.3 Proof of Corollary A.16 32

H Implementation Details **33**
H.1 NA-LoRA Training-Time Procedure 33
H.2 Hook Locations and Gradient Modification 33
H.3 Hyperparameters 34
H.4 Additional training state and inference cost 34

I Experimental Details **34**
I.1 Tasks and evaluation metrics 34
I.2 Adapters, optimization, and rank configurations 35
I.3 NLU and image-classification protocols 35
I.4 Result provenance and reporting 36

J Additional Experimental Results **37**

K Additional Scope of Gate-Trajectory Diagnostics **37**

L Analytical Scope and Extensions **37**

A Detailed Framework: Selection, Homogeneity, and Functional Approximation

The main text starts with selection mismatch in Figure 1, analyzes its nonlinear FFN mechanism, and then introduces MCDM to interpret evaluation and selection. A homogeneity-based response description connects this interpretation to controller design. This appendix supplies the complete architecture-aligned definitions and assumptions behind that progression. The target update serves as a reference for the mismatch analysis and is not required during NA-LoRA training.

After the gate/value decomposition, we characterize relative gate response and the aggregation of channel responses into the FFN output. A local functional metric makes the rank constraint explicit, and the optimizer analysis identifies how training-only feedback can affect gate-state motion. The exact identities identify the state variables and coupled-projection dependencies. The quantile mask and macro mapping instantiate these implications as the online policies developed in § B.

A.1 From self-gated activation to architecture-aligned selection

Decision-making lens and its scope. Following FleS (Cai et al., 2026b), we use multi-criteria decision making to interpret the affine–activation pipeline. Grey-systems analysis provides reference-based evaluation tools (Deng, 1982; Liu, 2025), while methods such as best–worst weighting estimate criteria weights (**?**). In the activation interpretation developed by IIEU, AdaShift, and FleS (Cai, 2023; 2024; Cai et al., 2026b), each input feature vector x is an alternative, its input channels are criteria, and a filter w is an updatable reference for an ideal pattern. The affine response

$$z = \langle w, x \rangle + b = \|w\|_2 \|x\|_2 \cos\theta_{w,x} + b \tag{12}$$

(the angular expression assumes nonzero w, x) serves as a filter-relative importance indication. Its magnitude combines feature–filter alignment, their norms, and the bias. The ideal pattern is a filter-level interpretive reference, whereas the later target update is a parameter-space analytical reference.

One self-gated scalar response. We retain FleS's distinction between a weighting function and its weighted response, writing its ϕ as ψ to match our controller notation:

$$\psi(z) = z\rho(z). \tag{13}$$

Here ρ is the *internal self-gating weight*, and ψ is the *scalar response*. SiLU uses $\rho(z) = \sigma(z)$ (Elfwing et al., 2018); exact GELU uses $\Phi(z)$ (Hendrycks & Gimpel, 2016); QuickGELU uses $\sigma(1.702z)$ in the original CLIP implementation (Radford et al., 2021); and ReLU uses the hard weight $\mathbb{I}(z > 0)$ (Nair & Hinton, 2010). ReLU fits the factorization, with its switch handled discretely. These weighting functions recalibrate affine responses, and their derivatives enter the loss gradient through the corresponding computation paths.

Definition A.1 (Operative multiplicative gate)**.** For an FFN value path v, the hidden channel is

$$\mathcal{U}(z, v) := g(z) \odot v. \tag{14}$$

The *operative gate* g is the coefficient that multiplies this value path. Its definition includes the computational dependence of v on z, rather than only an algebraic factorization of the output.

For SwiGLU (Shazeer, 2020), the scalar response gates a separate projection. For conventional self-gated FFNs, the response itself is the hidden channel. Allowing affine score/value projections,

$$\text{SwiGLU:} \quad z = W_g x + b_g, \quad v = u = W_u x + b_u, \quad g(z) = \psi(z) = \mathrm{SiLU}(z), \tag{15}$$

$$\text{QuickGELU:} \quad z = W_1 x + b_1, \quad v = z, \quad g(z) = \rho(z) = \sigma(\beta_{\mathrm{QG}} z), \tag{16}$$

$$\text{ReLU:} \quad z = W_1 x + b_1, \quad v = z, \quad g(z) = \rho(z) = \mathbb{I}(z > 0). \tag{17}$$

Here $\beta_{\mathrm{QG}} = 1.702$ for CLIP; exact GELU analogously has $g = \rho = \Phi$. A separate value branch means a distinct computational path, not statistical independence from the score. The branches may depend on the same input x. Table 6 lists the complete mapping.

Intuition A.1 (Activation-level recalibration and FFN-level selection)**.** *FleS's internal weight $\rho(z)$ recalibrates the score to produce $\psi(z)$. Our FFN-level extension analyzes the coefficient $g(z)$ that controls the specified transmitted value v: $g = \psi$ in SwiGLU but $g = \rho$ in conventional*

GELU/ReLU. SwiGLU's operative gate can be signed and unbounded, so its interpretation differs from that of a probability or a conventional nonnegative multi-criteria weight. The extension therefore preserves the source's score–weight–response distinction (Cai et al., 2026b) without transferring boundedness, monotonicity, or saturation assumptions from ρ to g. Selection alignment uses this architecture-specific coefficient; invariance under arbitrary product regroupings is not asserted.

A.2 Weight-space fit need not imply operative-gate alignment

Let target and LoRA-adapted gate-score projections produce z_{tar} and z_{lora}. With target-minus-LoRA weight residual E_{W_z} and possible bias residual e_{b_z}, define

$$\Delta z := z_{\mathrm{tar}} - z_{\mathrm{lora}} = E_{W_z} x + e_{b_z}. \tag{18}$$

For bias-free FFNs, or when the compared procedures leave the bias unchanged, $e_{b_z} = 0$ and this reduces to the usual data-aligned matrix residual $E_{W_z} x$. Similarly let

$$\Delta v := v_{\mathrm{tar}} - v_{\mathrm{lora}}. \tag{19}$$

For SwiGLU, $v = W_u x + b_u$ is an independent value branch and $\Delta v = E_{W_u} x + e_{b_u}$. For a conventional GELU/ReLU FFN, $v = z$ and therefore $\Delta v = \Delta z$.

Proposition A.2 (Exact operative-gate mismatch decomposition)**.** *For any FFN channel written as $\mathcal{U}(z, v) = g(z) \odot v$,*

$$\begin{aligned} &\mathcal{U}(z_{\mathrm{tar}}, v_{\mathrm{tar}}) - \mathcal{U}(z_{\mathrm{lora}}, v_{\mathrm{lora}}) \\ &= \underbrace{\big(g(z_{\mathrm{tar}}) - g(z_{\mathrm{lora}})\big) \odot v_{\mathrm{lora}}}_{\text{operative-gate selection mismatch}} + \underbrace{g(z_{\mathrm{tar}}) \odot \Delta v}_{\text{gated-value mismatch}}. \end{aligned} \tag{20}$$

This decomposition exposes the behavior-space term missed by a purely matrix-level comparison. Two low-rank updates with similar matrix residuals can place different data points in different gate states, and their impact depends on both the data-aligned score residual and the local geometry of the operative gate. In conventional GELU/ReLU FFNs, $v = z$, so the two terms in Eq. (20) are two algebraic pathways generated by the same score residual; they are not assumed to be statistically independent or orthogonal. § D further propagates this hidden-state decomposition through the FFN down projection, making explicit how selection mismatch enters the complete FFN output together with a separate down-projection residual.

Corollary A.3 (Data-aware behavior-space control)**.** *For any ℓ_p norm with $1 \le p \le \infty$,*

$$\begin{aligned} &\|\mathcal{U}(z_{\mathrm{tar}}, v_{\mathrm{tar}}) - \mathcal{U}(z_{\mathrm{lora}}, v_{\mathrm{lora}})\|_p \\ &\le \|v_{\mathrm{lora}}\|_\infty \|g(z_{\mathrm{tar}}) - g(z_{\mathrm{lora}})\|_p + \|g(z_{\mathrm{tar}})\|_\infty \|\Delta v\|_p. \end{aligned} \tag{21}$$

Thus a useful functional bound is necessarily data-aware: it depends not only on a weight residual but also on how the residual moves the operative gate on the inputs visited during adaptation.

Proposition A.4 (Sensitivity-weighted operative-gate mismatch)**.** *Assume g is continuous on the closed segment between $z_{\mathrm{lora},c}$ and $z_{\mathrm{tar},c}$ and differentiable on its interior. For each channel c, there exists an intermediate point $\bar{z}_c$ such that*

$$\big(g(z_{\mathrm{tar},c}) - g(z_{\mathrm{lora},c})\big) v_{\mathrm{lora},c} = g'(\bar{z}_c)\, \Delta z_c\, v_{\mathrm{lora},c}. \tag{22}$$

Consequently, the squared operative-gate mismatch energy is

$$\sum_c |g'(\bar{z}_c)|^2 |v_{\mathrm{lora},c}|^2 |\Delta z_c|^2. \tag{23}$$

For the hard ReLU gate, Proposition A.4 is applied only on segments that do not cross the switching point; crossing $z = 0$ is instead a discrete gate flip. The resulting sensitivity-weighted score residual provides the starting point for the rank-constrained functional approximation problem below.

A.3 Response sensitivity and homogeneity-based FFN characterization

The source of selection mismatch and the statistic that directly modulates back-propagation are related but not identical in every FFN architecture. We therefore distinguish two quantities.

Local optimization sensitivity. The scalar response is uniformly self-gated:

$$\psi(z) = z\rho(z). \tag{24}$$

With ρ fixed to the implemented activation, the practical first-order proxy is

$$k_{\text{opt}}(z) := |\psi'(z)| = |\rho(z) + z\rho'(z)|. \tag{25}$$

The product rule explains why its relation to g depends on the FFN. At smooth points,

$$\mathrm{d}h = v \odot g'(z) \odot \mathrm{d}z + g(z) \odot \mathrm{d}v. \tag{26}$$

For a SwiGLU score-projection perturbation at fixed x and fixed value projection, $\mathrm{d}v = 0$, and $g = \psi$. If $q = \partial\mathcal{L}/\partial h$ is the upstream hidden-state gradient, then

$$\nabla_{W_g}\mathcal{L} = (q \odot u \odot \psi'(z))x^\top. \tag{27}$$

For a conventional FFN, $v = z$ and $\mathrm{d}v = \mathrm{d}z$, so $\psi' = g + zg'$ and

$$\nabla_{W_1}\mathcal{L} = (q \odot \psi'(z))x^\top. \tag{28}$$

Both therefore use $|\rho + z\rho'|$. It omits upstream loss signals, input factors, and the separate SwiGLU value; it is neither the full gradient magnitude nor sufficient for functional importance. For SwiGLU,

$$k_{\text{opt}}(z) = \big|\sigma(z) + z\sigma(z)(1 - \sigma(z))\big|. \tag{29}$$

For ReLU, the derivative is $\mathbb{I}(z > 0)$ off zero; the implementation's autodifferentiation convention at zero is separate from the smooth theory.

Gate effective homogeneity.

Definition A.5 (Operative-gate effective homogeneity)**.** For differentiable g with $g(z) \neq 0$, define

$$\boxed{\kappa_g(z) := \frac{zg'(z)}{g(z)}}. \tag{30}$$

At removable singularities, κ_g is defined by continuity. Throughout the paper, *effective homogeneity* refers exclusively to κ_g. For the joint score–value radial path formalized below, we separately denote the full FFN-channel scaling elasticity by

$$\chi_{\text{FFN}}(z) := 1 + \kappa_g(z). \tag{31}$$

Thus gate effective homogeneity and FFN output scaling elasticity are related but named distinctly.

Architecture-dependent gate and response elasticities. Where ρ is differentiable and $z\rho(z) \neq 0$, Eq. (24) gives the exact identity

$$\frac{z\psi'(z)}{\psi(z)} = 1 + \frac{z\rho'(z)}{\rho(z)}. \tag{32}$$

The left-hand side equals κ_g for SwiGLU ($g = \psi$), but $1 + \kappa_g$ for conventional FFNs ($g = \rho$). Separately, $\chi_{\text{FFN}} = 1 + \kappa_g$ describes *joint* score–value scaling. In SwiGLU it includes the additional scaling of u; it is not the elasticity of ψ alone. Consequently, one self-gated response formula does not justify replacing the operative gate by the internal weight in regime analysis.

