LoRA-GACT: Gauge-Aware Continuation Transport for Optimizer-Preserving Resume

Anonymous authors
Paper under double-blind review

Abstract

LoRA checkpoint exchange has become routine infrastructure: adapters are merged, exported, compressed, and reloaded inside a single training process, and each of these handoffs is validated by one standard—function equivalence at the exchange point. This standard is silent about everything after the exchange. Real PEFT-SVD merge–export–reload–resume runs expose the gap: function-matched checkpoints follow different function-space trajectories, because the handoff changes factor coordinates and optimizer history while the training process continues. We define the missing second standard, optimizer-preserving resume correctness: agreement between the resumed continuation and the uninterrupted reference under a specified continuation system. Gauge-Aware Continuation Transport (LoRA-GACT) realizes this standard, lifting checkpoint conversion into continuation transport by jointly mapping factor-wise update rules, optimizer state, decay, and numerical metadata while retaining the original optimizer. Its theory derives gradient covariance, the global-learning-rate obstruction, and an update-commutation condition that yields multi-step restoration. Across tasks, optimizers, models, ranking evaluation, and real checkpoint exchange, handoffs that pass today's standard still produce trajectory and decision drift, whereas LoRA-GACT restores the specified continuation to numerical precision with zero candidate search.

1 Introduction

Parameter-efficient fine-tuning has turned adaptation into an exchangeable artifact. Low-Rank Adaptation (LoRA) is its most widely used ealization (Hu et al., 2022): the two-factor update supports adapter storage, task switching, merging, model serving, and continual training. In serving systems, one base model now hosts thousands of simultaneously loaded adapters (Sheng et al., 2023; Zhao et al., 2024); in training pipelines, adapters are merged and exported for serving, reconstructed by SVD before resumed use (Gabrielsson et al., 2025), merged into shared cores (Cao et al., 2026), re-expressed through polar decompositions (Lion et al., 2025), and exchanged across federated or continual rounds (Zhang et al., 2026; Jeon et al., 2026). A checkpoint is therefore no longer only a recovery point: it is a routine boundary between two segments of one training process, crossed in both directions.

Yet the standard for crossing this boundary has remained static: weight differences and fixed-input evaluations at the handoff point, plus verbatim store-and-reload of optimizer state. Both practices answer one question—does the exchange preserve the current function?—and are silent about the second: does it preserve the training process that continues after it? Inside one training process this is a no-op contract—exchanging a checkpoint must equal not exchanging it—and we call the property optimizer-preserving resume correctness (Section 3).

The silence is not hypothetical. PEFT's official SVD adapter reconstruction preserves the handoff function to approximately 10^{-6}, passes every standard check, and is accepted practice; yet naive resume progressively departs from the uninterrupted trajectory under both SGD and Momentum SGD. The cause is structural: the reconstructed factors define new coordinates, while the next update is still computed by the old factor-wise rule and optimizer history. The accepted standard therefore

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.

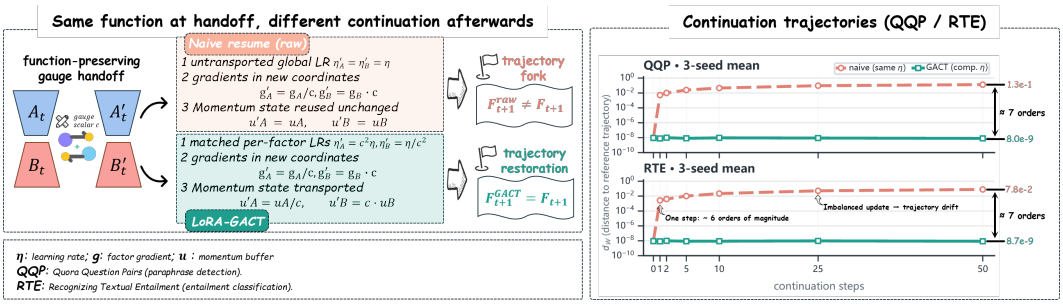

Figure 1: Problem overview. Function-preserving handoffs can still branch after resume when learning rates and optimizer state are reused unchanged; LoRA-GACT transports both to restore the reference continuation. Right: three-seed mean trajectories on QQP and RTE.

certifies handoffs that move SST-2 accuracy by 7.72 points at $c = 8$ within 200 steps; we call the resulting trajectory branch an optimizer-continuation failure.

Existing work studies factor optimization, canonical representations, federated alignment, and invariant perturbations (Han et al., 2026; Zhang et al., 2026; Chen et al., 2026; Wang & Zhang, 2026). Invariant optimizers such as LoRA-RITE install a dedicated recurrence from the start and preserve the continuation generated by that recurrence. Checkpoint handoff poses a different objective: an already trained checkpoint must retain its original optimizer continuation after its factor coordinates and state representation change. This distinction separates changing the continuation system from transporting the existing one. Figure 1 introduces the resulting correctness problem.

LoRA-GACT moves LoRA checkpoint handoff from static function equivalence to dynamic resume equivalence. It records the handoff map and transports the coordinate-induced update rule and optimizer state so that update and transport commute. Our contributions are: **First, the missing second standard:** We define resume correctness, the continuation-level extension of handoff function equivalence, as one evaluation object spanning factor coordinates, trajectories, and observable behavior. **Second, structural theory:** We characterize gradient covariance under scalar and positive-diagonal handoffs, derive the global-learning-rate obstruction, and prove multi-step restoration from one-step update commutation for the studied recurrences. **Third, LoRA-GACT continuation transport:** We derive a zero-search protocol that jointly transports factor update rules, optimizer state, decay, and numerical metadata at an existing checkpoint boundary without replacing the original optimizer; plain SGD retains zero persistent optimizer state. **Finally, empirical and practical validation:** We validate the coordinate-to-trajectory-to-behavior chain across optimizers, models, tasks, downstream ranking, independent implementations, disk-based fresh-process exchange, and the PEFT-SVD merge path.

2 Related Work

LoRA optimization and factor geometry. LoRA factor optimization studies learning-rate design, initialization, gradient construction, and low-rank geometry through LoRA+ (Hayou et al., 2024), LoRA-GA (Wang et al., 2024), PiSSA (Meng et al., 2024), LoRA-One (Zhang et al., 2025), factor-asymmetry analyses (Zhu et al., 2024), LoRA-Pro (Wang et al., 2025), and LoRA-TSD (Andriianov et al., 2026). Together, these works connect factor coordinates and update rules to LoRA training.

Invariant representations and adapter systems. Geometric low-rank optimization and canonical representations separate coordinate choices from intrinsic low-rank objects (Vandereycken, 2013; Edelman et al., 1998). Learning on LoRAs (Putterman et al., 2024), W2T (Han et al., 2026), and Spectral Rank Certification (AHNOUCH & lotfi, 2026) study invariant or canonical processing of low-rank weight spaces, and LoRA-RITE (Yen et al., 2025) installs a dedicated optimizer that keeps updates invariant from the start of training. At other system layers, FedRot-LoRA (Zhang et al., 2026), GLoRA (Chen et al., 2026), and FedPA-LoRA (Jeon et al., 2026) align factor gauges across federated aggregation rounds, PRISM (Wang & Zhang, 2026) imposes gauge-invariant differential

privacy in tangent space, and compression/merging methods (Gabrielsson et al., 2025; Cao et al., 2026) reshape factors for storage and serving.

Checkpoint resume and merge pipelines. These pipelines are now standard practice: SVD reconstructions compress factors before resumed use (Gabrielsson et al., 2025), expert adapters are merged into shared cores (Cao et al., 2026), and federated or continual stages exchange adapters across rounds (Zhang et al., 2026; Jeon et al., 2026). In all of them, resume reloads optimizer state unchanged, which preserves the continuation only when parameter coordinates are unchanged. Neither these pipelines nor adjacent continual-learning or reparameterization techniques—checkpoint averaging, model soups, reparameterized fine-tuning—specifies when a re-expressed checkpoint continues the same optimization process.

Positioning of this work. Existing pipelines enforce the first standard—handoff function equivalence—and are silent about the second. LoRA-GACT makes the checkpoint boundary a transport boundary: function, factor coordinates, optimizer recurrence, state, decay, and numerical metadata move as one continuation object, so that the exchanged checkpoint preserves the incoming optimizer's continuation rather than redefining it.

3 Background and Problem Setting

3.1 LoRA Representation and Function-Preserving Handoff

For frozen base weights $W_0 \in \mathbb{R}^{m\times n}$, LoRA writes

$$W = W_0 + \Delta W, \qquad \Delta W = sBA, \tag{1}$$

where $A \in \mathbb{R}^{r\times n}$, $B \in \mathbb{R}^{m\times r}$, and $s = \alpha/r$, with $r \ll \min(m, n)$ in the low-rank regime. We use BA as the functional update object and distinguish raw factor coordinates (A, B) from functional objects such as ΔW and fixed-input outputs, with the continuation sequence forming the behavior trajectory; coordinate evidence and initialization analysis are collected in Appendix D.

We study function-preserving handoffs induced by positive scalar or positive-diagonal transformations. A positive scalar transformation is

$$A' = cA, \qquad B' = B/c, \qquad c > 0, \tag{2}$$

whereas a positive-diagonal transformation is $A' = CA$, $B' = BC^{-1}$, with $C = \mathrm{diag}(c_1, \ldots, c_r)$ and $c_i > 0$. Both satisfy $B'A' = BA$, preserving ΔW and the linear-layer response while changing factor-coordinate scales; we call them function-preserving gauge handoffs.