Lemma A.6 (Relative gate-response bridge)**.** *Let $z \neq 0$ and $g(z) \neq 0$, and assume g is twice continuously differentiable in a neighborhood of z. For a sufficiently small score perturbation δz,*

$$\frac{g(z + \delta z) - g(z)}{g(z)} = \kappa_g(z)\frac{\delta z}{z} + \mathcal{O}(\delta z^2). \tag{33}$$

Equivalently, the first-order operative-gate selection mismatch is

$$\big(g(z + \delta z) - g(z)\big)v = \frac{g(z)}{z}\kappa_g(z)\, v\, \delta z + \mathcal{O}(\delta z^2). \tag{34}$$

Thus κ_g converts relative score perturbations into relative operative-gate perturbations, directly linking the selection-mismatch term to macro gate geometry.

Proposition A.7 (Channel-wise score–value radial scaling sensitivity)**.** *Fix a channel state* (z_c, v_c) *and define the score–value radial path*

$$h_c(\alpha) := g(\alpha z_c)(\alpha v_c), \qquad \alpha > 0. \tag{35}$$

Whenever the logarithmic derivative is defined,

$$\left.\frac{\partial \log |h_c(\alpha)|}{\partial \log \alpha}\right|_{\alpha=1} = 1 + \kappa_g(z_c) = \chi_{\mathrm{FFN}}(z_c). \tag{36}$$

For bias-free linear FFN branches this path coincides with the input scaling ray $x \mapsto \alpha x$*; for affine FFNs it describes a local radial score/value perturbation, not literal input scaling.*

Corollary A.8 (Aggregation into the FFN output)**.** *Under the smooth radial-path assumptions of Proposition A.7, fix* W_d *and set* $h = g(z) \odot v$ *and* $y(\alpha) = W_d[g(\alpha z) \odot (\alpha v)]$*. Where the gate ratios are defined, with removable extensions taken by continuity,*

$$y'(1) = W_d\big[h \odot (\mathbf{1} + \kappa_g(z))\big] = \sum_c W_{d,:,c}\, h_c\, [1 + \kappa_g(z_c)]. \tag{37}$$

The vector output response depends on channel values, elasticities, and down-projection mixing; it is not generally a scalar multiple of y *with degree given by the unweighted mean of* κ_g*.*

This corollary is the output-level part of Proposition 3.1. It motivates evaluating channels in the context of their population without identifying any single channel, or an unweighted average, with the homogeneity of the whole FFN. The controller uses inexpensive population summaries; the actual task gradient retains the value and downstream factors.

The definition is deliberately gate-level. For SwiGLU the operative gate is $g(z) = \mathrm{SiLU}(z)$, so

$$\kappa_{\mathrm{SwiGLU}}(z) = 1 + z(1 - \sigma(z)). \tag{38}$$

For QuickGELU, the operative gate is the sigmoid factor itself, not the complete activation output,

$$\kappa_{\mathrm{QGELU}}(z) = \beta_{\mathrm{QG}} z\big(1 - \sigma(\beta_{\mathrm{QG}} z)\big). \tag{39}$$

For ReLU, $g(z) = \mathbb{I}(z > 0)$. Hence κ_g is undefined for every $z < 0$ because $g(z) = 0$, equals 0 for $z > 0$, and cannot be smoothly extended across the hard switch at $z = 0$.

NA-LoRA derives coarse regimes from characteristic points of κ_g, so they depend on the operative gate, not a universal activation template. SwiGLU yields a finite nonzero symmetric threshold; QuickGELU/exact GELU do not yield the same nonzero symmetric threshold under this construction; and ReLU collapses to the hard boundary at zero. The detailed derivation is given in § F.

Definition A.9 (Effective-homogeneity drift diagnostic)**.** For consecutive states θ_t, θ_{t+1}, define

$$\mathcal{D}^{(1)}_{\mathrm{EH}}(t) := \mathbb{E}_x\left[\frac{1}{d_h}\sum_{c=1}^{d_h} |\kappa_g(z_{t+1,c}(x)) - \kappa_g(z_{t,c}(x))|\right], \tag{40}$$

which is the mean absolute adjacent-state drift plotted in our diagnostics. We use $\mathcal{D}^{(1)}_{\mathrm{EH}}(t)$ to characterize the magnitude of nonlinear-state reconfiguration along training trajectories. Cumulative summaries sum this quantity over adjacent saved-checkpoint pairs.

A.4 A LOCAL FUNCTIONAL METRIC UNDER THE RANK CONSTRAINT

Fix a reference SwiGLU layer with scores $z_0(x)$, independent values $v_0(x)$, and output projection W_d. To isolate the gate-score contribution, hold v_0, W_d, and the score bias fixed, and compare score-projection perturbations ξT and ξR, where $\xi \geq 0$ is a local expansion parameter and $\mathrm{rank}(R) \leq r$. The effective LoRA update can represent any such ξR with a fixed nonzero LoRA coefficient and unconstrained factors. This comparison concerns an update from a common reference, not an individual optimizer increment, whose matrix rank can exceed r.

Proposition A.10 (Gate-induced local functional error)**.** *Let* x *and* $v_0(x)$ *be bounded, and let* g *be* C^2 *with bounded first and second derivatives on all score segments considered for* $0 \leq \xi \leq \xi_0$*. For fixed finite matrices* T, R, W_d*, define* $E = T - R$ *and*

$$J_g(x) := W_d\, \mathrm{diag}\big(v_0(x) \odot g'(z_0(x))\big). \tag{41}$$

With $y_T(\xi, x) = W_d[g(z_0(x) + \xi Tx) \odot v_0(x)]$ and y_R defined analogously,

$$y_T - y_R = \xi J_g(x)Ex + \mathcal{O}(\xi^2), \tag{42}$$

$$\mathbb{E}_x\|y_T - y_R\|_2^2 = \xi^2 \mathcal{Q}(E) + \mathcal{O}(\xi^3), \tag{43}$$

$$\mathcal{Q}(E) := \mathbb{E}_x\|J_g(x)Ex\|_2^2. \tag{44}$$

The remainder bounds are uniform over the stated input distribution for fixed T, R.

The metric includes value magnitudes, cross-channel interactions through W_d, and input statistics. It can be singular because some hidden changes are canceled by the down projection. In general $J_g(x)$ depends on x, so its contribution cannot be separated from the input covariance. A tractable special case makes the rank dependence transparent.

Proposition A.11 (Conditional weighted rank-r approximation)**.** *Suppose $J_g(x)^\top J_g(x) = D \succeq 0$ almost surely, and let $\Sigma = \mathbb{E}[xx^\top] \succeq 0$. Set $C = D^{1/2}T\Sigma^{1/2}$. Then*

$$\min_{\mathrm{rank}(R)\le r} \mathcal{Q}(T - R) = \min_{\mathrm{rank}(R)\le r} \|D^{1/2}(T - R)\Sigma^{1/2}\|_F^2 = \sum_{i>r} \sigma_i(C)^2. \tag{45}$$

One minimizer is

$$R_\star = D^{\dagger/2} C_r \Sigma^{\dagger/2}, \tag{46}$$

where C_r is a truncated SVD using the largest r nonzero singular components (or all of them if fewer exist), and $D^{\dagger/2}$ denotes the pseudoinverse of $D^{1/2}$. For positive definite D and Σ, the pseudoinverses reduce to ordinary inverse square roots.

The classical low-rank approximation theorem (Eckart & Young, 1936) gives the solution in the FFN-induced weighted coordinates. The constant-Gram condition makes this case exactly solvable; the general Jacobian produces an input-dependent quadratic form. A separable surrogate admits the same algebra when explicitly adopted as an approximation. § E gives proofs and the general form.

A rank-one counterexample. Take $T = \mathrm{diag}(2, 1)$, $D = \mathrm{diag}(1, 9)$, $\Sigma = I_2$, and $r = 1$. The unique Frobenius-optimal approximation is $R_W = \mathrm{diag}(2, 0)$, with squared matrix residual 1 but functional residual $\mathcal{Q}(T - R_W) = 9$. The weighted optimum is $R_\star = \mathrm{diag}(0, 1)$, with squared matrix residual 4 but functional residual 4. Scaling both updates by ξ places the example arbitrarily close to the reference. Thus the local functional geometry reverses the ranking of these candidate rank-one updates while preserving continuity as their scale tends to zero.

From functional geometry to online feedback. The metric shows why a score residual should not be evaluated in isolation from the gate, transmitted values, and output mixing. In SwiGLU, $g' = (g/z)\kappa_g = \sigma\kappa_g$ (with the removable value extended at zero), so response sensitivity is the absolute-response counterpart of the relative homogeneity description. NA-LoRA extracts this inexpensive nonlinear factor for channel prioritization and uses the population of homogeneity-derived regimes for block-level coordination. The full back-propagated gradient retains the value, downstream, input, and loss factors. The policy seeks useful reconfiguration of nonlinear selection under the rank constraint; it is not the weighted-SVD solution or an approximation guarantee for the full functional metric. § B distinguishes geometry from quantile, history, and multiplier choices.

For conventional GELU/ReLU FFNs, $v = z$ cannot be held fixed independently of z. The corresponding full-response Jacobian is $J_\psi(x) = W_d\mathrm{diag}(\psi'(z_0(x)))$, with $\psi(z) = g(z)z$; the same local metric construction applies on smooth segments. ReLU boundary crossings remain outside the smooth expansion. Joint SwiGLU value- and down-projection changes add first-order terms (§ E), extending the gate-only comparison to the complete local FFN differential.

A.5 Optimizer-consistent gate-state motion

Gradient modulation before AdamW. The mask and macro scales act on gradients before the optimizer. The local score analysis uses deterministic adapter evaluation, as in probes with LoRA dropout disabled. We retain the fixed LoRA coefficient γ in $W_{\mathrm{eff}} = W_0 + \gamma BA$ rather than absorbing it into an adaptively optimized factor. Let $G_{P,t}$ be a raw gradient for factor $P \in \{A, B\}$. For one controller evaluation per optimizer minibatch, the inputs are

$$H_{A,t} = s_t G_{A,t}, \qquad H_{B,t} = s_t M_t G_{B,t}, \qquad M_t = \mathrm{diag}(m_t). \tag{47}$$

If hooks run on separate microbatches, their modulated contributions must first be accumulated; any clipping or other gradient processing must also be included in the actual optimizer input $H_{P,t}$. In that case a single factorization as in Eq. (47) need not hold. The analysis below uses the final $H_{P,t}$ supplied to AdamW in either case.

Definition A.12 (Optimizer-consistent local factor update)**.** With zero-initialized moments, standard AdamW (Kingma & Ba, 2015; Loshchilov & Hutter, 2019) updates pre-step factor P at $t \geq 1$ as

$$\begin{aligned} \mu_{P,t} &= \beta_1 \mu_{P,t-1} + (1-\beta_1) H_{P,t}, \\ \nu_{P,t} &= \beta_2 \nu_{P,t-1} + (1-\beta_2) H_{P,t}^{\odot 2}, \\ \widehat{\mu}_{P,t} &= \frac{\mu_{P,t}}{1-\beta_1^t}, \qquad \widehat{\nu}_{P,t} = \frac{\nu_{P,t}}{1-\beta_2^t}, \end{aligned} \tag{48}$$

$$U_{P,t} = \frac{\widehat{\mu}_{P,t}}{\sqrt{\widehat{\nu}_{P,t}} + \epsilon_{\mathrm{Adam}}} + \lambda_P P, \qquad P^+ = P - \eta_t U_{P,t}. \tag{49}$$

Squares, roots, and division are elementwise; $\epsilon_{\mathrm{Adam}} > 0$, $0 \leq \beta_1, \beta_2 < 1$, and λ_P is the decoupled weight-decay coefficient. For the local formulas below, both factors use learning rate η_t and the pretrained weight, bias, and γ are fixed. Different factor learning rates can instead be handled using the actual increments $\Delta A = A^+ - A$ and $\Delta B = B^+ - B$.

Proposition A.13 (A local bound using actual optimizer directions)**.** *Fix the input x to the analyzed layer and disable adapter dropout in this local evaluation. Let $\delta z_{t,c}(x) = z_{t+1,c}(x) - z_{t,c}(x)$, and suppose κ_g is L_κ-Lipschitz on every segment between the compared channel scores. Then*

$$|\kappa_g(z_{t+1,c}) - \kappa_g(z_{t,c})| \leq L_\kappa |\delta z_{t,c}|. \tag{50}$$

Under Definition A.12, define

$$R^B_{t,c} = (U_{B,t} A)_{c,:}, \quad R^A_{t,c} = (B U_{A,t})_{c,:}, \quad R^{BA}_{t,c} = (U_{B,t} U_{A,t})_{c,:}. \tag{51}$$

The score change is exactly

$$\delta z_{t,c}(x) = -\eta_t \gamma (R^B_{t,c} + R^A_{t,c}) x + \eta_t^2 \gamma R^{BA}_{t,c} x. \tag{52}$$

Consequently,

$$\begin{aligned} \mathbb{E}_x |\kappa_g(z_{t+1,c}) - \kappa_g(z_{t,c})|^2 \leq 3\gamma^2 \eta_t^2 L_\kappa^2 \, \mathbb{E}_x \Big[\big(\|R^B_{t,c}\|_2^2 + \|R^A_{t,c}\|_2^2 \\ + \eta_t^2 \|R^{BA}_{t,c}\|_2^2 \big) \|x\|_2^2 \Big]. \end{aligned} \tag{53}$$

The expectation may be conditioned on the current training and optimizer state. The mask and scales remain inside the moment-dependent directions $U_{P,t}$; they generally cannot be factored out.

Interpretation of the optimizer-consistent bound. The bound relates actual optimizer directions to score and gate-homogeneity changes. Controller multipliers enter these directions through both moment buffers. A fixed positive coordinate multiplier applied to an entire prescribed gradient history cancels between Adam's first and second moments when $\epsilon_{\mathrm{Adam}} = 0$ and the second moment is nonzero. With finite ϵ_{Adam} the cancellation is approximate when the denominator is dominated by the second moment. Time-varying controls can change the moment history. Gradient multipliers therefore need not induce monotone changes in parameter-step magnitude; § F.5 derives the exact scalar response. The mask attenuates the current B-gradient input, and s_t scales the gradients supplied to AdamW, whose moment normalization determines the adaptive update.

The fixed-layer-input calculation isolates the direct adapter contribution. Keeping the same external probe examples across checkpoints does not keep each internal layer input fixed. For a changing layer input x^+ and a fixed bias, the exact score difference additionally contains $W^+_{\mathrm{eff}}(x^+ - x)$; § F.5 states this identity. Full-network probe drift must therefore include propagated input changes and cannot be attributed solely to the local gate adapter.

A.6 Temporal control as a design principle

Gate-state changes accumulate over optimization. Let $D(\theta; x)$ denote a scalar gate-derived margin evaluated at LoRA parameters θ; write $D_t(x) := D(\theta_t; x)$. The following elementary bound motivates controlling one-step movement without requiring gate states to remain fixed. In the SwiGLU three-region controller, the relevant margins are the distances to the two characteristic boundaries, not the aggregate block score a in Eq. (71).

Proposition A.14 (Margin-based flip bound). *For any threshold $\tau > 0$,*

$$\mathbb{P}[\operatorname{sign}(D_{t+1}(x)) \neq \operatorname{sign}(D_t(x))] \leq \mathbb{P}[|D_t(x)| \leq \tau] + \mathbb{P}[|D_{t+1}(x) - D_t(x)| > \tau]\,. \tag{54}$$

Corollary A.15 (SwiGLU three-region flip bound). *For SwiGLU threshold $\tau_z > 0$, define*

$$R_{\tau_z}(z) = \begin{cases} -1, & z < -\tau_z, \\ 0, & |z| \leq \tau_z, \\ +1, & z > \tau_z, \end{cases} \tag{55}$$

and the two boundary margins $D_t^- = z_t + \tau_z$ and $D_t^+ = z_t - \tau_z$. For any auxiliary margin width $\tau > 0$,

$$\begin{aligned} \mathbb{P}[R_{\tau_z}(z_{t+1}) \neq R_{\tau_z}(z_t)] \leq & \sum_{\nu \in \{-,+\}} \mathbb{P}(|D_t^\nu| \leq \tau) \\ & + 2\,\mathbb{P}(|z_{t+1} - z_t| > \tau). \end{aligned} \tag{56}$$

Thus regime changes require either proximity to one of the two characteristic boundaries or sufficiently large one-step score motion.

Corollary A.16 (Actual-step control of one-step movement). *Let θ_t collect the LoRA parameters affecting a fixed margin map $D(\theta; x)$, and write the actual optimizer step as $\theta_{t+1} = \theta_t - \eta_t U_t$. Assume $D(\cdot; x)$ is $L_D(x)$-Lipschitz on the entire update segment. For any $\tau > 0$,*

$$\mathbb{P}\big(|D_{t+1}(x) - D_t(x)| > \tau\big) \leq \frac{\eta_t^2}{\tau^2} \mathbb{E}\big[L_D(x)^2 \|U_t\|_2^2\big]. \tag{57}$$

For AdamW, U_t includes the moment-normalized directions and decoupled weight decay. No separate factor of s_t^2 or $m_{t,c}^2$ follows from pre-optimizer gradient modulation.

Together, these results connect nonlinear-state feedback to two measurable sources of temporal variation: proximity to characteristic boundaries and the size of the actual optimizer-induced score change. The quantile mask and smoothed macro scale turn the sensitivity and geometry signals into online controls, whose empirical contributions are examined separately in the ablations.

B Detailed Controller Construction

B.1 From theoretical implications to two control resolutions

Theoretical implications. The mismatch decomposition identifies the gate-selection and value pathways. Proposition A.7 and Corollary A.8 describe channel scaling and its output-level aggregation. The local metric identifies response derivatives as one component of functional sensitivity. Together, they suggest controlling the adapted FFN at two resolutions: selecting where to emphasize nonlinear reconfiguration, and coordinating how its coupled projections remain responsive as the gate regime changes. These implications motivate a family of geometry-guided controllers.

Intuition B.1 (Population context and history for micro control). *Intuition 4.1 interprets channel responsiveness in the context of the FFN's aggregated output. SiLU's κ_g diverges negatively in a tail where the absolute response derivative tends to zero. The micro controller therefore uses $|\psi'|$, linked to κ_g through $g' = \sigma\kappa_g$ in SwiGLU. It averages over samples and tokens, compares the resulting channel statistics against the population through a soft quantile, and smooths the mask assignments over time. Together, these operations supply a bounded, history-aware preference for response-sensitive score channels while retaining the task gradient's other factors. This design seeks more effective corrective state reconfiguration; § F.6 details its analytic motivation.*

Intuition B.2 (Regime feedback for the coupled decision unit). *Beyond SwiGLU's positive characteristic point, the relative nonlinear-reweighting share of the local derivative decreases (Corollary F.3). Intuition 4.2 interprets such states as increasingly established transmission, motivating reduced selection-focused modulation. This regime serves as a proxy for the current transmission mechanism. Eq.* (79) *shows that target agreement depends jointly on the gate, value, and output paths, motivating the use of a shared state signal to inform all three projections. We therefore coordinate* `up`*,* `gate`*, and* `down` *through positive, bounded gradient multipliers, allowing all three roles to remain subject to task supervision.*

Controller instantiation. Micro uses a soft quantile and EMA; Macro uses a difference of regime occupancies, projection-specific affine maps, clipping, and EMA. These inexpensive policies instantiate the implications above; the theory characterizes response and optimizer dependence but does not specify a unique optimal feedback law. Both act before AdamW and preserve the frozen backbone, uniform adapter rank, and LoRA forward computation. The mask-only specialization applies when the operative gate lacks the characteristic geometry needed for the present macro rule. Task comparisons, component ablations, and fixed-probe diagnostics evaluate the instantiated design at distinct levels.

B.2 TEMPORAL-IMPORTANCE CHANNEL MASK

Unified response-based optimization sensitivity. Following Eq. (24), we use

$$k_{\mathrm{opt}}(z) = |\psi'(z)| = |\rho(z) + z\rho'(z)| \tag{58}$$

as micro-level responsiveness. For SwiGLU, $\psi(z) = g(z) = \mathrm{SiLU}(z)$, so

$$k_{\mathrm{opt}}(z) = \big|\sigma(z) + z\sigma(z)(1-\sigma(z))\big|. \tag{59}$$

In a conventional single-branch FFN, $g = \rho$ and $\psi = z\rho$, so the definition retains gate and tied-value derivative paths. QuickGELU uses $\psi(z) = z\sigma(\beta_{\mathrm{QG}} z)$, whereas ReLU uses $\psi(z) = \max(z, 0)$.

For FFN pre-activations $z \in \mathbb{R}^{N_b \times S \times d_h}$ with minibatch size N_b, the channel score is

$$k_{\mathrm{eff}}(c) := \mathbb{E}_{b,s}[k_{\mathrm{opt}}(z_{b,s,c})]\,, \qquad c \in [d_h]. \tag{60}$$

Large $k_{\mathrm{eff}}(c)$ indicates a larger average intrinsic response derivative for channel c. Full-output and loss sensitivity additionally depend on the remaining factors in Eq. (41) and the loss gradient.

Soft top-k mask with temporal smoothing. Given a keep ratio $\rho_{\mathrm{keep}} \in (0, 1]$, let

$$\tau_k = \mathrm{Quantile}\big(k_{\mathrm{eff}}, 1 - \rho_{\mathrm{keep}}\big), \tag{61}$$

and define

$$m_c^{\mathrm{new}} = \sigma\big(\beta_m(k_{\mathrm{eff}}(c) - \tau_k)\big) \in (0, 1), \tag{62}$$

where $\beta_m > 0$ controls mask-boundary sharpness. We smooth the mask over optimization steps,

$$m_t = \eta_m m_{t-1} + (1 - \eta_m) m^{\mathrm{new}}, \qquad \eta_m \in [0, 1), \tag{63}$$

so relative channel weights retain mask history. Repeated substitution gives the exact expansion

$$m_t = \eta_m^t m_0 + (1 - \eta_m) \sum_{j=1}^{t} \eta_m^{t-j} m_j^{\mathrm{new}}. \tag{64}$$

Thus temporal memory acts on mask assignments after the nonlinear quantile-to-mask map. With $m_0 \in [0, 1]^{d_h}$, the update remains bounded in $[0, 1]^{d_h}$ by convexity. The sigmoid mask is soft: the keep ratio sets a quantile threshold rather than an exact number of active channels. These properties establish bounded, history-dependent gradient-input weights; their optimizer-dependent effect is analyzed in § A.5.

Gradient use. The mask is applied along the output-channel dimension of the LoRA factor associated with the projection that produces the relevant score z. For SwiGLU this is the `gate_proj`; for a conventional GELU/ReLU FFN it is the first FFN projection whose output is passed through the activation. With $B_z \in \mathbb{R}^{d_h \times r}$,

$$(\nabla B_z)_{c,:} \leftarrow m_{t,c}(\nabla B_z)_{c,:}. \tag{65}$$

This rowwise operation leaves the A-gradient unmasked. AdamW converts the modified gradient inputs into parameter increments as in Definition A.12. The control acts on FFN score projections, while attention projections retain standard LoRA.

B.3 HOMOGENEITY-CONDITIONED GRADIENT SCALING

Trajectory shaping from gate geometry. The temporal mask answers where response-sensitive score updates should be emphasized relative to other channels. Macro answers how to coordinate the three FFN projections as the gate population changes. Its regime partition follows from operative-gate geometry, while its modulation rule is a policy motivated by the declining relative nonlinear-reweighting contribution and the coupled output differential. Proposition A.13 and Corollary A.16 connect the resulting optimizer inputs to state movement through the actual, moment-dependent parameter step.