3.2 Continuation Restoration

Suppose that the reference reaches (A_{t_0}, B_{t_0}) with optimizer state S_{t_0} at step t_0, and the converted checkpoint is (A'_{t_0}, B'_{t_0}). Under the same subsequent batches, randomness, optimizer steps, and training budget, let the update sequences be $\{\Delta W_t^{\mathrm{ref}}\}$ and $\{\Delta W_t^{\mathrm{cont}}\}$. We require

$$\Delta W_t^{\mathrm{cont}} \approx \Delta W_t^{\mathrm{ref}}, \qquad f_{\mathrm{cont}}(x) \approx f_{\mathrm{ref}}(x), \tag{3}$$

to hold simultaneously on fixed evaluation inputs and held-out downstream inputs.

4 Theoretical Guarantees and LoRA-GACT

4.1 Continuation System and Transport Objective

This section connects gradient covariance to update commutation and multi-step restoration. A handoff changes factor gradients, and reciprocal factor changes require distinct update coefficients. LoRA-GACT transports update coefficients, optimizer state, decay, and numerical metadata within a registered continuation system; continuous-flow results locate the dynamic breakpoint, while registered recurrences provide the discrete transport rules.

Let $F = BA$ and $\Delta W = sF$. For the batch and random state ξ_t at training step t, define

$$M_t := \frac{\partial \ell_t}{\partial F} = s\frac{\partial L_{\xi_t}}{\partial \Delta W}. \tag{4}$$

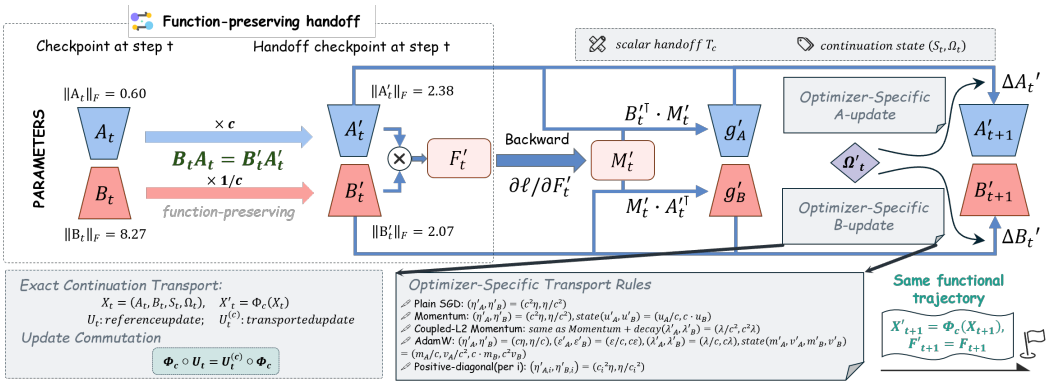

Figure 2: LoRA-GACT continuation transport. Factor coordinates, optimizer state, and update metadata are transported jointly so that the resumed update commutes with the handoff map and preserves the reference functional trajectory.

The reference and converted paths share the frozen base model, subsequent batches, random states, step counter, and scheduler metadata.

Definition 4.1 (Registered continuation system)**.** A registered continuation system contains the loss sequence $\ell_t(F)$, frozen base parameters, LoRA scaling, factor layout, handoff transformation, optimizer recurrence, and subsequent batch and random-state sequence $\{\xi_t\}$. The recurrence fixes parameter groups, state, learning rates, epsilon, decay, scheduler, and step counter, including whether decay is coupled or decoupled. The loss depends on the factors through $F = BA$, and its parameter-group interface exposes the factor-wise or latent-direction-wise metadata required by transport.

Figure 2 summarizes the transport; complete recurrence-level proofs are provided in Appendix A.

4.2 Dynamic Breakpoint: Function Equivalence and Update Dynamics

Proposition 4.2 (Separation of Loss Symmetry and Factor Dynamics)**.** *For continuous gradient flow with the factor-regularization coefficient set to zero,*

$$\dot{A} = -B^{\top}M, \qquad \dot{B} = -MA^{\top}. \tag{5}$$

The latent-space imbalance $D := AA^{\top} - B^{\top}B \in \mathbb{R}^{r\times r}$ is conserved, while the product velocity $\dot{F} = -MA^{\top}A - BB^{\top}M$ depends on the factor coordinates. Consequently, distinct factor coordinates representing the same F generally induce distinct function-space velocities.

Under pure GD, D also develops a second-order discrete drift; its full expression is given in Appendix A. Stateful optimizers and decay follow their registered recurrences below.

Lemma 4.3 (Scalar Covariance of Factor Gradients)**.** *Let $T_c(A, B) = (cA, B/c)$ for $c > 0$. For a loss that depends on the factors through $F = BA$,*

$$g_A = B^{\top}M, \qquad g_B = MA^{\top}, \tag{6}$$

and

$$g'_A = g_A/c, \qquad g'_B = cg_B. \tag{7}$$

Proposition 4.4 (Structural Mismatch under a Shared Global Learning Rate)**.** *Suppose that the reference uses a shared learning rate $\eta > 0$, the handoff path uses a shared learning rate $\bar{\eta} > 0$, and the scalar handoff has $c > 0$. Relative to the target coordinates $T_c(A^+, B^+)$, the factor defects are*

$$\delta_A = \left(c\eta - \frac{\bar{\eta}}{c}\right) g_A, \tag{8}$$

$$\delta_B = \left(\frac{\eta}{c} - \bar{\eta}c\right) g_B. \tag{9}$$

Here δ_A and δ_B use raw-minus-target coordinates, consistent with the function-space residual $F'_{\mathrm{raw}} - F^+$ below. For $c \neq 1$ with nonzero g_A and g_B, the A and B requirements select distinct zero-defect learning rates, producing a joint mismatch. If Bg_A, g_BA, and g_Bg_A are linearly independent, the function-space residual is nonzero for every shared learning rate. The complete product expansion is given in Appendix A.

Corollary 4.5 (Minimum One-Step Exchange Residual of a Global Learning Rate). *Let $a = \|g_A\|_F^2 > 0$ and $b = \|g_B\|_F^2 > 0$, with $c > 0$ and $\eta > 0$. Using $J(\bar{\eta}) = \|\delta_A\|_F^2 + \|\delta_B\|_F^2$ to measure the factor-update exchange residual, the optimal shared learning rate and its minimum residual are*

$$\bar{\eta}^* = \eta \frac{a+b}{a/c^2 + bc^2}, \qquad \min_{\bar{\eta}} J(\bar{\eta}) = \eta^2 \frac{ab(c^2 - c^{-2})^2}{a/c^2 + bc^2}. \tag{10}$$

For $c \neq 1$, this minimum is strictly positive. The residual in this corollary is the factor-update residual J, distinct from the function-space product residual in Proposition 4.4. Thus global-learning-rate (global-LR) search supplies one shared scalar, whereas reciprocal factor updates require two update coefficients.

4.3 Update Commutation and Plain-SGD Restoration

Definition 4.6 (Exact continuation transport). Let $X_t = (A_t, B_t, S_t, \Omega_t)$ collect the factors, optimizer state, learning rates, decay, epsilon, scheduler, and step counter. A joint transport Φ_c defines an exact continuation transport when it satisfies

$$\Phi_c \circ U_t = U_t^{(c)} \circ \Phi_c, \tag{11}$$

for the registered optimizer recurrence.

Lemma 4.7 (Stepwise Commutation Implies Multi-Step Trajectory Restoration). *Suppose that the factor component of Φ_c preserves the product, so $B'A' = BA$ whenever $X' = \Phi_c(X)$. Given the handoff state $X'_{t_0} = \Phi_c(X_{t_0})$, stepwise satisfaction of equation 11 yields*

$$X'_t = \Phi_c(X_t), \qquad B'_t A'_t = B_t A_t, \qquad t_0 \leq t \leq T. \tag{12}$$

The two paths therefore share the same $\Delta W_t = sB_tA_t$ at every step and produce identical model responses in exact arithmetic.

Theorem 4.8 (Exact and Unique Scalar Plain-SGD Transport). *The reference plain-SGD updates are*

$$A_{t+1} = A_t - \eta_t g_{A,t}, \qquad B_{t+1} = B_t - \eta_t g_{B,t}.$$

For the scalar handoff T_c, LoRA-GACT uses

$$\eta'_{A,t} = c^2 \eta_t, \qquad \eta'_{B,t} = \eta_t / c^2 \tag{13}$$

This choice satisfies equation 11, and nonzero gradients on both factors make this coefficient pair unique.

4.4 Optimizer-Specific Transport Rules

For stateful optimizers, transport covers the full continuation state together with the factor learning rates. Table 1 summarizes the registered rules for plain SGD, Momentum SGD, coupled-L_2, decoupled decay, and AdamW (Kingma & Ba, 2015; Loshchilov & Hutter, 2019); the complete coupled-L_2 and decay statements and proofs are in Appendix A.

Theorem 4.9 (Scalar Momentum SGD State and Trajectory Restoration). *For the recurrence $u_p^+ = \mu u_p + g_p$, $p^+ = p - \eta_t u_p^+$, choose*

$$u'_A = u_A / c, \qquad u'_B = c u_B, \qquad \eta'_{A,t} = c^2 \eta_t, \quad \eta'_{B,t} = \eta_t / c^2 \tag{14}$$

These assignments make the factor and momentum-state updates commute, and Lemma 4.7 restores the trajectory.