SwiGLU gate-regime statistic. The SwiGLU operative gate $g(z) = \mathrm{SiLU}(z)$ has homogeneity

$$\kappa_g(z) = 1 + z(1 - \sigma(z)). \tag{66}$$

The negative zero of κ_g and its positive stationary point occur at $-\tau_z$ and $+\tau_z$, respectively, where

$$\tau_z \sigma(\tau_z) = 1, \qquad \tau_z \approx 1.27846. \tag{67}$$

These two different characteristic points—a negative zero and a positive stationary point—motivate a finite nonzero symmetric *characteristic partition*:

$$p_{\mathrm{sup}} = \mathbb{E}[\mathbb{I}(z < -\tau_z)], \quad p_{\mathrm{res}} = \mathbb{E}[\mathbb{I}(|z| \le \tau_z)], \quad p_{\mathrm{pos}} = \mathbb{E}[\mathbb{I}(z > \tau_z)]. \tag{68}$$

The shape of κ_g distinguishes a negative suppressive regime ($\kappa_g < 0$), a central responsive regime, and a positive asymptotic regime beyond the positive maximum of κ_g. In the last region, gate effective homogeneity remains positive and decreases toward SiLU's linear-tail value of 1. The partition gives the controller an activation-specific summary of the current score distribution. Specifically, $z\sigma'(z)/g'(z) = (\kappa_g - 1)/\kappa_g$ decreases on $z > \tau_z$, while $g'(z)$ tends to 1. This *relative* decline of nonlinear reweighting motivates discounting positive-region occupancy. The central characteristic band contains scores with varying absolute sensitivity; Micro supplies that complementary derivative information.

GELU and ReLU boundary cases for macro control. For the QuickGELU used by the CLIP implementation, the operative gate is

$$g_{\mathrm{QGELU}}(z) = \sigma(\beta_{\mathrm{QG}} z), \qquad \beta_{\mathrm{QG}} = 1.702, \tag{69}$$

whose gate homogeneity is

$$\kappa_g(z) = \beta_{\mathrm{QG}} z\big(1 - \sigma(\beta_{\mathrm{QG}} z)\big). \tag{70}$$

This curve has its only finite zero at $z = 0$; on the positive side it has a stationary point at $z = \tau_z/\beta_{\mathrm{QG}}$ and then decays toward 0, while on the negative side it has no nonzero finite zero paired with that stationary point. Hence the *symmetric nonzero threshold* construction in Eq. (68) has no valid QuickGELU solution. Exact GELU, with $g(z) = \Phi(z)$, has the same relevant asymmetry: $\kappa_g(z) = z\varphi(z)/\Phi(z)$ has no nonzero negative zero that can form the SwiGLU-style symmetric pair.

For ReLU, the operative gate is $g(z) = \mathbb{I}(z > 0)$. The smooth ratio $\kappa_g = zg'/g$ is undefined on the negative half-line because $g(z) = 0$, equals 0 on the positive half-line, and cannot be smoothly extended through the discontinuity at $z = 0$. Its unique regime boundary is therefore the hard switch $z = 0$. If one inserts this boundary into the symmetric three-way construction, the responsive band collapses to zero width. Thus QuickGELU/GELU and ReLU are two degenerate boundary cases of the *specific macro statistic*: the former provides no admissible finite nonzero symmetric threshold, while the latter collapses to the hard threshold 0. In both cases we disable homogeneity-conditioned gradient scaling and retain only the temporal-importance mask.

We summarize the SwiGLU block state by

$$a = \mathrm{clamp}\big(p_{\mathrm{res}} - \lambda_{\mathrm{pos}} p_{\mathrm{pos}},\, 0,\, 1\big), \tag{71}$$

where $\lambda_{\mathrm{pos}} > 0$ discounts occupancy of the positive asymptotic region relative to the central responsive region. This score converts the geometry-derived partition into a bounded block descriptor, which the projection-specific mappings below translate into gradient multipliers. Expectations pool samples, tokens, and channels within one block. Each score contributes through a bounded regime indicator, so magnitudes within a regime carry the same weight; suppressive-tail scores contribute zero to both p_{res} and p_{pos}. The subtraction and penalty λ_{pos} encode the proposed balance between responsive-state emphasis and established-transmission discounting.

Projection-wise LoRA scaling for the coupled FFN decision unit. The regime statistic is measured from the gate score, while the macro controller modulates all three SwiGLU FFN adapters. The `gate_proj` determines the operative transmission coefficient, `up_proj` supplies its independent value, and `down_proj` maps the gated feature to the residual stream. We use the gate statistic as a shared block-level signal with projection-specific scale ranges. This block-level design coordinates the gate, value, and output roles using one measured state. At a fixed layer input, $z = W_g x$ depends directly on the gate projection, while the value and down projections contribute to the coupled FFN output. The current loss gradient supplies supervision to all three paths, and the controller adjusts their relative weighting before optimizer normalization. Positive scales preserve the signs of individual current-gradient coordinates but generally change the joint gradient direction; AdamW's moment history further determines the actual step.

For each FFN projection $\star \in \{\mathrm{up}, \mathrm{gate}, \mathrm{down}\}$ in a SwiGLU block, we compute

$$s_\star^{\mathrm{new}} = \mathrm{clip}\big(1 + \alpha_\star(2a-1),\ s_{\min}^\star,\ s_{\max}^\star\big), \tag{72}$$

and optionally smooth it by

$$s_{\star,t} = \beta_s s_{\star,t-1} + (1-\beta_s) s_\star^{\mathrm{new}}. \tag{73}$$

The scale is applied only to the corresponding LoRA gradients,

$$\nabla\theta_\star \leftarrow s_{\star,t} \nabla\theta_\star. \tag{74}$$

The frozen backbone and forward computation remain unchanged. The gate-derived score is shared across the three projections, while $\alpha_\star$ and the clipping ranges specify their different gradient modulation strengths. The resulting parameter increments depend on AdamW's accumulated moments and weight decay. § F.5 gives moment-dependent score-motion identities for its temporal effect.

Summary. The micro controller is architecture-general once k_{opt} is defined from the actual FFN response. The macro controller is deliberately activation-specific at the sensing level but block-level at the actuation level: the present three-way statistic is enabled for SwiGLU because its operative SiLU gate provides the required finite nonzero symmetric characteristic points, and the resulting gate state supplies a common signal for the `up`/`gate`/`down` gradient multipliers. For ReLU and GELU boundary cases, we disable macro control rather than impose an artificial shared threshold.

C Notation and Architecture-Level Gate Mapping

C.1 Internal weight, scalar response, and operative gate

Our ρ and ψ correspond to the weighting function ρ and activation ϕ in FleS (Cai et al., 2026b). The additional symbol g identifies the coefficient relative to an FFN's specified value path:

$$\psi(z) = z\rho(z), \qquad h = g(z) \odot v, \qquad y = W_d h. \tag{75}$$

Let $x \in \mathbb{R}^{d_{\mathrm{in}}}$ and $z, h, v \in \mathbb{R}^{d_h}$. The score projection is in $\mathbb{R}^{d_h \times d_{\mathrm{in}}}$, W_u has the same shape when present, and $W_d \in \mathbb{R}^{d_{\mathrm{out}} \times d_h}$. Scalar functions act channel-wise; ρ_{keep} is the unrelated scalar mask keep ratio, not the function $\rho(z)$.

Table 6: Architecture-aligned correspondence between the internal self-gating weight ρ, scalar response ψ, and operative gate g. SiLU/SwiGLU, GELU, and ReLU follow their original definitions (Elfwing et al., 2018; Shazeer, 2020; Hendrycks & Gimpel, 2016; Nair & Hinton, 2010); QuickGELU uses the CLIP implementation (Radford et al., 2021).

FFN	**Weight** $\rho(z)$	**Response** $\psi(z)$	**Gate** $g(z)$	**Value / channel**
SwiGLU	$\sigma(z)$	$z\sigma(z)$	$\psi(z)$	$v = u,\ h = \psi(z) \odot u$
QuickGELU	$\sigma(\beta_{\mathrm{QG}} z)$	$z\sigma(\beta_{\mathrm{QG}} z)$	$\rho(z)$	$v = z,\ h = \psi(z)$
Exact GELU	$\Phi(z)$	$z\Phi(z)$	$\rho(z)$	$v = z,\ h = \psi(z)$
ReLU	$\mathbb{I}(z > 0)$	$\max(z, 0)$	$\rho(z)$	$v = z,\ h = \psi(z)$

For SwiGLU, $z = W_g x + b_g$ and $u = W_u x + b_u$; conventional FFNs use $z = W_1 x + b_1$. Here $\beta_{\mathrm{QG}} = 1.702$, Φ is the standard normal CDF, and φ denotes its density. QuickGELU is not the tanh-based GELU approximation: the actual implemented ρ must be used when evaluating ψ' or κ_g. Exact GELU is an analytical comparison; the reported CLIP experiments use QuickGELU.

C.2 What is inherited from the decision-making interpretation

The source interpretation regards feature vectors as alternatives, input channels as criteria, filters as updatable reference patterns, z as an importance indication, and ρ as sign-aware response recalibration (Cai et al., 2026b). The grey-systems and multi-criteria references (Deng, 1982; Liu, 2025; **?**) provide the conceptual background for this reference-based evaluation and weighting interpretation. Our extra step is to identify the FFN transmission coefficient g relative to v. In SwiGLU that coefficient is signed and unbounded, so it should not be identified with a conventional nonnegative decision weight or with a probability.

FleS's prototype replaces $z\rho(z)$ by $\kappa_{\mathrm{ve}}\rho(\kappa_{\mathrm{ho}}z)z$ to modify forward recalibration. In contrast, NA-LoRA preserves ρ and the LoRA forward parameterization and uses $|\psi'|$ and κ_g only for training-time gradient modulation. We carry the score–weight–response lens into the distinct problem of fixed-rank adaptation. The local results are established using the explicit differentiability and bounded-derivative assumptions stated for the analyzed score segments.

C.3 Algebraic regrouping and value-path dependencies

For a SwiGLU channel, the regrouping $h = \sigma(z)(zu)$ is valid, but makes the new value zu depend on z. At fixed u, its derivative is $u[\sigma(z) + z\sigma'(z)]$, not $zu\sigma'(z)$ alone. Thus the unified self-gated representation remains valid, but changing the operative-gate/value split requires tracking its dependencies. The gate-level mismatch and regime geometry in this paper always use the original architecture-aligned split in Table 6.

D Proofs: Operative-Gate Selection Mismatch

D.1 Proof of Proposition A.2

Proof. By definition,

$$\mathcal{U}(z_{\mathrm{tar}}, v_{\mathrm{tar}}) - \mathcal{U}(z_{\mathrm{lora}}, v_{\mathrm{lora}}) = g(z_{\mathrm{tar}}) \odot v_{\mathrm{tar}} - g(z_{\mathrm{lora}}) \odot v_{\mathrm{lora}}.$$

Add and subtract $g(z_{\mathrm{tar}}) \odot v_{\mathrm{lora}}$:

$$= \big(g(z_{\mathrm{tar}}) - g(z_{\mathrm{lora}})\big) \odot v_{\mathrm{lora}} + g(z_{\mathrm{tar}}) \odot (v_{\mathrm{tar}} - v_{\mathrm{lora}}),$$

which is Eq. (20). □

Corollary D.1 (Propagation to the full FFN output)**.** *Let h_{tar} and h_{lora} denote the pre-output-projection FFN channels and let $y_{\mathrm{tar}} = W_{d,\mathrm{tar}}h_{\mathrm{tar}}$ and $y_{\mathrm{lora}} = W_{d,\mathrm{lora}}h_{\mathrm{lora}}$. With $\Delta h = h_{\mathrm{tar}} - h_{\mathrm{lora}}$ and $\Delta W_d = W_{d,\mathrm{tar}} - W_{d,\mathrm{lora}}$,*

$$y_{\mathrm{tar}} - y_{\mathrm{lora}} = W_{d,\mathrm{tar}}\Delta h + \Delta W_d\, h_{\mathrm{lora}}. \tag{76}$$

Hence, for any induced operator norm compatible with ℓ_p,

$$\|y_{\mathrm{tar}} - y_{\mathrm{lora}}\|_p \le \|W_{d,\mathrm{tar}}\|_{p\to p}\|\Delta h\|_p + \|\Delta W_d\|_{p\to p}\|h_{\mathrm{lora}}\|_p. \tag{77}$$

Thus the hidden selection mismatch identified in Proposition A.2 enters the complete FFN output through the target down projection, while a separate down-projection residual contributes additively. The statement is an upper-bound decomposition; it does not claim that every nonzero hidden mismatch must survive a particular down projection without cancellation.

Proof. Add and subtract $W_{d,\mathrm{tar}}h_{\mathrm{lora}}$ and then apply the induced-operator-norm inequality. □

D.2 Proof of Corollary A.3

Proof. Apply the triangle inequality to Eq. (20) and apply $\|a \odot b\|_p \le \|a\|_\infty \|b\|_p$ to each product. □

D.3 Proof of Proposition A.4

Proof. For each differentiable channel c, the mean-value theorem gives an intermediate point $\bar{z}_c$ between $z_{\mathrm{lora},c}$ and $z_{\mathrm{tar},c}$ such that

$$g(z_{\mathrm{tar},c}) - g(z_{\mathrm{lora},c}) = g'(\bar{z}_c)(z_{\mathrm{tar},c} - z_{\mathrm{lora},c}) = g'(\bar{z}_c)\Delta z_c.$$

Multiplying by $v_{\mathrm{lora},c}$ yields Eq. (22); squaring and summing yields Eq. (23). □

D.4 Why the bound is data-aware

Under bounded inputs and locally Lipschitz gates, sufficiently small parameter residuals also control functional change. The relevant weighting depends on the data-aligned score residual $E_{W_z}x + e_{b_z}$, local operative-gate sensitivity, and transmitted values. This data- and gate-state-dependent weighting gives Corollary A.3 its functional interpretation.

E Local Functional Metrics and Rank-Constrained Approximation

E.1 Proof of Proposition A.10

Proof. Apply Taylor's theorem componentwise at the common reference $z_0(x)$. For $S \in \{T, R\}$,

$$g(z_0 + \xi Sx) = g(z_0) + \xi \,\mathrm{diag}(g'(z_0))Sx + q_S, \qquad \|q_S\|_2 \le \frac{L_{g''}}{2}\xi^2\|Sx\|_2^2,$$

where $L_{g''}$ bounds $|g''|$ on the score segments. The remainder bound follows from $\|(Sx)^{\odot 2}\|_2 \le \|Sx\|_2^2$. Multiplying the difference of the two expansions by $W_d\mathrm{diag}(v_0)$ gives

$$y_T - y_R = \xi J_g(x)(T-R)x + q, \qquad \|q\|_2 \le \frac{L_{g''}}{2}\xi^2\|W_d\|_2\|v_0(x)\|_\infty\big(\|Tx\|_2^2 + \|Rx\|_2^2\big).$$

The stated boundedness assumptions make $\|q\|_2 = \mathcal{O}(\xi^2)$ and $\|J_g(x)Ex\|_2 = \mathcal{O}(1)$ uniformly in x for fixed matrices. Expanding the squared norm yields the leading term $\xi^2\|J_g(x)Ex\|_2^2$, a cross term of order ξ^3, and a squared remainder of order ξ^4. Taking expectations proves Eq. (43). A fixed finite empirical probe distribution is one setting satisfying the required boundedness when the gate is smooth on the specified segments. □

E.2 General quadratic form and the separable special case

Writing $\mathrm{vec}(E)$ by stacking the columns of E, the exact local quadratic form is

$$\mathcal{Q}(E) = \mathrm{vec}(E)^\top \mathcal{H}\,\mathrm{vec}(E), \qquad \mathcal{H} = \mathbb{E}_x\big[(xx^\top) \otimes (J_g(x)^\top J_g(x))\big] \succeq 0. \tag{78}$$

Neither channel diagonality nor a Kronecker factorization is assumed in this formula. If $J_g(x)^\top J_g(x) = D$ almost surely, then $\mathcal{H} = \Sigma \otimes D$ and

$$\mathcal{Q}(E) = \mathrm{tr}(DE\Sigma E^\top) = \|D^{1/2}E\Sigma^{1/2}\|_F^2.$$

Replacing $J_g(x)^\top J_g(x)$ by its expectation does not generally yield this identity, because the Gram matrix can correlate with $xx^\top$. This separable case gives an exactly solvable conditional model.

E.3 Proof of Proposition A.11

Proof. For every feasible R, the transformed matrix $X = D^{1/2}R\Sigma^{1/2}$ has rank at most r. The classical truncated-SVD approximation theorem (Eckart & Young, 1936) therefore gives

$$\|C - X\|_F^2 \ge \sum_{i>r}\sigma_i(C)^2.$$

A truncated SVD C_r is attainable even for singular D or Σ. Let $P_D = D^{1/2}D^{\dagger/2}$ and $P_\Sigma = \Sigma^{\dagger/2}\Sigma^{1/2}$ be the orthogonal projectors onto their respective ranges. Since $C = D^{1/2}T\Sigma^{1/2}$, every

nonzero left singular vector of C lies in the range of D, and every nonzero right singular vector lies in the range of Σ. Thus $P_D C_r P_\Sigma = C_r$. For $R_\star = D^{\dagger/2} C_r \Sigma^{\dagger/2}$ we have $\mathrm{rank}(R_\star) \le r$ and

$$D^{1/2} R_\star \Sigma^{1/2} = P_D C_r P_\Sigma = C_r.$$

This attains the lower bound. Null-space components may make the optimizer nonunique, which does not change the minimum value. □

E.4 Counterexample and architectural scope

For the example in the main text, $C = \mathrm{diag}(2, 3)$, so its leading singular component retains the second coordinate. Direct substitution gives

$$\|T - R_W\|_F^2 = 1, \quad \mathcal{Q}(T - R_W) = 9, \qquad \|T - R_\star\|_F^2 = 4, \quad \mathcal{Q}(T - R_\star) = 4.$$

One local independent-value gated-layer realization sets $z_0 = 0$, $g = \mathrm{SiLU}$, $v_0 = (2, 2)^\top$, $W_d = \mathrm{diag}(1, 3)$, and uses bounded inputs with covariance I_2. Since $g'(0) = 1/2$, this realizes $J_g = W_d$ and $D = \mathrm{diag}(1, 9)$. A fixed affine value branch can supply this v_0. This realization belongs to the general affine gated-layer model; bias-free Llama layers retain the input-dependent value branch in the general analysis. For sufficiently small $\xi > 0$, the strict ordering $4\xi^2 + \mathcal{O}(\xi^3) < 9\xi^2 + \mathcal{O}(\xi^3)$ persists. Both functional errors still tend to zero with the perturbation scale.

For conventional self-gated FFNs the value is tied to the score, so the full channel derivative is $(g(z)z)' = g(z) + zg'(z)$. Substituting $J_\psi = W_d \mathrm{diag}(\psi'(z_0))$ gives the analogous local quadratic form. For ReLU this statement is restricted to segments contained in one smooth region; it does not differentiate the hard gate across zero.

With joint SwiGLU gate, value, and output-projection perturbations, the local differential is

$$\mathrm{d}y = W_d \mathrm{diag}(v_0 \odot g'(z_0))\,\mathrm{d}z + W_d \mathrm{diag}(g(z_0))\,\mathrm{d}v + (\mathrm{d}W_d)\,[g(z_0) \odot v_0]. \tag{79}$$

The squared discrepancy contains interactions among these terms. This joint differential makes explicit the value and downstream contributions that complement the score-path geometry.

F Gate Homogeneity, Characteristic Points, and Practical Sensitivity

F.1 Proof of Lemma A.6

Proof. Because g is twice continuously differentiable near z, Taylor's theorem gives

$$g(z + \delta z) = g(z) + g'(z)\delta z + \mathcal{O}(\delta z^2).$$

Divide by $g(z) \neq 0$ and use $g'(z)/g(z) = \kappa_g(z)/z$ for $z \neq 0$ to obtain Eq. (33); multiply by v for Eq. (34). □

F.2 Proof of Proposition A.7

Proof. Fix a channel state (z, v) and define the radial path from Proposition A.7,

$$q(\alpha) = h_c(\alpha) = g(\alpha z)(\alpha v).$$

Then

$$\frac{\partial q(\alpha)}{\partial \alpha} = z g'(\alpha z)(\alpha v) + g(\alpha z) v.$$

Therefore, wherever the logarithmic derivative is defined,

$$\frac{\alpha}{q(\alpha)} \frac{\partial q(\alpha)}{\partial \alpha}\bigg|_{\alpha=1} = 1 + \frac{z g'(z)}{g(z)} = 1 + \kappa_g(z) = \chi_{\mathrm{FFN}}(z).$$

If both score and value branches are bias-free linear maps of x, then $z(\alpha x) = \alpha z(x)$ and $v(\alpha x) = \alpha v(x)$, so this radial state path is induced by input scaling. With affine biases, the proposition concerns local score–value radial geometry, not literal input homogeneity. □

Proof of Corollary A.8 and the output part of Proposition 3.1. Differentiating the vector hidden response componentwise yields

$$h_c'(1) = v_c\big[g(z_c) + z_c g'(z_c)\big] = h_c\,[1 + \kappa_g(z_c)].$$

Multiplication by the fixed linear map W_d gives Eq. (37). At a removable zero of g, the expression is interpreted through the continuous product $v_c[g(z_c) + z_c g'(z_c)]$. The logarithmic expression is not used when $h_c = 0$. Different channel coefficients and possible output-space cancellation prevent replacing this vector sum by an unweighted average elasticity in general.

F.3 Activation-specific gate geometry

The macro regime statistic is derived from the operative gate g, not from the full activation unless the full activation is itself the FFN gate. The next proposition specifies the three experimental cases.

Proposition F.1 (Gate-specific characteristic geometry)**.** *Let $\tau_\star > 0$ be the unique positive solution of*

$$\tau_\star \sigma(\tau_\star) = 1, \tag{80}$$

so $\tau_\star \approx 1.2784645$.

1. *For a SwiGLU operative gate $g(z) = \mathrm{SiLU}(z) = z\sigma(z)$,*

$$\kappa_g(z) = 1 + z(1 - \sigma(z)). \tag{81}$$

   Its unique negative zero is $-\tau_\star$, and its unique positive stationary point is $+\tau_\star$. Hence these two different characteristic points motivate the finite nonzero symmetric characteristic threshold $\tau_z = \tau_\star$ used by our controller.

2. *For a QuickGELU operative gate $g(z) = \sigma(\beta_{\mathrm{QG}} z)$ with $\beta_{\mathrm{QG}} > 0$,*

$$\kappa_g(z) = \beta_{\mathrm{QG}} z\big(1 - \sigma(\beta_{\mathrm{QG}} z)\big). \tag{82}$$

   Its only finite zero is $z = 0$. Its positive stationary point is $z = \tau_\star/\beta_{\mathrm{QG}}$, with no paired nonzero negative zero. Thus QuickGELU has no SwiGLU-style finite nonzero symmetric *threshold.*

3. *For the ReLU operative gate $g(z) = \mathbb{I}(z > 0)$, κ_g is undefined on $z < 0$ because the gate itself is zero, equals 0 on $z > 0$, and cannot be smoothly extended across the discontinuity at 0. The unique switching boundary is $z = 0$, so the symmetric responsive band collapses to zero width.*

Proof. For SwiGLU, differentiating Eq. (81) gives

$$\kappa_g'(z) = (1 - \sigma(z))(1 - z\sigma(z)).$$

Let $f(z) = z\sigma(z)$ for $z > 0$. Since

$$f'(z) = \sigma(z)\big(1 + z(1 - \sigma(z))\big) > 0,$$

f is strictly increasing on $(0, \infty)$, with $f(0) = 0$ and $f(z) \to \infty$ as $z \to \infty$. Hence $f(z) = 1$ has a unique positive solution $\tau_\star$. The positive stationary point therefore satisfies $z\sigma(z) = 1$, giving the unique point $z = \tau_\star$. For a negative zero, write $z = -a$ with $a > 0$:

$$0 = 1 - a(1 - \sigma(-a)) = 1 - a\sigma(a),$$

which is the same equation $a\sigma(a) = 1$, hence $z = -\tau_\star$.