Theorem 4.10 (Exact Coordinate-wise AdamW Transport). *For*

$$\begin{aligned} m_p^+ &= \beta_1 m_p + (1-\beta_1) g_p, \\ v_p^+ &= \beta_2 v_p + (1-\beta_2) g_p \odot g_p, \\ p^+ &= (1 - \eta_t \lambda_t) p - \eta_t \frac{\widehat{m}_p^+}{\sqrt{\widehat{v}_p^+} + \epsilon}, \end{aligned}$$

Synchronize the step and bias correction, and set

$$m'_A = m_A / c, \quad v'_A = v_A / c^2, \qquad m'_B = c m_B, \quad v'_B = c^2 v_B, \tag{15}$$

$$\eta'_A = c\eta, \quad \epsilon'_A = \epsilon / c, \quad \lambda'_A = \lambda / c,$$
$$\eta'_B = \eta / c, \quad \epsilon'_B = c\epsilon, \quad \lambda'_B = c\lambda. \tag{16}$$

These assignments make the AdamW parameter and state updates commute step by step.

Table 1: Registered LoRA-GACT transport rules for the optimizer recurrences. The table uses the scalar handoff $A' = cA$, $B' = B/c$; positive-diagonal transport replaces c with c_i for each latent direction.

Recurrence	Factor and state transport	Synchronized update metadata
Plain SGD	$\eta'_A = c^2\eta, \eta'_B = \eta/c^2$	Decoupled decay preserves $\eta'_A\lambda'_A = \eta\lambda$, $\eta'_B\lambda'_B = \eta\lambda$
Momentum SGD	$u'_A = u_A/c, u'_B = cu_B$; learning rates follow plain-SGD transport	Momentum buffer and factor learning rates
Coupled-L_2 Momentum	State and learning rates as above; $\lambda'_A = \lambda/c^2, \lambda'_B = c^2\lambda$	Coefficients of coupled-L_2 regularization in the momentum buffer
Coordinate-wise AdamW	$m'_A = m_A/c, v'_A = v_A/c^2$; $m'_B = cm_B, v'_B = c^2 v_B$	$\eta'_A = c\eta, \eta'_B = \eta/c$; $\epsilon'_A = \epsilon/c, \epsilon'_B = c\epsilon$; $\lambda'_A = \lambda/c, \lambda'_B = c\lambda$

4.5 Positive-Diagonal Closure and Checkpoint Protocol

For the positive-diagonal handoff defined in Section 3, let $C = \mathrm{diag}(c_1, \ldots, c_r)$ with $c_i > 0$. The corresponding gradient transformation is

$$g'_A = C^{-1} g_A, \qquad g'_B = g_B C.$$

Each latent direction therefore follows the scalar transport rule:

$$\eta'_{A,i} = c_i^2\eta, \quad \eta'_{B,i} = \eta/c_i^2, \quad u'_{A,i} = u_{A,i}/c_i, \quad u'_{B,:,i} = c_i u_{B,:,i}. \tag{17}$$

Corollary 4.11 (Exact Positive-Diagonal Continuation Transport)**.** *Applying the corresponding scalar rule to the factors, state, epsilon, and decay of each direction assembles the exchange relation into multi-step restoration for a positive-diagonal handoff.*

Theorem 4.12 (LoRA-GACT Checkpoint Continuation Restoration)**.** *Consider the system in Definition 4.1, with a positive scalar handoff T_c or a positive-diagonal handoff T_C that supports per-direction metadata, and choose any registered recurrence in Table 1. When the handoff transports the factors, optimizer state, and update metadata according to the corresponding rule, every continuation length $T - t_0$ satisfies*

$$X'_t = \Phi(X_t), \qquad sB'_t A'_t = sB_t A_t, \qquad t_0 \le t \le T. \tag{18}$$

The frozen-base-model weights, model responses, and downstream decisions determined by those responses are consequently restored step by step.

Appendix A.5 derives the scaling-aware transport law for a general invertible G with matrix-valued update coefficients; Section 5.2 evaluates its approximate finite-precision realization in PEFT-SVD reconstruction.

Corollary 4.13 (Closure for Multi-Module LoRA)**.** *When the optimizer recurrence of multiple LoRA modules separates by factor tensor or latent direction, applying Theorem 4.12 to each module restores the continuation trajectory of the joint model.*

Checkpoint protocol. LoRA-GACT follows four steps: save factors, optimizer state, and step metadata; apply the known gauge transform to factors and registered update metadata; verify the handoff; then resume on the same batches and evaluate the restoration gates.

5 Experimental Evaluation

This section follows the evidence chain from real handoff-induced drift through the failure of naive repair paths to continuation restoration, behavioral propagation, and cross-system validation. Figure 3 summarizes the main results; the stepwise PEFT-SVD trajectory, extended rules, and execution records are provided in Appendix F.4.

5.1 Protocol and Fair Comparison

The comparison includes raw continuation, global-LR, LoRA+ (Hayou et al., 2024), LoRA-RITE (Yen et al., 2025), and LoRA-GACT, all paired with the same uninterrupted reference and sharing the model, task, seed, warm checkpoint, handoff point, subsequent batches, and held-out inputs; differences therefore arise only from the post-handoff coordinates, update rules, and state

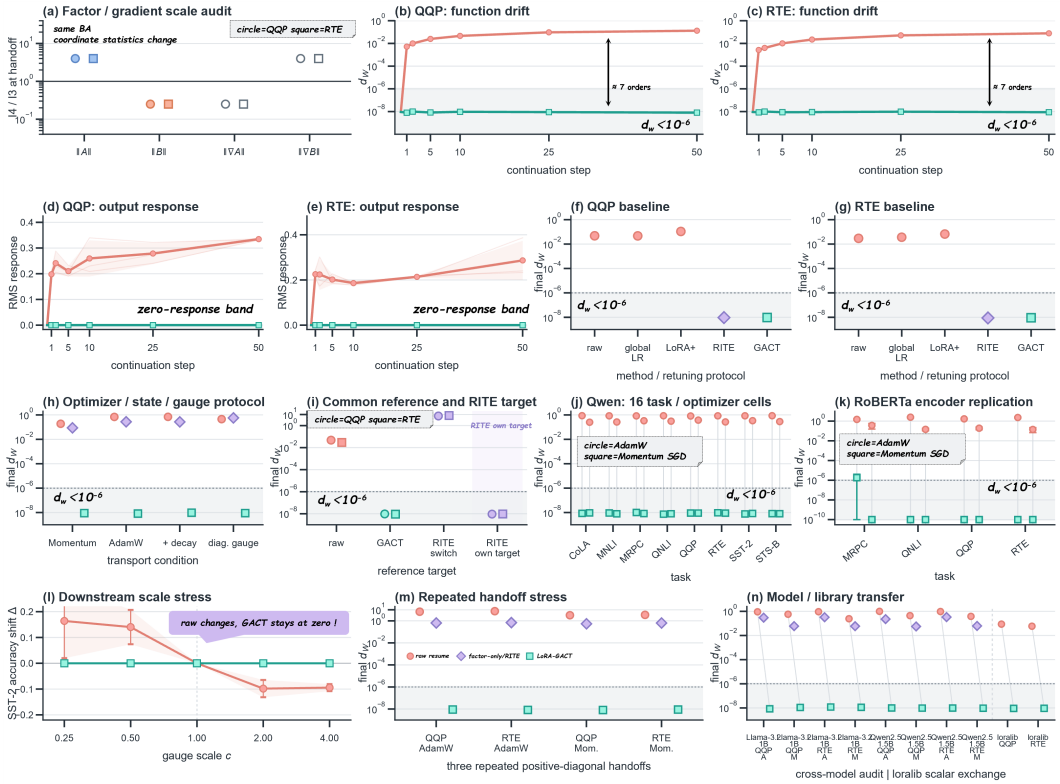

Figure 3: Experimental overview. Panels (a)–(e) trace reciprocal coordinate changes to function-space trajectory drift and output response; (f)–(i) compare global-LR, LoRA+, LoRA-RITE, and transport rules; (j)–(n) extend the evidence across tasks, models, downstream scale, repeated handoffs, and independent implementations. Markers, colors, restoration bands, and error bars encode methods, optimizers, tasks, and seed variation.

transport. The main matrix covers a Qwen decoder, RoBERTa-base, Qwen2.5-1.5B, and Llama-3.2-1B across NLP and recommendation ranking tasks, and validates plain SGD, Momentum-SGD, and AdamW continuation systems; scalar, positive-diagonal, and initialization protocols are given in Appendix F.2. LoRA-RITE is paired with its own uninterrupted reference produced by running RITE from training start; common-reference arms target the original plain-SGD continuation. I3 denotes the nonzero native-scale reference, and I4 the reciprocal-rescaled arm matched to I3 at handoff. Candidate grids, state cost, module layout, and execution steps are given in Appendix F.2.

Metrics. Restoration is first evaluated with relative functional-update distance and output response:

$$d_W = \frac{\left\|\Delta W^{\mathrm{cont}} - \Delta W^{\mathrm{ref}}\right\|_F}{\max(\left\|\Delta W^{\mathrm{ref}}\right\|_F, 10^{-12})}, \qquad d_{\mathrm{out}} = \mathrm{RMS}_{x\in Q}\left(f_{\mathrm{cont}}(x) - f_{\mathrm{ref}}(x)\right). \tag{19}$$

We also report the same RMS response on held-out inputs as held-out output response. The main restoration gate requires $d_W \le 10^{-3}$, fixed-input output response $\le 10^{-3}$, and decision disagreement $\le 1\%$; downstream evaluation additionally reports held-out output response and downstream metrics, and checks paired metrics and decision hashes. Complete definitions, gates, resampling rules, and machine-readable fields are given in Appendix F.

5.2 Function-Preserving Handoff Changes Ordinary Continuation

We first evaluate PEFT 0.20.0's official weighted-SVD adapter reconstruction path through a real merge–export–reload–resume cycle. The handoff has relative function error of approximately 6×10^{-7}; relative ΔW and fixed-input output errors are at the 10^{-7} level, and decision disagreement

Table 2: Scalar-continuation baselines under own and common references. The own-reference block uses each method's uninterrupted run; LoRA-RITE's target is its RITE-from-start run, while common-reference rows target plain SGD. QQP/RTE values are 3-seed means; *Restored* reports passing seeds in QQP/RTE order, and global-LR selects the best of six candidates per seed. Full output response, decision disagreement, and per-task results are in Appendix B.1.

Condition	Target	d_W		Restored		Trials/seed	State MiB
		QQP	RTE	QQP	RTE		
Own-reference baseline (each method's optimizer)							
raw	own	0.0477	0.0294	0/3	0/3	0	0
global-LR selected	own	0.0468	0.0364	0/3	0/3	6	0
LoRA+	own	0.1077	0.0661	0/3	0/3	0	0
LoRA-RITE (own optimizer)	own	9.46×10^{-9}	8.85×10^{-9}	3/3	3/3	0	12.28
LoRA-GACT	own	9.59×10^{-9}	9.13×10^{-9}	3/3	3/3	0	0
Common-reference comparison							
raw gauge	plain SGD	0.04765	0.02939	0/3	0/3	–	0
LoRA-GACT	plain SGD	9.59×10^{-9}	9.13×10^{-9}	3/3	3/3	–	0
RITE-after-SGD	plain SGD	7.4137	8.1985	0/3	0/3	–	12.28
RITE-after-SGD+gauge	plain SGD	7.4137	8.1985	0/3	0/3	–	12.28
RITE-from-start	RITE	9.29×10^{-9}	9.54×10^{-9}	3/3	3/3	–	12.28

remains within the 1% gate, so static validation accepts it. Naive resume nevertheless reaches terminal d_W values of 0.08295, 0.05429, and 0.77680 for QQP/SGD, RTE/SGD, and QQP/Momentum SGD, with nonzero Momentum state producing the largest continuation break. On the same reconstructed checkpoints, LoRA-GACT transports the full-matrix update rule and Momentum state, keeping the three terminal values in 6.81×10^{-7}–9.50×10^{-7} and restoring reference outputs and decisions. Thus, function matching at handoff does not determine the resumed training process; stepwise trajectories, terminal statistics, and the trajectory figure (Figure 12) are given in Appendix F.4.

We evaluate three local-repair attempts, each modifying one component of the continuation system. Selecting the best global learning rate from six candidates per seed leaves terminal d_W at $0.0468/0.0364$ for QQP/RTE. Proposition 4.4 and Corollary 4.5 show that this residual is structural: the zero-defect conditions for A and B require incompatible learning rates, so every shared rate has a positive residual lower bound and cannot yield exact restoration. Zero-state resumption yields terminal $d_W = 0.34283$ for Momentum SGD and 0.78108 for AdamW. Applying the plain-SGD $c^2\eta$ rule to AdamW increases the unrepaired zero-state AdamW baseline from 0.78108 to 2.64114, a $3.4\times$ increase. Full values are reported in Table 2 and Appendix C; LoRA-GACT transports this joint object in the next subsection.

5.3 LoRA-GACT Restores Function-Space Continuation

By Lemma 4.3, $c = 4$ scales ∇_A down by a factor of 4 and ∇_B up by a factor of 4. Figure 10 separately records checkpoint parameter norms and next-batch gradient norms; the $\eta'_A = c^2\eta$, $\eta'_B = \eta/c^2$ rule from Theorem 4.8 compensates the two update coefficients. Panels (b)–(e) of Figure 3 place the rule beside raw resume: LoRA-GACT follows the reference trajectory from steps 50–100, raw resume develops ΔW and output drift, and all six task–seed cells meet the restoration gate.

Panels (d), (e), (l), and (m) of Figure 3 connect trajectory drift to observable behavior: raw and factor-only produce nonzero response after the first update, while LoRA-GACT remains at zero; repeated handoffs amplify raw decision disagreement, while LoRA-GACT remains at zero. In the SST-2 $c = 8$ stress test, I3 and I4 are function-matched at handoff up to the reported numerical construction error; after 200 steps, raw-I4 changes accuracy by -7.72 ± 2.64 percentage points and produces 8.26%–11.81% decision disagreement. Matched-initialization independent endpoints show no significant difference (Appendix D), whereas a mid-training handoff breaks continuation fidelity and decision stability. LoRA-GACT restores d_W to approximately 10^{-8} and returns output, decisions, and downstream metrics to the reference; the complete scale and QQP records are in Appendix F.3.

5.4 Update Rules, Optimizer State, Epsilon, and Decay Determine Restoration

Under the zero-state, $c = 4$ protocol, LoRA-GACT restores the mean d_W to 8.98×10^{-9} for Momentum SGD and 8.25×10^{-9} for AdamW, while the SGD-law control's 2.64114 shows the

transport rule must match the specified recurrence (panels (h) and (i) of Figure 3). With nonzero state, all 12 task–seed–optimizer combinations pass the complete restoration gate; at $\lambda = 0.01$, shared-decay controls retain 0.0164–0.1036 deviation, whereas full transport restores to approximately 10^{-8}. Positive-diagonal handoffs follow the same pattern, showing that the update rule, state, decay, and per-direction metadata form one continuation transport; complete cell values are in Appendix C.

5.5 Generalization Across Tasks, Models, and Downstream

The Qwen matrix spans 16 task–optimizer cells (8 tasks, 2 optimizers, and 3 seeds); raw cross-cell d_W/response averages are approximately $0.57/0.32$, while the mean of the 16 cell-level three-seed means for LoRA-GACT is 8.63×10^{-9} and every cell is restored. Architectural transfer covers eight RoBERTa-base cells and eight Qwen2.5/Llama cross-decoder cells: raw retains measurable deviation, factor-only retains a residual, and full LoRA-GACT restores all cells with zero response and decision disagreement. Recommendation ranking covers six Beauty, Sports, and Toys cells; raw-I4 produces a nonzero ranking-hash change relative to the uninterrupted reference in every evaluated cell, whereas LoRA-GACT produces none. Per-task, per-cell, HR/NDCG/MRR, and protocol details are given in Appendix F.

5.6 Baselines and Checkpoint Exchange

Table 2 separates the two continuations: LoRA-RITE preserves its RITE-from-start continuation, whereas switching a plain-SGD checkpoint into RITE leaves QQP/RTE d_W at $7.4137/8.1985$ relative to the original plain-SGD target. LoRA-GACT transports the incoming continuation system at the checkpoint boundary and restores the original plain-SGD target with zero search and no persistent state. Disk export, CPU reload, and fresh-process resume through PEFT-SVD and Microsoft loralib show the same separation: raw drifts, LoRA-GACT restores. Per-seed values, resource costs, independent implementations, and package entry points are in Appendices F.4 and F.

6 Discussion

Theorem 4.12 casts checkpoint continuation as an exchange diagram: when update and transport commute, this yields multi-step trajectory restoration. The global-LR bound shows why a shared learning rate cannot reproduce reciprocal factor updates. In Section 1's terminology, existing pipelines enforce the first standard, whereas LoRA-GACT supplies the second: it preserves the incoming optimizer continuation, while an invariant optimizer such as LoRA-RITE preserves only the continuation generated by its recurrence from initialization. The `lora-gact` release implements GACT transport and records the LoRA-RITE comparison protocol.

Limitations and Future Work. Transport laws are derived per specified recurrence (Table 1). For a new optimizer, LoRA-GACT derives its own transport rule by enforcing the same update-commutation condition on the optimizer recurrence and transports the corresponding optimizer state, decay, and numerical metadata; the framework therefore extends by rule derivation rather than special-case enumeration. Full-matrix transport covers plain SGD and Momentum; AdamW currently supports only scalar and positive-diagonal gauges. Restoration is exact in exact arithmetic and reaches numerical precision in FP64; FP32 effects on second-moment metadata are quantified in Appendix A.10. Behavioral evidence covers the evaluated classification and ranking tasks.

7 Conclusion

Correct LoRA checkpoint handoff requires function equivalence at exchange and resume equivalence thereafter. LoRA-GACT treats handoff as continuation transport: it jointly maps the incoming factor-coordinate update rule, optimizer state, decay, and numerical metadata so that the resumed update commutes with the handoff. It restores the reference trajectory, outputs, decisions, and downstream behavior with zero search, no persistent plain-SGD state, and a reusable Python package, as validated across PEFT-SVD reconstruction, disk-based fresh-process exchange, cross-model transfer, and independent implementations.

MANDATORY STATEMENTS

AI Use Statement. Generative AI tools assisted this work in five areas: (i) editing the manuscript for readability, including polishing selected paragraphs and condensing verbose passages while preserving their information; (ii) suggesting an organizational structure for the paper; (iii) locating and identifying literature relevant to the study, while the authors assessed differences and novelty relative to prior work through their own reading of the retrieved literature; (iv) drafting and refining plotting code used to visualize experimental results produced by the authors; and (v) drafting, debugging, and refactoring selected code fragments in the experiment runners and reproducibility utilities. The authors wrote and controlled the core implementation and experimental logic. Each AI-assisted fragment was checked against the intended protocol and tested through the executable experiment or unit-test pipeline before use. AI tools did not make the final decisions about the theoretical model, methodology, experimental protocol, reported results, or scientific interpretation. All experimental data and figures derive from the authors' executed runs and recorded outputs. The authors read and verified all cited references, confirmed the behavior of AI-assisted code and outputs, and take responsibility for the final manuscript, including its text, claims, and artifacts produced with AI assistance.