For QuickGELU, Eq. (82) follows directly from $g'(z) = \beta_{\mathrm{QG}}\sigma(\beta_{\mathrm{QG}} z)(1 - \sigma(\beta_{\mathrm{QG}} z))$. Since $0 < \sigma(\beta_{\mathrm{QG}} z) < 1$ at every finite z, $\kappa_g(z) = 0$ implies $z = 0$. Moreover,

$$\kappa_g'(z) = \beta_{\mathrm{QG}}(1 - \sigma(\beta_{\mathrm{QG}} z))\big(1 - \beta_{\mathrm{QG}} z\sigma(\beta_{\mathrm{QG}} z)\big),$$

so the positive stationary point satisfies $(\beta_{\mathrm{QG}} z)\sigma(\beta_{\mathrm{QG}} z) = 1$, i.e., $z = \tau_\star/\beta_{\mathrm{QG}}$. No nonzero negative root exists. The ReLU claim follows from its piecewise-constant gate: $g = 0$ on the negative half-line, $g = 1$ on the positive half-line, and the only switching boundary is zero. □

Remark F.2 (Exact GELU). For exact GELU, $g(z) = \Phi(z)$ and

$$\kappa_g(z) = \frac{z\varphi(z)}{\Phi(z)}. \tag{83}$$

Because $\Phi(z) > 0$ and $\varphi(z) > 0$ for all finite z, the only finite zero is again $z = 0$. Thus exact GELU has the same property relevant to our macro controller: it does not provide the nonzero negative characteristic point required to form the SwiGLU-style symmetric pair. We therefore do not infer GELU thresholds by rescaling SiLU characteristic points.

Corollary F.3 (Declining relative nonlinear contribution in the positive tail)**.** *For the SwiGLU gate $g(z) = z\sigma(z)$ and $z > \tau_\star$, define the score derivative's relative nonlinear-reweighting contribution by*

$$\omega(z) := \frac{z\sigma'(z)}{g'(z)} = \frac{\kappa_g(z) - 1}{\kappa_g(z)}. \tag{84}$$

Then $\omega(z) > 0$, $\omega'(z) < 0$, and $\omega(z) \to 0$ as $z \to +\infty$. Meanwhile $g'(z) \to 1$, $\kappa_g(z) \to 1$, and the joint score–value elasticity $\chi_{\mathrm{FFN}}(z) \to 2$.

Proof. The identity $g'(z) = \sigma(z) + z\sigma'(z) = \sigma(z)\kappa_g(z)$ and $\kappa_g(z) - 1 = z(1 - \sigma(z))$ give Eq. (84). For $z > \tau_\star$, $\kappa_g(z) > 1$ and Proposition F.1 gives $\kappa_g'(z) < 0$. Hence

$$\omega'(z) = \frac{\kappa_g'(z)}{\kappa_g(z)^2} < 0.$$

Since $z(1 - \sigma(z)) \to 0$ and $z\sigma'(z) \to 0$ as $z \to +\infty$, the stated limits follow. This is a statement about the *relative share* of the nonlinear term; it does not assert that $z\sigma'(z)$ decreases immediately after $\tau_\star$, or that the full task gradient vanishes. □

Interpretation of the three SwiGLU regions. For $z < -\tau_\star$, $\kappa_g(z) < 0$, corresponding to a suppressive negative-tail geometry. On $[-\tau_\star, \tau_\star]$, κ_g is nonnegative and increases toward its positive maximum, yielding the central responsive region used by the controller. These characteristic points give the controller its geometry-based partition. For $z > \tau_\star$, κ_g decreases toward 1, the degree-one linear-tail value of SiLU. We therefore call this the *positive asymptotic* region rather than claiming that the gate is non-amplifying there. Corollary F.3 gives the precise gradient-contribution interpretation: the nonlinear part accounts for a decreasing fraction of the local derivative. Macro consequently discounts this occupancy as a design proxy for increasingly established transmission. A target may nonetheless require a different gate, value, or output map; the regime cannot certify successful escape from mismatch. Likewise, a larger drift $|\Delta\kappa_g|$ is movement of a response descriptor, not an objective whose increase proves better adaptation.

This establishes the boundary-case interpretation used in the experiments: ReLU compresses the transition to a hard threshold at 0, whereas GELU/QuickGELU does not admit the finite nonzero symmetric threshold required by Eq. (68). Both boundary cases select mask-only control.

F.4 Proof of Proposition A.13

Proof. Using the actual directions from Definition A.12,

$$B^+A^+ - BA = -\eta_t(U_{B,t}A + BU_{A,t}) + \eta_t^2 U_{B,t}U_{A,t}.$$

Multiplication by the fixed LoRA coefficient γ and fixed input x, followed by restriction to row c, gives Eq. (52). Cauchy–Schwarz and $(a + b + c)^2 \le 3(a^2 + b^2 + c^2)$ imply

$$|\delta z_{t,c}(x)|^2 \le 3\gamma^2\eta_t^2\big(\|R_{t,c}^B\|_2^2 + \|R_{t,c}^A\|_2^2 + \eta_t^2\|R_{t,c}^{BA}\|_2^2\big)\|x\|_2^2.$$

Squaring the Lipschitz inequality in Eq. (50) and taking expectations yields Eq. (53). Weight decay is included in $U_{P,t}$. The multipliers enter through the full moment-normalized directions. □

F.5 AdamW response to a gradient multiplier

Consider one coordinate with fixed current raw gradient q and fixed previous moment buffers μ_{t-1}, ν_{t-1}. Let $c > 0$ multiply only the current gradient supplied to Adam. Define

$$a = \beta_1 \mu_{t-1}, \quad b = (1-\beta_1)q, \quad d = \beta_2 \nu_{t-1}, \quad e = (1-\beta_2)q^2,$$

$$K_t = \frac{\sqrt{1-\beta_2^t}}{1-\beta_1^t}, \qquad \varepsilon_t = \epsilon_{\text{Adam}}\sqrt{1-\beta_2^t}.$$

Here a, b, d, e are local algebraic abbreviations, unrelated to the macro regime score. The bias-corrected adaptive direction, excluding decoupled weight decay, is exactly

$$u_t(c) = K_t \frac{a + bc}{\sqrt{d + ec^2} + \varepsilon_t}. \tag{85}$$

For $d + ec^2 > 0$, differentiation gives

$$u_t'(c) = K_t \frac{bd - aec + b\varepsilon_t\sqrt{d+ec^2}}{\sqrt{d+ec^2}\big(\sqrt{d+ec^2} + \varepsilon_t\big)^2}. \tag{86}$$

In the idealized $\epsilon_{\text{Adam}} = 0$ case this reduces to

$$u_t'(c) = K_t \frac{bd - aec}{(d+ec^2)^{3/2}}.$$

Thus the signed response, and even its magnitude when the direction remains positive, need not increase with c. For example, take $\beta_1 = 0.9$, $\beta_2 = 0.999$, an initial gradient 1 from zero moments, and a second raw gradient $q = 1$. At the second step, $a = 0.09$, $b = 0.1$, $d = 0.000999$, $e = 0.001$. With zero epsilon the positive direction increases for $c < 1.11$ and decreases for $c > 1.11$. The derivative formula describes the finite-epsilon correction directly. Weight decay adds $\lambda_P P$ to this direction and is independent of the current c when the parameter is fixed.

There is a separate invariance for rescaling an entire prescribed gradient history. If every gradient of a coordinate is multiplied by the same constant $c > 0$ from zero moments, induction gives

$$\mu_t^{(c)} = c\mu_t, \qquad \nu_t^{(c)} = c^2\nu_t, \qquad \frac{\widehat{\mu}_t^{(c)}}{\sqrt{\widehat{\nu}_t^{(c)}} + \epsilon_{\text{Adam}}} = \frac{\widehat{\mu}_t}{\sqrt{\widehat{\nu}_t} + \epsilon_{\text{Adam}}/c}.$$

For zero epsilon and nonzero second moments, the adaptive direction is invariant. For a time-varying coefficient, the two buffers instead accumulate $c_k q_k$ and $c_k^2 q_k^2$ with different decay rates, so the cancellation generally fails. These identities expose the history dependence of gradient feedback: accumulated momentum contributes alongside the current masked gradient, and subsequent gradients depend on the evolving parameters.

Actual increments and changing internal inputs. For arbitrary factor learning rates, set $\Delta B = B^+ - B$ and $\Delta A = A^+ - A$. Then

$$W_{\text{eff}}^+ - W_{\text{eff}} = \gamma\big(\Delta B A + B\Delta A + \Delta B \Delta A\big).$$

With a fixed bias and possibly changing internal input, the exact score difference is

$$z^+(x^+) - z(x) = \gamma\big(\Delta B A + B\Delta A + \Delta B\Delta A\big)x + W_{\text{eff}}^+(x^+ - x). \tag{87}$$

If a bias is also updated, its increment must be added. The local gate-adapter bound uses the first term with $x^+ = x$; checkpoint diagnostics on fixed external examples generally include both terms. With a fixed adapter-dropout mask, the direct adapter term instead uses the corresponding masked adapter input; Eq. (87) is stated for dropout-disabled evaluation.

F.6 Response sensitivity for temporal channel control

The temporal mask senses the actual scalar response $\psi = z\rho$, whereas the macro statistic describes the specified operative gate g. Differentiation gives

$$\psi'(z) = \rho(z) + z\rho'(z). \tag{88}$$

For SwiGLU, $h = \psi(z) \odot u$ and $g = \psi$. With input and value projection fixed,

$$\nabla_{W_g} \mathcal{L} = (q \odot u \odot \psi'(z))x^\top, \tag{89}$$

where $q = \partial \mathcal{L}/\partial h$. For conventional self-gated FFNs, $h = \psi(z)$ and $g = \rho$, so

$$\nabla_{W_1} \mathcal{L} = (q \odot \psi'(z))x^\top, \qquad \psi'(z) = g(z) + zg'(z). \tag{90}$$

Both use $|\psi'|$; the differing formulas in terms of g arise from the independent versus tied value path. For example, on ReLU's positive half-line $g' = 0$ but $\psi' = 1$: using the gate derivative alone would omit the entire tied-value contribution. At zero, smooth differentiation is not invoked.

These chain rules are exact and require no large-score approximation or saturation argument. As in FleS's self-gated factorization (Cai et al., 2026b), the response derivative contains both ρ and $z\rho'$; NA-LoRA retains both terms. The statistic isolates the architecture-local derivative from u, q, and input factors, which remain present in the back-propagated gradient. Operative-gate homogeneity $\kappa_g = zg'/g$ remains a separate relative-response quantity, related to the response elasticity by Eq. (32).

Bounded response sensitivity and population-relative scoring. For SiLU, $\kappa_g(z) = 1 + z(1 - \sigma(z)) \to -\infty$ as $z \to -\infty$, but $|g'(z)| \to 0$. More generally, the derivative is uniformly bounded because

$$|z|\sigma(z)(1 - \sigma(z)) \le |z|e^{-|z|} \le e^{-1}, \qquad |g'(z)| \le 1 + e^{-1}.$$

The first inequality follows from $\sigma(z)(1-\sigma(z)) = e^{-|z|}/(1+e^{-|z|})^2$; the second uses the maximum of te^{-t} on $t \ge 0$. Thus a large raw score or homogeneity does not generate an unbounded micro responsiveness statistic. Population-relative quantile comparison and the mask-history expansion in Eq. (64) then supply channel context and temporal memory. This is an analytic motivation for the existing controller, not an additional measured outlier-removal result.

G Margin-Based Temporal Bounds

This appendix relates actual parameter motion to one-step margin movement. The bounds identify boundary proximity and the full optimizer-induced update as the quantities governing local gate-state transitions, complementing the trajectory diagnostics.