Reproducibility Statement. The reproducibility artifact `lora_gact_artifact.zip` provides the source-only `lora-gact` implementation under `src/lora_gact/`, the public handoff protocol and reproduction instructions, deterministic smoke and unit tests under `scripts/` and `tests/`, machine-readable core results and a SHA-256 manifest under `results/`, and a native cross-library checkpoint runner with its recorded results under `experiments/cross_library/`. Appendices E and F define the release boundary, execution protocol, seeds, restoration gates, data and model fingerprints, result schema, and reproduction entry points; Appendix F.4 documents the PEFT-SVD and independent-implementation exchange results. The artifact reproduces the package-level protocol checks and cross-library smoke experiment, while the appendix reports the full model, task, optimizer, ranking, and checkpoint-exchange matrices generated from the recorded experiments.

REFERENCES

Mohammed AHNOUCH and Elaachak lotfi. Spectral rank certification for foundation model adapters. In *KDD 2026 Workshop on Interplay Between Classical Tensor Methods And Foundation Models*, 2026. URL https://openreview.net/forum?id=T15YnsaP1b.

Dmitrii Andriianov, Andrey Veprikov, and Aleksandr Beznosikov. Lora-tsd: Tangent-space spectral descent for lora via muon-style updates. *arXiv preprint arXiv:2609.02734*, 2026.

Jie Cao, Zhenxuan Fan, Zhuonan Wang, Tianwei Lin, Ziyuan Zhao, Rolan Yan, Wenqiao Zhang, Feifei Shao, Hongwei Wang, Jun Xiao, and Siliang Tang. CoMoL: Efficient mixture of LoRA experts via dynamic core space merging. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 16471–16481, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.811. URL https://aclanthology.org/2026.findings-acl.811/.

Jinqian Chen, Chang Liu, and Jihua Zhu. Beyond factor aggregation: Gauge-aware low-rank server representations for federated lora. *arXiv preprint arXiv:2605.06733*, 2026.

Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

Rickard Brüel Gabrielsson, Jiacheng Zhu, Onkar Bhardwaj, Leshem Choshen, Kristjan Greenewald, Mikhail Yurochkin, and Justin Solomon. Compress then serve: Serving thousands of loRA adapters with little overhead. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=3XMA8RDJu2.

Xiaolong Han, Ferrante Neri, Zijian Jiang, Fang Wu, Yanfang Ye, Lu Yin, and Zehong Wang. W2t: Lora weights already know what they can do. *arXiv preprint arXiv:2603.15990*, 2026.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. LoRA+: Efficient low rank adaptation of large models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=NEv8YqBROO.

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

Juseok Jeon, Ramy E Ali, Doyun Kwon, Myungbeom Her, Jinhwi Kim, and Jinhyun So. Fedpa-lora: Product-aligned framework for mitigating aggregation and initialization errors in heterogeneous federated lora. *arXiv preprint arXiv:2608.15381*, 2026.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, 2015. URL https://mlanthology.org/iclr/2015/kingma2015iclr-adam/.

Kai Lion, Liang Zhang, Bingcong Li, and Niao He. Polar: Polar-decomposed low-rank adapter representation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 149313–149358. Curran Associates, Inc., 2025. doi: 10.52202/085713-4996. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/db79b6c50430a52fbab9d63efb8433ca-Paper-Conference.pdf.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. Pissa: Principal singular values and singular vectors adaptation of large language models. *Advances in Neural Information Processing Systems*, 37:121038–121072, 2024.

Theo Putterman, Derek Lim, Yoav Gelberg, Stefanie Jegelka, and Haggai Maron. Learning on loras: Gl-equivariant processing of low-rank weight spaces for large finetuned models. *arXiv preprint arXiv:2410.04207*, 2024.

Ying Sheng, Shiyi Cao, Dacheng Li, Coleman Hooper, Nicholas Lee, Shuo Yang, Christopher Chou, Banghua Zhu, Lianmin Zheng, Kurt Keutzer, et al. S-lora: Serving thousands of concurrent lora adapters. *arXiv preprint arXiv:2311.03285*, 2023.

Bart Vandereycken. Low-rank matrix completion by riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013.

Shaowen Wang, Linxi Yu, and Jian Li. Lora-ga: Low-rank adaptation with gradient approximation. *Advances in Neural Information Processing Systems*, 37:54905–54931, 2024.

Shihao Wang and Xueru Zhang. PRISM: Gauge-invariant tangent-space differentially private loRA. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=SiCjKmArjQ.

Zhengbo Wang, Jian Liang, Ran He, Zilei Wang, and Tieniu Tan. LoRA-pro: Are low-rank adapters properly optimized? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=gTwRMU3lJ5.

Jui-Nan Yen, Si Si, Zhao Meng, Felix Yu, Venkata Sai Surya Subramanyam Duvvuri, Inderjit Dhillon, Cho-Jui Hsieh, and Sanjiv Kumar. Lora done rite: Robust invariant transformation equilibration for lora optimization. In *International Conference on Learning Representations*, volume 2025, pp. 75815–75834, 2025.

Haoran Zhang, Dongjun Kim, Seohyeon Cha, and Haris Vikalo. Fedrot-loRA: Mitigating rotational misalignment in federated loRA. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=2X8Qi3VdjA.

Yuanhe Zhang, Fanghui Liu, and Yudong Chen. LoRA-one: One-step full gradient could suffice for fine-tuning large language models, provably and efficiently. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=KwIlvmLDLm.

Justin Zhao, Timothy Wang, Wael Abid, Geoffrey Angus, Arnav Garg, Jeffery Kinnison, Alex Sherstinsky, Piero Molino, Travis Addair, and Devvret Rishi. Lora land: 310 fine-tuned llms that rival gpt-4, a technical report. *arXiv preprint arXiv:2405.00732*, 2024.

Jiacheng Zhu, Kristjan Greenewald, Kimia Nadjahi, Haitz Sáez de Ocáriz Borde, Rickard Brüel Gabrielsson, Leshem Choshen, Marzyeh Ghassemi, Mikhail Yurochkin, and Justin Solomon. Asymmetry in low-rank adapters of foundation models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=txRZBD8tBV.

A Continuation Theory: Complete Proofs

This appendix gives complete proofs of the theoretical results in the main text. To keep concrete experimental models and batch notation out of the theory, all equations are written at one fixed continuation step and the step subscript is omitted unless needed. Let $A \in \mathbb{R}^{r\times n}$, $B \in \mathbb{R}^{m\times r}$, and $F = BA$, $M = \partial\ell/\partial F$, with $g_A = B^\mathsf{T}M$ and $g_B = MA^\mathsf{T}$. Matrix inner products use $\langle X, Y\rangle_F = \mathrm{tr}(X^\mathsf{T}Y)$. "Equality" in these proofs means equality in exact arithmetic; finite-precision errors are measured by the continuation gates in the main text.

A.1 Factor Flow, Imbalance, and Discrete Drift

Proposition (Factor Flow, Imbalance, and Discrete Drift (corresponding to Proposition 4.2))**. Assumptions and notation.** Let $A \in \mathbb{R}^{r\times n}$, $B \in \mathbb{R}^{m\times r}$, $F = BA$, and suppose that the loss depends on the factors only through F. Let $M = \partial\ell/\partial F$. With the factor-regularization coefficient set to zero, the continuous gradient flow is

$$\dot{A} = -B^\mathsf{T}M, \qquad \dot{B} = -MA^\mathsf{T}.$$

Define the latent-space imbalance $D = AA^\mathsf{T} - B^\mathsf{T}B \in \mathbb{R}^{r\times r}$. Then D is conserved under this flow, while the product velocity is

$$\frac{\mathrm{d}}{\mathrm{d}t}D = 0, \qquad \dot{F} = -MA^\mathsf{T}A - BB^\mathsf{T}M.$$

Thus, distinct factor coordinates representing the same F generally have distinct function-space velocities. If pure GD uses the same learning rate η, $A^+ = A - \eta B^\mathsf{T}M$ and $B^+ = B - \eta MA^\mathsf{T}$, the imbalance has the discrete change

$$D^+ - D = \eta^2\left(B^\mathsf{T}MM^\mathsf{T}B - AM^\mathsf{T}MA^\mathsf{T}\right).$$

Complete proof of Proposition 4.2. Because the loss depends on the factors only through $F = BA$, the chain rule gives the factor gradients $\nabla_A\ell = B^\mathsf{T}M$ and $\nabla_B\ell = MA^\mathsf{T}$. With no factor regularizer, the continuous gradient flow is therefore

$$\dot{A} = -B^\mathsf{T}M, \qquad \dot{B} = -MA^\mathsf{T}.$$

First compute the derivative of AA^T:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(AA^\mathsf{T}) &= \dot{A}A^\mathsf{T} + A\dot{A}^\mathsf{T}\\ &= (-B^\mathsf{T}M)A^\mathsf{T} + A(-MA^\mathsf{T})^\mathsf{T}\\ &= -B^\mathsf{T}MA^\mathsf{T} - AM^\mathsf{T}B.\end{aligned}$$

Next compute the derivative of $B^\mathsf{T}B$:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(B^\mathsf{T}B) &= \dot{B}^\mathsf{T}B + B^\mathsf{T}\dot{B}\\ &= (-MA^\mathsf{T})^\mathsf{T}B + B^\mathsf{T}(-MA^\mathsf{T})\\ &= -AM^\mathsf{T}B - B^\mathsf{T}MA^\mathsf{T}.\end{aligned}$$

The right-hand sides are identical, so

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(AA^{\mathsf{T}} - B^{\mathsf{T}}B\right) = 0.$$

This proves that $D = AA^{\mathsf{T}} - B^{\mathsf{T}}B$ is conserved under this continuous flow.

Applying the product rule gives

$$\begin{aligned}\dot{F} &= \dot{B}A + B\dot{A} \\ &= (-MA^{\mathsf{T}})A + B(-B^{\mathsf{T}}M) \\ &= -MA^{\mathsf{T}}A - BB^{\mathsf{T}}M.\end{aligned}$$

Now let $A' = cA$ and $B' = B/c$ with $c > 0$. Then $B'A' = BA = F$, so $M' = M$ for the same input and base model. However, the product velocity in the new coordinates is

$$\begin{aligned}\dot{F}' &= -M(A')^{\mathsf{T}}A' - B'(B')^{\mathsf{T}}M \\ &= -c^2MA^{\mathsf{T}}A - c^{-2}BB^{\mathsf{T}}M.\end{aligned}$$

The original velocity is $-MA^{\mathsf{T}}A - BB^{\mathsf{T}}M$. Unless the terms satisfy a special cancellation relation or $c = 1$, the two velocities differ; static equality of F therefore does not imply equality of its continuous function-space velocity.

Finally, prove the discrete identity in the main text. Let $A^+ = A - \eta B^{\mathsf{T}}M$ and $B^+ = B - \eta MA^{\mathsf{T}}$. Direct expansion gives

$$\begin{aligned}A^+(A^+)^{\mathsf{T}} = AA^{\mathsf{T}} &- \eta B^{\mathsf{T}}MA^{\mathsf{T}} - \eta AM^{\mathsf{T}}B \\ &+ \eta^2 B^{\mathsf{T}}MM^{\mathsf{T}}B,\end{aligned}$$

and

$$\begin{aligned}(B^+)^{\mathsf{T}}B^+ = B^{\mathsf{T}}B &- \eta AM^{\mathsf{T}}B - \eta B^{\mathsf{T}}MA^{\mathsf{T}} \\ &+ \eta^2 AM^{\mathsf{T}}MA^{\mathsf{T}}.\end{aligned}$$

Subtracting cancels the two first-order cross terms term by term, leaving

$$D^+ - D = \eta^2\left(B^{\mathsf{T}}MM^{\mathsf{T}}B - AM^{\mathsf{T}}MA^{\mathsf{T}}\right),$$

which is the discrete balance-drift identity above. This second-order result uses the same η in pure GD; recurrences with historical state or nonuniform preconditioning require separate commutation checks. □

A.2 Gradient Covariance and the Global-Learning-Rate Obstruction

Lemma (Scalar Covariance of Factor Gradients (corresponding to Lemma 4.3))**. Assumptions and notation.** Let $T_c(A, B) = (cA, B/c)$ with $c > 0$ and let the loss depend on the factors through $F = BA$. If $M = \partial\ell/\partial F$, then

$$g_A = \nabla_A\ell = B^{\mathsf{T}}M, \qquad g_B = \nabla_B\ell = MA^{\mathsf{T}}.$$

On the same input, T_c preserves F and M and gives

$$g'_A = g_A/c, \qquad g'_B = cg_B.$$

Proposition (Structural Mismatch under a Shared Global Learning Rate (corresponding to Proposition 4.4))**. Assumptions and notation.** The reference uses a shared learning rate $\eta > 0$, the handoff path uses a shared learning rate $\bar{\eta} > 0$, and both paths start from the same state under the scalar handoff $T_c(A, B) = (cA, B/c)$ with $c > 0$. Relative to the target coordinates $T_c(A^+, B^+)$, define the factor defects as the raw update minus the target:

$$\delta_A = \left(c\eta - \frac{\bar{\eta}}{c}\right)g_A, \qquad \delta_B = \left(\frac{\eta}{c} - \bar{\eta}c\right)g_B.$$

The one-step products on the reference and raw paths are

$$F^+ = (B - \eta g_B)(A - \eta g_A), \qquad F'_{\mathrm{raw}} = (B/c - \bar{\eta}cg_B)(cA - \bar{\eta}g_A/c).$$

Their difference is

$$F'_{\text{raw}} - F^+ = \left(\eta - \frac{\bar{\eta}}{c^2}\right) Bg_A + \left(\eta - \bar{\eta}c^2\right) g_B A + (\bar{\eta}^2 - \eta^2) g_B g_A.$$

When $c \neq 1$ and g_A, g_B are both nonzero, the A and B zero-defect requirements select different $\bar{\eta}$. When Bg_A, $g_B A$, and $g_B g_A$ are linearly independent, the function-space residual is nonzero for every shared learning rate.

Corollary (Minimum One-Step Exchange Residual of a Global Learning Rate (corresponding to Corollary 4.5))**. Assumptions and notation.** Let $a = \|g_A\|_F^2 > 0$ and $b = \|g_B\|_F^2 > 0$, with $c > 0$ and $\eta > 0$, and use

$$J(\bar{\eta}) = \|\delta_A\|_F^2 + \|\delta_B\|_F^2$$

to measure the factor-update residual. For $c > 0$, the optimal shared learning rate and minimum residual are

$$\bar{\eta}^\star = \eta \frac{a+b}{a/c^2 + bc^2}, \qquad \min_{\bar{\eta}} J(\bar{\eta}) = \eta^2 \frac{ab(c^2 - c^{-2})^2}{a/c^2 + bc^2}.$$

When $c \neq 1$, the minimum residual is strictly positive.

Complete proof of Lemma 4.3. Differentiate $F = BA$:

$$\mathrm{d}F = (\mathrm{d}B)A + B(\mathrm{d}A).$$

Therefore,

$$\begin{aligned} \mathrm{d}\ell &= \langle M, \mathrm{d}F \rangle_F \\ &= \mathrm{tr}\big(M^\mathsf{T}(\mathrm{d}B)A\big) + \mathrm{tr}\big(M^\mathsf{T}B(\mathrm{d}A)\big). \end{aligned}$$

By cyclic invariance of the trace, the first term can be written as $\mathrm{tr}((MA^\mathsf{T})^\mathsf{T}\mathrm{d}B)$, and the second term can be written as $\mathrm{tr}((B^\mathsf{T}M)^\mathsf{T}\mathrm{d}A)$. By the definition of the Frobenius gradient,

$$g_A = \nabla_A \ell = B^\mathsf{T}M, \qquad g_B = \nabla_B \ell = MA^\mathsf{T}.$$

After handoff, let $A' = cA$ and $B' = B/c$. Since $B'A' = BA$, the loss derivative with respect to F remains $M' = M$. Substituting into the gradient expressions above gives

$$g'_A = (B')^\mathsf{T}M = \frac{1}{c}B^\mathsf{T}M = \frac{g_A}{c},$$

and

$$g'_B = M(A')^\mathsf{T} = cMA^\mathsf{T} = cg_B.$$

This proves reciprocal covariance of the two factor gradients. □

Complete proof of Proposition 4.4. The reference one-step plain-SGD update is

$$A^+ = A - \eta g_A, \qquad B^+ = B - \eta g_B.$$

First apply the handoff to this reference update:

$$T_c(A^+, B^+) = \left(cA - c\eta g_A,\ \frac{B}{c} - \frac{\eta}{c} g_B\right).$$

Conversely, when the handoff is applied to the current checkpoint first, Lemma 4.3 gives $g'_A = g_A/c$ and $g'_B = cg_B$. If the converted path still uses the same global learning rate $\bar{\eta}$, its raw update is

$$(A'_{\text{raw}}, B'_{\text{raw}}) = \left(cA - \frac{\bar{\eta}}{c} g_A,\ \frac{B}{c} - \bar{\eta}cg_B\right).$$

Define the one-step defect as raw coordinates minus target coordinates:

$$\begin{aligned} \delta_A &= \left(cA - \frac{\bar{\eta}}{c} g_A\right) - (cA - c\eta g_A) \\ &= \left(c\eta - \frac{\bar{\eta}}{c}\right) g_A, \end{aligned}$$

and

$$\begin{aligned}\delta_B &= \left(\frac{B}{c} - \bar{\eta} c g_B\right) - \left(\frac{B}{c} - \frac{\eta}{c} g_B\right) \\ &= \left(\frac{\eta}{c} - \bar{\eta} c\right) g_B.\end{aligned}$$

Thus δ_A, δ_B use the same raw-minus-target orientation as the function-space defect $F'_{\text{raw}} - F^+$ in the next proof. If $g_A \neq 0$, $\delta_A = 0$ requires $c\eta - \bar{\eta}/c = 0$, giving $\bar{\eta} = c^2\eta$. If $g_B \neq 0$, $\delta_B = 0$ requires $\eta/c - \bar{\eta}c = 0$, giving $\bar{\eta} = \eta/c^2$. When $c \neq 1$, $c^2\eta \neq \eta/c^2$, so one shared $\bar{\eta}$ cannot eliminate both defects. □