Remark G.1 (Local sign preservation away from the boundary). For continuous decision scores with margins bounded away from zero, sufficiently small perturbations cannot change the sign.

G.1 Proof of Proposition A.14

Proof. If $\operatorname{sign}(D_{t+1}(x)) \ne \operatorname{sign}(D_t(x))$, then either $|D_t(x)| \le \tau$ or $|D_{t+1}(x) - D_t(x)| > \tau$. Otherwise, if both $|D_t(x)| > \tau$ and $|D_{t+1}(x) - D_t(x)| \le \tau$, then $D_{t+1}(x)$ cannot cross zero and must keep the same sign as $D_t(x)$. Taking probabilities yields Eq. (54). □

G.2 Proof of Corollary A.15

Proof. If $R_{\tau_z}(z_{t+1}) \ne R_{\tau_z}(z_t)$, then at least one boundary indicator changes: either the trajectory crosses the lower boundary $-\tau_z$, which is a sign change of $D_t^- = z_t + \tau_z$, or it crosses the upper boundary $+\tau_z$, which is a sign change of $D_t^+ = z_t - \tau_z$. A union bound and Proposition A.14 give

$$\mathbb{P}[R_{\tau_z}(z_{t+1}) \ne R_{\tau_z}(z_t)] \le \sum_{\nu \in \{-,+\}} \left(\mathbb{P}(|D_t^\nu| \le \tau) + \mathbb{P}(|D_{t+1}^\nu - D_t^\nu| > \tau)\right).$$

Since $D_{t+1}^\nu - D_t^\nu = z_{t+1} - z_t$ for both boundaries, the two movement terms are identical, giving Eq. (56). □

G.3 Proof of Corollary A.16

Proof. By Lipschitzness of the fixed margin map on the full update segment,

$$|D_{t+1}(x) - D_t(x)| \le L_D(x)\|\theta_{t+1} - \theta_t\|_2 = \eta_t L_D(x)\|U_t\|_2.$$

Markov's inequality applied to the squared nonnegative variable $\eta_t^2 L_D(x)^2 \|U_t\|_2^2$ gives Eq. (57). The actual direction U_t may depend on the controller, data, and optimizer history; no independence is required. The bound alone does not determine how changing a gradient multiplier changes U_t. □

H Implementation Details

H.1 NA-LoRA Training-Time Procedure

Algorithm 1 summarizes the full training-time procedure for the non-degenerate SwiGLU case. The micro controller extends to other FFNs by replacing k_{opt} with the architecture-specific response sensitivity in Eq. (25). When the operative gate lacks the finite nonzero symmetric threshold required by macro control, the regime-statistic and scaling lines are skipped and all $s_\star$ are set to 1.

Algorithm 1 NA-LoRA training-time procedure.

Input: FFN-block gate pre-activations z_t, keep ratio ρ_{keep}, mask sharpness β_m, mask EMA η_m, activation-specific regime specification τ_z when available (code argument `t`), positive-region penalty λ_{pos}, gradient-scale parameters $\{\alpha_\star, s^\star_{\min}, s^\star_{\max}\}_{\star\in\{\text{up,gate,down}\}}$, and scale EMA β_s.
State: mask buffer m_{t-1} and scale buffers $\{s_{\star,t-1}\}$.
(1) Temporal-importance channel mask
Compute channel scores $k_{\text{eff}}(c) = \mathbb{E}_{b,s}[k_{\text{opt}}(z_{t,b,s,c})]$ for $c \in [d_h]$.
Set $\tau_k \leftarrow \text{Quantile}(k_{\text{eff}}, 1-\rho_{\text{keep}})$.
$m_c^{\text{new}} \leftarrow \sigma\big(\beta_m(k_{\text{eff}}(c) - \tau_k)\big)$ for all c.
Update EMA mask: $m_t \leftarrow \eta_m m_{t-1} + (1-\eta_m)m^{\text{new}}$.
(2) Regime-aware statistics
$p_{\text{sup}} \leftarrow \mathbb{E}[\mathbb{I}(z_t < -\tau_z)]$; $p_{\text{res}} \leftarrow \mathbb{E}[\mathbb{I}(|z_t| \le \tau_z)]$; $p_{\text{pos}} \leftarrow \mathbb{E}[\mathbb{I}(z_t > \tau_z)]$.
$a \leftarrow \text{clamp}(p_{\text{res}} - \lambda_{\text{pos}} p_{\text{pos}}, 0, 1)$.
(3) Projection-wise LoRA gradient scaling
for $\star \in \{\text{up}, \text{gate}, \text{down}\}$ **do**
  $s_\star^{\text{new}} \leftarrow \text{clip}(1 + \alpha_\star(2a-1),\ s^\star_{\min},\ s^\star_{\max})$.
  $s_{\star,t} \leftarrow \beta_s s_{\star,t-1} + (1-\beta_s)s_\star^{\text{new}}$.
  Apply LoRA-only scaling: $\nabla\theta_\star \leftarrow s_{\star,t}\nabla\theta_\star$.
end for
(4) Mask application
Apply m_t to the output-channel rows of the score-producing LoRA B-gradient; A is not masked.
Output: modulated LoRA gradients, supplied to the standard AdamW update.
AdamW accumulates their first/second moments and applies bias correction, normalization, and decoupled weight decay as in Eqs. (48)–(49).

This section supplements Algorithm 1 with specific implementation details for reproducibility.

H.2 Hook Locations and Gradient Modification

NA-LoRA uses PyTorch forward/backward hooks, requiring no custom optimizer.

1. Mask Application (Micro). The mask m_t (Eq. (63)) is calculated during the forward pass of the FFN block and is applied to the **LoRA B-gradient along the output-channel dimension of the projection that produces** z. For SwiGLU this projection is `gate_proj`; for conventional GELU/ReLU FFNs it is the first FFN projection before the activation. Let the relevant adapter be $W + \gamma BA$, with fixed LoRA coefficient γ, and $B \in \mathbb{R}^{d_h \times r}$ and $A \in \mathbb{R}^{r\times d}$.

$$\text{Hook on } \nabla B_z: \quad (\nabla B_z)_{c,:} \leftarrow m_{t,c}(\nabla B_z)_{c,:}.$$

With LoRA factors stored in transposed layout, the same mask applies along the corresponding output-channel dimension. The forward gate and pretrained weights are unchanged.

2. Gradient Scaling (Macro). The scalars $s_{\text{up}}, s_{\text{gate}}, s_{\text{down}}$ are computed from the gate-regime statistics of the *same forward pass*, but they actuate the FFN as a coupled decision unit. The gate projection sets the selection coefficient, the up projection supplies a value, and the down projection returns the gated value to the residual stream; consequently, the gate state is used as a shared block-level gradient-modulation signal with projection-specific scale ranges. They are applied to the gradients of **both A and B matrices** of the corresponding LoRA modules.

$$\nabla A_\star \leftarrow s_\star \cdot \nabla A_\star, \quad \nabla B_\star \leftarrow s_\star \cdot \nabla B_\star.$$

H.3 HYPERPARAMETERS

We report the default hyperparameters used in the Llama experiments, which correspond directly to the `NALoRAController` implementation. The parameters are grouped into (i) the **Micro-level Mask**, which governs channel-selection sparsity and temporal smoothness, and (ii) the **Macro-level Gradient Scaling**, which governs regime-based gradient modulation. Table 7 lists the corresponding pre-optimizer gradient controls.

Notable settings include a high mask sharpness ($\beta_m = 20$) to approximate a binary selection mask, and distinct scaling bounds ($s_{\min}, s_{\max}$) for the `gate`, `up`, and `down` projections to respect their different roles in the FFN mechanism. Table 7 lists the default values used for Llama experiments.

Table 7: Default NA-LoRA hyperparameters for Llama experiments.

Component	Parameter / Symbol	Code Arg	Value
Micro: Temporal-Importance Mask			
Selection	Keep Ratio (ρ_{keep})	`gate_keep_ratio`	0.30
	Sharpness (β_m)	`gate_beta`	20.0
	Mask EMA (η_m)	`gate_ema`	0.90
Macro: Homogeneity-Conditioned Gradient Scaling			
Regime Stats	Regime Threshold (τ_z)	`t`	$1.27846^{\dagger}$
	Positive-region Penalty (λ_{pos})	`lambda_ov`	1.0
	Scale EMA (β_s)	`beta_s`	0.95
Projection-Specific Sensitivity & Bounds			
`gate_proj`	Sensitivity (α_{gate})	`alpha_gate`	0.40
	Range $[s_{\min}, s_{\max}]$	`smin/max_gate`	$[0.80, 1.50]$
`up_proj`	Sensitivity (α_{up})	`alpha_up`	0.30
	Range $[s_{\min}, s_{\max}]$	`smin/max_up`	$[0.80, 1.40]$
`down_proj`	Sensitivity (α_{down})	`alpha_down`	0.20
	Range $[s_{\min}, s_{\max}]$	`smin/max_down`	$[0.85, 1.30]$

† The threshold is specific to the SwiGLU operative gate $g(z) = \text{SiLU}(z)$ and is derived from $\kappa_g(z) = zg'(z)/g(z)$; see § F.3. For ReLU-FFN T5 the boundary collapses to 0, while QuickGELU-based CLIP does not admit the same finite nonzero symmetric threshold; in both cases macro gradient scaling is disabled. The code uses `t`; the paper uses τ_z to avoid confusion with the optimization step index.

H.4 ADDITIONAL TRAINING STATE AND INFERENCE COST

NA-LoRA stores one channel-mask vector and, when Macro is active, three projection-wise scalars per FFN block. Its additional persistent state scales as $\mathcal{O}(d_h)$. Computing response-sensitivity and regime statistics scans the gate-score tensor at cost $\mathcal{O}(N_b S d_h)$. These statistics and gradient multipliers are training-only; the resulting LoRA factors merge into the base projections for inference. These costs describe storage and arithmetic; wall-clock overhead depends on implementation.

I EXPERIMENTAL DETAILS

I.1 TASKS AND EVALUATION METRICS

We use the benchmark families established in LoRA-GA, LoRA-Pro, and GoRA (Wang et al., 2024; 2025; He et al., 2025): natural language generation (NLG) with Llama models, natural language understanding (NLU) with T5-Base (Raffel et al., 2020), and image-classification transfer with CLIP-ViT-B/16 (Radford et al., 2021).

We fine-tune the pretrained Llama-2-7B (Touvron et al., 2023) and Llama-3.1-8B (Grattafiori et al., 2024) checkpoints, rather than their chat- or instruction-tuned variants, on the same three NLG task pairs below. For mathematical reasoning and code generation, each task's training and evaluation

recipe is shared across the two backbones, except that each backbone uses its corresponding pre-trained tokenizer. The shared all-linear adapter scope and optimization settings are specified in § I.2.

1. **Mathematical reasoning:** We fine-tune on 100K MetaMathQA examples (Yu et al., 2024) and evaluate on GSM8K (Cobbe et al., 2021), reporting exact-answer accuracy for both backbones.
2. **Code generation:** We fine-tune on 100K Code-Feedback examples (Zheng et al., 2024) with code-only supervision and report HumanEval (Chen et al., 2021) pass@1 for both backbones.
3. **Conversational instruction following:** We fine-tune on a 52K WizardLM subset (Xu et al., 2024) and evaluate on MT-Bench (Zheng et al., 2023), reporting the first-turn score with the backbone-specific judge described below.

MT-Bench judge conventions. For Llama-2-7B, we follow LoRA-GA and LoRA-Pro (Wang et al., 2024; 2025) and use GPT-4 (OpenAI et al., 2023). For Llama-3.1-8B, we use GPT-4o (OpenAI, 2024), motivated by its use in related evaluations (He et al., 2025; Yu et al., 2025) and the judge endpoint available when the experiments were conducted. For Llama-3.1-8B, the reported metric is the GPT-4o first-turn score, not an average across judges. The two backbones therefore share task definitions, but not an identical MT-Bench judge; their dialogue scores are compared within each backbone's evaluation setting, not across judge conventions.