One-step function-space defect proof for Proposition 4.4. First expand the raw-continuation product:

$$\begin{aligned}F'_{\text{raw}} &= \left(\frac{B}{c} - \bar{\eta} c g_B\right)\left(cA - \frac{\bar{\eta}}{c} g_A\right) \\ &= BA - \frac{\bar{\eta}}{c^2} B g_A - \bar{\eta} c^2 g_B A + \bar{\eta}^2 g_B g_A.\end{aligned}$$

The product after the reference update is

$$\begin{aligned}F^+ &= (B - \eta g_B)(A - \eta g_A) \\ &= BA - \eta B g_A - \eta g_B A + \eta^2 g_B g_A.\end{aligned}$$

Subtracting term by term cancels the BA term and gives

$$\begin{aligned}F'_{\text{raw}} - F^+ = &\left(\eta - \frac{\bar{\eta}}{c^2}\right) B g_A + \left(\eta - \bar{\eta} c^2\right) g_B A \\ &+ (\bar{\eta}^2 - \eta^2) g_B g_A,\end{aligned}$$

which is the function-space defect identity above. If the three matrices Bg_A, g_BA, and g_Bg_A are linearly independent and the difference is zero, their coefficients must vanish individually. The first two coefficients give $\bar{\eta} = c^2\eta$ and $\bar{\eta} = \eta/c^2$. When $c \neq 1$ and $\eta > 0$, these conditions conflict, so no shared learning rate makes the two function-space products equal. The third coefficient adds a constraint and cannot remove the conflict between the first two. □

Complete proof of Corollary 4.5. Let $x = \bar{\eta}$, $a = \|g_A\|_F^2$, and $b = \|g_B\|_F^2$. From Eqs. equation 8–equation 9,

$$\begin{aligned}J(x) &= a\left(c\eta - \frac{x}{c}\right)^2 + b\left(\frac{\eta}{c} - cx\right)^2 \\ &= \left(\frac{a}{c^2} + bc^2\right) x^2 - 2\eta(a+b)x + \eta^2\left(ac^2 + \frac{b}{c^2}\right).\end{aligned}$$

Let $q = a/c^2 + bc^2$. For $a, b > 0$ and $c > 0$, $q > 0$, so J is a strictly convex quadratic. Differentiating and setting the derivative to zero gives

$$J'(x) = 2qx - 2\eta(a+b) = 0,$$

the unique minimizer

$$x^\star = \eta \frac{a+b}{a/c^2 + bc^2}.$$

Writing the quadratic at its minimizer as a constant minus one squared term gives

$$\begin{aligned}J(x^\star) &= \eta^2\left(ac^2 + \frac{b}{c^2}\right) - \eta^2 \frac{(a+b)^2}{a/c^2 + bc^2} \\ &= \eta^2 \frac{\left(ac^2 + b/c^2\right)\left(a/c^2 + bc^2\right) - (a+b)^2}{a/c^2 + bc^2}.\end{aligned}$$

The numerator expands as

$$\begin{aligned}&\left(a^2 + abc^4 + abc^{-4} + b^2\right) - (a^2 + 2ab + b^2)\\&= ab(c^4 + c^{-4} - 2) = ab(c^2 - c^{-2})^2.\end{aligned}$$

Therefore,

$$\min_x J(x) = \eta^2 \frac{ab(c^2 - c^{-2})^2}{a/c^2 + bc^2}.$$

When $c \neq 1$ and $a, b > 0$, the numerator and denominator are positive, so the minimum residual is strictly positive. Thus, even continuous search over a shared learning rate cannot yield exact factor-coordinate restoration. □

A.3 Commutation and Plain SGD

Notation and applicability. Let $X_t = (A_t, B_t, S_t, \Omega_t)$ denote the joint state of the factors, optimizer state, learning rates, decay, epsilon, scheduler, and step counter. Let U_t and $U_t^{(c)}$ denote the registered updates in the reference and handoff coordinates, respectively. The joint transport Φ_c acts on these states and update metadata.

Lemma (Stepwise Commutation Implies Multi-Step Trajectory Restoration (corresponding to Lemma 4.7))**. Assumption.** Suppose that the factor component of Φ_c preserves the product, so $B'A' = BA$ whenever $X' = \Phi_c(X)$. Suppose also that the handoff state satisfies $X'_{t_0} = \Phi_c(X_{t_0})$ and that every subsequent step satisfies

$$\Phi_c \circ U_t = U_t^{(c)} \circ \Phi_c.$$

Then, for all $t_0 \leq t \leq T$,

$$X'_t = \Phi_c(X_t), \qquad B'_t A'_t = B_t A_t.$$

Therefore, $\Delta W_t = sB_tA_t$, the model responses, and the decisions determined by those responses agree step by step in exact arithmetic.

Theorem (Exact and Unique Scalar Plain-SGD Transport (corresponding to Theorem 4.8))**. Assumptions and notation.** The reference uses

$$A^+ = A - \eta_t g_A, \qquad B^+ = B - \eta_t g_B,$$

and the handoff is $A' = cA$, $B' = B/c$ with $c > 0$. Let the handoff path use factor-level learning rates

$$\eta'_{A,t} = c^2 \eta_t, \qquad \eta'_{B,t} = \eta_t / c^2$$

Then these coefficients satisfy one-step update commutation. When g_A and g_B are both nonzero, this coefficient pair is unique among scalar learning rates assigned separately to the two factors.

Complete proof of Lemma 4.7. Let the reference state start from X_{t_0} and the converted path start from $X'_{t_0} = \Phi_c(X_{t_0})$. At $t = t_0$, the conclusion $X'_{t_0} = \Phi_c(X_{t_0})$ follows directly from the handoff definition.

Assume inductively for some $t < T$ that $X'_t = \Phi_c(X_t)$. The converted path uses $U_t^{(c)}$ at the next step, so

$$\begin{aligned}X'_{t+1} &= U_t^{(c)}(X'_t)\\&= U_t^{(c)}(\Phi_c(X_t))\\&= \Phi_c(U_t(X_t))\\&= \Phi_c(X_{t+1}),\end{aligned}$$

The third line is exactly the one-step commutation relation equation 11. The inductive hypothesis therefore propagates from t to $t + 1$, and mathematical induction gives $X'_t = \Phi_c(X_t)$ for all $t_0 \leq t \leq T$.

For a scalar handoff, the factor part of Φ_c is $A \mapsto cA$, $B \mapsto B/c$; for a positive-diagonal handoff, it is $A \mapsto CA$, $B \mapsto BC^{-1}$. In both cases,

$$(B/c)(cA) = BA, \qquad (BC^{-1})(CA) = BA.$$

Thus the inductive state relation also gives $B'_t A'_t = B_t A_t$. If the actual weight update is $\Delta W_t = sB_t A_t$, fixed s gives $\Delta W'_t = \Delta W_t$. With the same base parameters, batches, random states, and deterministic forward computation, the model responses are identical because the weights are identical. □

Complete proof of Theorem 4.8. First prove sufficiency. The reference update is

$$A^+ = A - \eta_t g_A, \qquad B^+ = B - \eta_t g_B.$$

After handoff, the current factors are $A' = cA$, $B' = B/c$, with gradients $g'_A = g_A/c$ and $g'_B = cg_B$. LoRA-GACT sets $\eta'_A = c^2\eta_t$ and $\eta'_B = \eta_t/c^2$, and therefore

$$\begin{aligned}(A')^+ &= cA - \eta'_A g'_A \\ &= cA - (c^2\eta_t)\frac{g_A}{c} \\ &= c(A - \eta_t g_A) = cA^+,\end{aligned}$$

and

$$\begin{aligned}(B')^+ &= \frac{B}{c} - \eta'_B g'_B \\ &= \frac{B}{c} - \frac{\eta_t}{c^2}(cg_B) \\ &= \frac{B - \eta_t g_B}{c} = \frac{B^+}{c}.\end{aligned}$$

This proves the factor part of one-step commutation.

Next prove uniqueness. Allow arbitrary scalar learning rates $\widetilde{\eta}_A$ and $\widetilde{\eta}_B$ on the A and B sides. Requiring the A side to equal the transformed reference, with $g_A \neq 0$, is equivalent to

$$cA - \widetilde{\eta}_A \frac{g_A}{c} = cA - c\eta_t g_A.$$

Canceling cA and comparing scalar coefficients of the nonzero matrix g_A gives $\widetilde{\eta}_A/c = c\eta_t$, namely $\widetilde{\eta}_A = c^2\eta_t$. Similarly, for $g_B \neq 0$, the B side requires

$$\frac{B}{c} - \widetilde{\eta}_B(cg_B) = \frac{B}{c} - \frac{\eta_t}{c} g_B,$$

which gives $\widetilde{\eta}_B c = \eta_t/c$ and hence $\widetilde{\eta}_B = \eta_t/c^2$. Thus, when both factor gradients are nonzero, this coefficient pair is unique. Sufficiency gives one-step commutation, and Lemma 4.7 gives the full multi-step result. □

A.4 Momentum SGD State and Trajectory Restoration

Assumptions and notation. Consider the scalar handoff $A' = cA$, $B' = B/c$ with $c > 0$. For $p \in \{A, B\}$, let u_p denote a Momentum buffer with the same shape as p, let g_p be the factor gradient at the current continuation step, and let μ be the momentum coefficient. The reference recurrence is

$$u_p^+ = \mu u_p + g_p, \qquad p^+ = p - \eta_t u_p^+.$$

The two paths use the same batch, random state, step, and μ.