I.2 Adapters, optimization, and rank configurations

Both Llama-2-7B and Llama-3.1-8B use *all-linear* LoRA by default: adapters are attached to the attention projections `q_proj`, `k_proj`, `v_proj`, and `o_proj`, and the FFN projections `gate_proj`, `up_proj`, and `down_proj`. Embeddings, normalization layers, and output heads are excluded from adapter targeting. The target-module ablation in Table 5 is the explicit exception; full fine-tuning is a separate reference and is not an adapter-target configuration. For attention projections, NA-LoRA uses vanilla LoRA; its response-sensitive mask and homogeneity-conditioned gradient scales operate on FFN adapters. For T5 and CLIP, the architecture-specific controller settings follow § B.3. Both controls supply gradients to the standard AdamW moment update, leaving the base learning-rate schedule unchanged. Unless otherwise stated, Llama NA-LoRA runs use AdamW (Loshchilov & Hutter, 2019) with learning rate 2×10^{-5} and the LoRA-GA/LoRA-Pro-style recipe (Wang et al., 2024; 2025).

The rank study uses $r \in \{8, 32, 128\}$, following the low-, medium-, and higher-rank configurations examined in LoRA-Pro and GoRA (Wang et al., 2025; He et al., 2025). Rank 8 is the default comparison setting, while the additional NA-LoRA rows assess its behavior as adapter capacity grows. NA-LoRA retains uniform adapter rank; GoRA uses a reference-rank parameter to set its adaptive allocation budget. RaLoRA increases effective rank through gradient-informed block-diagonal decomposition, and RaLoRA-Pro additionally reallocates parameters across layers (Ye et al., 2026). A shared nominal/reference rank identifies the comparison configuration; effective ranks and per-layer allocations remain method-specific. Controller hyperparameters are listed in § H.3.

I.3 NLU and image-classification protocols

Five-task GLUE evaluation. Following the T5 protocol in LoRA-GA and LoRA-Pro (Wang et al., 2024; 2025), we evaluate `t5-base v1.0` on MNLI, SST-2, CoLA, QNLI, and MRPC from GLUE (Wang et al., 2019). Each task is evaluated by *development-set classification accuracy*, including CoLA and MRPC. The Average column is the unweighted arithmetic mean of these five accuracies. It is a five-task accuracy summary, distinct from the official full-GLUE aggregate with task-specific metrics. This convention preserves comparability with inherited baselines in Table 2.

Training settings. For T5 and CLIP we adopt the task-specific settings reported by LoRA-Pro (Wang et al., 2025): default $r = 8$, LoRA $\alpha = 16$, AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and zero weight decay, and a cosine schedule with 3% warmup. T5 uses learning rate 10^{-4}, maximum sequence length 128, and batch size 32. CLIP-ViT-B/16 uses learning rate 10^{-4} and batch size 64; only the visual backbone is adapted and the prompt-derived classifier (e.g., "a photo of a {class}") is frozen. These comparisons use the single-epoch benchmark configuration described

by GoRA (He et al., 2025). CLIP evaluation reports classification accuracy on each of the seven datasets and their arithmetic mean.

I.4 Result provenance and reporting

Our comparisons combine this work's measurements with published reference results. Shared baseline values appearing in LoRA-GA, LoRA-Pro, and GoRA are treated as the same reference measurement, rather than independent runs. The Llama-3.1-8B Full FT MT-Bench entry, 6.02 ± 0.12, is measured in this work using GPT-4o first-turn scoring. Other reporting sources are specified below.

T5-Base / GLUE. The Full FT, LoRA, PiSSA, rsLoRA, LoRA+, DoRA, AdaLoRA, and LoRA-GA reference rows originate in the LoRA-GA comparison (Wang et al., 2024) and are also reported in LoRA-Pro (Wang et al., 2025). The LoRA-Pro row is from LoRA-Pro, and the GoRA row is from GoRA (He et al., 2025). RaLoRA and RaLoRA-Pro are quoted from Table 1 of Ye et al. (2026), using their default rank/reference-rank 8 configuration and the same five-task accuracy protocol with all non-head linear targets. Their original means and standard deviations are retained; these are not reruns. NA-LoRA is evaluated in this work.

Llama-2-7B. The shared LoRA, PiSSA, rsLoRA, LoRA+, DoRA, AdaLoRA, and LoRA-GA reference results trace to LoRA-GA (Wang et al., 2024) and the later comparison tables in LoRA-Pro and GoRA (Wang et al., 2025; He et al., 2025). The Full FT and LoRA-Pro rows follow LoRA-Pro's reported values; the GoRA row follows GoRA. For the PiSSA HumanEval entry, we retain GoRA's reported 16.02 ± 0.17, making the reporting source explicit. We evaluate the NA-LoRA rows.

RaLoRA-family NLG results. The RaLoRA and RaLoRA-Pro rows in Table 1 are evaluated in this work on both Llama-3.1-8B and Llama-2-7B under the shared all-linear NLG protocol. These runs use rank/reference rank 8, target both attention and FFN projections (§ I.2), and follow the backbone-specific MT-Bench judge conventions in § I.1. Their means and standard deviations are reported under this work's seed convention below.

CLIP-ViT-B/16. The Zero-shot, Full FT, LoRA, rsLoRA, LoRA+, DoRA, LoRA-GA, and LoRA-Pro rows are taken from the image-classification comparison in LoRA-Pro (Wang et al., 2025). GoRA's own row is taken from GoRA (He et al., 2025), whose CLIP table also reuses the LoRA-Pro baselines. NA-LoRA is evaluated in this work.

Runs and uncertainty. For runs conducted in this work, we report mean$\pm$std over three seeds $\{0, 1, 2\}$ unless explicitly stated otherwise; the $4\times$A100 training setup applies to these runs. Quoted baseline rows retain the means, standard deviations, and experimental settings of their reporting sources, including the original seed and hardware conventions. The inherited Llama-2-7B, T5, and CLIP comparisons follow the provenance above; Llama-3.1-8B uses the task and metric description in § I.1. The gate-trajectory diagnostics use the seed-0 training trajectory with a separate fixed GSM8K probe sampled using seed 42 (§ K).

Gate-dynamics diagnostic definitions. For the fixed SwiGLU probe in Figures 3 and 4, let i index retained scalar gate scores (sample, token, channel, and, for global pooling, layer), and define

$$R_t(i) = \begin{cases} -1, & z_t(i) < -\tau_z, \\ 0, & |z_t(i)| \le \tau_z, \\ +1, & z_t(i) > \tau_z. \end{cases} \tag{91}$$

For each adjacent checkpoint pair $(t, t+1)$, the reported per-pair diagnostics are

$$S_t := \mathbb{E}_i\big[|\sigma(z_{t+1}(i)) - \sigma(z_t(i))|\big], \quad \text{auxiliary sigmoid-factor change,} \tag{92}$$

$$F_t := \mathbb{E}_i\big[\mathbb{I}(R_{t+1}(i) \neq R_t(i))\big], \quad \text{regime flip rate,} \tag{93}$$

$$D^{(1)}_{\mathrm{EH},t} := \mathbb{E}_i\big[|\kappa_g(z_{t+1}(i)) - \kappa_g(z_t(i))|\big], \quad \text{effective-homogeneity drift,} \tag{94}$$

$$C^{\mathrm{sup}}_t := \mathbb{E}_i\big[\mathbb{I}((R_t(i) = -1) \oplus (R_{t+1}(i) = -1))\big], \quad \text{suppressive-boundary crossing,} \tag{95}$$

$$J^{\mathrm{sev}}_t := \mathbb{E}_i\big[\mathbb{I}((R_t(i), R_{t+1}(i)) \in \{(-1,+1),(+1,-1)\})\big], \quad \text{severe direct jump.} \tag{96}$$

Here $\oplus$ denotes exclusive OR. Cumulative summaries sum the adjacent-checkpoint values. The S_t statistic is an internal-factor diagnostic only: for SwiGLU the operative gate is the whole $\mathrm{SiLU}(z)$, not its sigmoid factor. The direct $D^{(1)}_{\mathrm{EH},t}$ series was regenerated from stored checkpoint probe scores with the corrected analytic $\kappa_g(z) = 1+z(1-\sigma(z))$, with no denominator regularization or clipping. Layer-wise diagnostics restrict i to one FFN layer using the same definitions.

J Additional Experimental Results

Auxiliary internal-factor movement diagnostic. Figure 4 reports two complementary levels of nonlinear-state movement on the Llama-3.1-8B / MetaMathQA probe. Cumulative adjacent-checkpoint movement of the internal self-gating weight $\rho(z) = \sigma(z)$ is larger for NA-LoRA than LoRA (0.3744 vs. 0.3199). The architecture-aligned direct quantity $D^{(1)}_{\mathrm{EH},t}$ instead measures changes in the homogeneity of the full operative gate $g(z) = \mathrm{SiLU}(z)$. Keeping these levels separate connects the auxiliary internal-weight measurement to the gate-level analysis.

K Additional Scope of Gate-Trajectory Diagnostics

Figures 3 and 4 use a fixed GSM8K-train probe subset of 128 examples (seed 42) during MetaMathQA fine-tuning of Llama-3.1-8B at rank $r = 8$, with `mlp.gate_proj` pre-activations probed from step 250 to 50,000 in 250-step increments. For the SwiGLU operative gate,

$$\kappa_g(z) = 1 + z(1 - \sigma(z)), \tag{97}$$

with the removable value at $z = 0$ defined by continuity, and gate states are partitioned by $\pm\tau_z$ with $\tau_z = 1.27846$.

The direct diagnostic is computed from the stored checkpoint probe scores: each transition reports $D^{(1)}_{\mathrm{EH},t} = \mathbb{E}|\Delta\kappa_g|$ with no epsilon denominator and no κ_g clipping. Summed across the 199 adjacent checkpoint transitions, cumulative effective-homogeneity drift is 0.7394 for LoRA and 0.9072 for NA-LoRA. The auxiliary sigmoid-factor cumulative movement is 0.3199 versus 0.3744. Averaged layer-wise across transitions, NA-LoRA has larger $D^{(1)}_{\mathrm{EH}}$ in all 32 FFN layers, complementing the stronger middle/deep-layer regime-crossing motion visible in the diagnostic records.

The boundary statistics remain complementary: cumulative regime-flip rates are 0.1421 and 0.1609, cumulative suppressive-boundary crossings are 0.1279 and 0.1449, and severe direct jumps across distant regimes are nearly zero (1.10×10^{-7} and 3.36×10^{-7} for LoRA and NA-LoRA, respectively). At the saved checkpoint endpoints, these statistics characterize more frequent neighboring-region transitions and greater homogeneity movement under NA-LoRA.

Input dependence and interpretation. The probe fixes external examples, while internal layer inputs evolve across checkpoints. Recorded score drift therefore includes changes propagated from upstream layers in addition to the local adapter contribution (§ F.5). These single-trajectory diagnostics track changes in the proposed state variables; task comparisons and component ablations provide separate evidence of practical usefulness.

Regime crossings and absolute homogeneity drift count movement regardless of its direction. They do not establish monotone homogeneity growth, target-directed mismatch reduction, or causal mediation of task gains. The experiments do not separately isolate the quantile rule, EMA, positive-region penalty, or individual projection-scale ranges.

L Analytical Scope and Extensions

The framework combines gate/value mismatch identities, channel-to-output response relations, conditional weighted low-rank approximation, gate geometry, and optimizer-consistent motion. Homogeneity is a local, channel-wise descriptor aggregated through the value and output paths (§ A.3). Smooth results retain their assumptions; ReLU switching is treated separately.

The quantile mask, EMA, occupancy difference, and clipped projection maps implement response-based design principles; the full functional metric provides the analytical characterization. § B.1

details this scope, and § F.5 explains optimizer dependence. Task gradients retain value, downstream, and loss information.

Task results, component ablations, and fixed-probe diagnostics assess practical utility, controller complementarity, and state reconfiguration, respectively. § K details the diagnostic scope and causal limits. Incorporating more downstream geometry and directly measuring target alignment are natural extensions.