Theorem (Scalar Momentum SGD State and Trajectory Restoration (corresponding to Theorem 4.9))**.** If the handoff also transports

$$u'_A = u_A/c, \qquad u'_B = cu_B,$$

and uses

$$\eta'_{A,t} = c^2\eta_t, \qquad \eta'_{B,t} = \eta_t/c^2,$$

then

$$(u'_A)^+ = u_A^+/c, \qquad (u'_B)^+ = cu_B^+,$$

and

$$(A')^+ = cA^+, \qquad (B')^+ = B^+/c.$$

Thus the state transport and factor update commute step by step, and Lemma 4.7 restores a continuation trajectory of arbitrary length.

Complete proof of Theorem 4.9. Each reference factor p uses

$$u_p^+ = \mu u_p + g_p, \qquad p^+ = p - \eta_t u_p^+.$$

Consider the A side. After handoff, set $u'_A = u_A/c$ and $g'_A = g_A/c$. Then

$$\begin{aligned}(u'_A)^+ &= \mu u'_A + g'_A \\ &= \mu \frac{u_A}{c} + \frac{g_A}{c} \\ &= \frac{\mu u_A + g_A}{c} = \frac{u_A^+}{c}.\end{aligned}$$

Similarly, on the B side $u'_B = cu_B$ and $g'_B = cg_B$, so

$$(u'_B)^+ = \mu c u_B + c g_B = c u_B^+.$$

The parameter update uses $\eta'_A = c^2\eta_t$ and $\eta'_B = \eta_t/c^2$:

$$\begin{aligned}(A')^+ &= cA - c^2\eta_t \frac{u_A^+}{c} \\ &= c(A - \eta_t u_A^+) = cA^+, \\ (B')^+ &= \frac{B}{c} - \frac{\eta_t}{c^2}(cu_B^+) \\ &= \frac{B - \eta_t u_B^+}{c} = \frac{B^+}{c}.\end{aligned}$$

Thus the state and factors both satisfy the Φ_c transformation at the same step. Repeating the calculation at the next step, or applying Lemma 4.7, gives full trajectory restoration. □

A.5 Scaling-Aware Full-Matrix Transport

Assumptions and notation. Let $G \in \mathbb{R}^{r\times r}$ be invertible, let the LoRA scaling before and after handoff be $s, s' > 0$, and set $q = s/s'$. Consider

$$A' = GA, \qquad B' = qBG^{-1}.$$

Then $s'B'A' = sBA$. Let $N = \partial L/\partial \Delta W$, where $\Delta W = sBA = s'B'A'$, and let g_A, g_B, g'_A, g'_B denote the parameter gradients produced by the corresponding scaled LoRA modules. Thus the gradients include the active LoRA scaling rather than an unscaled gradient of BA.

Theorem (Exact Scaling-Aware Full-Matrix Plain-SGD and Momentum Transport)**.** Under the assumptions above, the factor gradients satisfy

$$g'_A = G^{-\mathsf{T}} g_A, \qquad g'_B = q^{-1} g_B G^{\mathsf{T}}.$$

For reference plain SGD, $A^+ = A - \eta_t g_A$ and $B^+ = B - \eta_t g_B$. Define

$$P_A = GG^{\mathsf{T}}, \qquad R_B = q^2 G^{-\mathsf{T}} G^{-1},$$

and update the transformed factors by

$$(A')^+ = A' - \eta_t P_A g'_A, \qquad (B')^+ = B' - \eta_t g'_B R_B.$$

Then

$$(A')^+ = GA^+, \qquad (B')^+ = qB^+G^{-1}, \qquad s'(B')^+(A')^+ = sB^+A^+.$$

For linear Momentum SGD, transport the buffers as

$$u'_A = G^{-\mathsf{T}} u_A, \qquad u'_B = q^{-1} u_B G^{\mathsf{T}}.$$

The same P_A, R_B make both the Momentum recurrence and parameter update commute step by step.

Proof. Because the model uses the scaled update $\Delta W = sBA$, the reference parameter gradients are

$$g_A = sB^\mathsf{T}N, \qquad g_B = sNA^\mathsf{T}.$$

Since $s' = s/q$, the transformed gradients are

$$g'_A = s'(B')^\mathsf{T}N = \frac{s}{q}\left(qG^{-\mathsf{T}}B^\mathsf{T}\right)N = G^{-\mathsf{T}}g_A,$$
$$g'_B = s'N(A')^\mathsf{T} = \frac{s}{q}NA^\mathsf{T}G^\mathsf{T} = q^{-1}g_BG^\mathsf{T}.$$

Consequently,

$$(A')^+ = GA - \eta_t GG^\mathsf{T}G^{-\mathsf{T}}g_A = G(A - \eta_t g_A) = GA^+,$$
$$\begin{aligned}(B')^+ &= qBG^{-1} - \eta_t(q^{-1}g_BG^\mathsf{T})(q^2G^{-\mathsf{T}}G^{-1}) \\ &= q(B - \eta_t g_B)G^{-1} = qB^+G^{-1}.\end{aligned}$$

Multiplying the two identities and using $s'q = s$ proves equality of the scaled functional updates.

For Momentum SGD, let $u_p^+ = \mu u_p + g_p$. Gradient covariance and the transported buffers give

$$(u'_A)^+ = G^{-\mathsf{T}}u_A^+, \qquad (u'_B)^+ = q^{-1}u_B^+G^\mathsf{T}.$$

Replacing g'_A, g'_B by these updated buffers in the calculation above gives $(A')^+ = GA^+$ and $(B')^+ = qB^+G^{-1}$ at the current step. Induction therefore gives exact multi-step restoration for this registered full-matrix plain-SGD or Momentum recurrence in exact arithmetic. □

Existence and recovery of the coordinate map. The theorem assumes that G is known and invertible. If the two checkpoints are minimal rank-r factorizations of the same rank-r scaled update, their factor row spaces coincide, so an invertible G exists. With full-row-rank A, it can be recovered as $G = A'A^\dagger$ and must be verified by both $A' = GA$ and $B' = qBG^{-1}$. Without these rank and verification conditions, equality of two products alone does not justify an invertible coordinate map.

Approximate reconstruction. PEFT-SVD reconstruction is finite precision rather than an exact coordinate identity. Write

$$A' = GA + E_A, \qquad B' = qBG^{-1} + E_B.$$

Then

$$\begin{aligned}\|s'B'A' - sBA\|_F \le s'\big(&q\|BG^{-1}\|_F\|E_A\|_F + \|E_B\|_F\|GA\|_F \\ &+ \|E_B\|_F\|E_A\|_F\big).\end{aligned}$$

Thus the exact theorem applies to the ideal full-matrix coordinate map, while the measured reconstruction and rounding errors enter the initial handoff error and local exchange residual defined in Appendix A.10. The PEFT-SVD results report this approximate finite-precision realization rather than an exact-arithmetic identity.

A.6 Coupled-L_2 State Recurrence and Exact Transport

Assumptions and notation. Under the same scalar handoff and Momentum recurrence, suppose that the reference forms the effective buffer gradients

$$q_A = g_A + \lambda_t A, \qquad q_B = g_B + \lambda_t B,$$

and uses $u_p^+ = \mu u_p + q_p$, $p^+ = p - \eta_{p,t}u_p^+$. Here λ_t is the coupled-L_2 coefficient in the reference; the converted path may use separate λ' values for the two factor groups.

Proposition (Exact Coupled-L_2 Transport under Momentum SGD)**.** Under

$$\lambda'_{A,t} = \lambda_t/c^2, \qquad \lambda'_{B,t} = c^2\lambda_t$$

and the state and learning-rate transport from the Momentum SGD subsection,

$$q'_A = q_A/c, \qquad q'_B = cq_B.$$

Therefore the coupled gradient, Momentum buffer, and factor update satisfy

$$(u'_A)^+ = u_A^+/c, \qquad (u'_B)^+ = cu_B^+, \qquad (A')^+ = cA^+, \qquad (B')^+ = B^+/c.$$

When $\mu = 0$, the result reduces to coupled-L_2 plain SGD.

Complete proof of the coupled-L_2 transport proposition. Write coupled-L_2 as the effective gradient entering the momentum buffer:

$$q_A = g_A + \lambda_t A, \qquad q_B = g_B + \lambda_t B.$$

In transformed coordinates use $A' = cA$, $B' = B/c$, $\lambda'_A = \lambda_t/c^2$, $\lambda'_B = c^2\lambda_t$. By gradient covariance,

$$\begin{aligned} q'_A &= g'_A + \lambda'_A A' \\ &= \frac{g_A}{c} + \frac{\lambda_t}{c^2}(cA) \\ &= \frac{q_A}{c}, \end{aligned}$$

and

$$\begin{aligned} q'_B &= g'_B + \lambda'_B B' \\ &= c g_B + (c^2\lambda_t)\frac{B}{c} \\ &= c q_B. \end{aligned}$$

The momentum recurrence then satisfies

$$(u'_A)^+ = \mu(u_A/c) + q_A/c = u_A^+/c, \qquad (u'_B)^+ = \mu(cu_B) + cq_B = cu_B^+.$$

Using the same $\eta'_A = c^2\eta_t$ and $\eta'_B = \eta_t/c^2$ as in Theorem 4.9, we obtain

$$(A')^+ = cA^+, \qquad (B')^+ = B^+/c.$$

Thus the coupled gradient, momentum state, and factor-parameter update commute in one step. When $\mu = 0$, the buffer retains only the current effective gradient and the derivation still holds; the multi-step conclusion follows from Lemma 4.7. □

A.7 Decoupled Decay Exchange Condition

Assumptions and notation. Write the decoupled-decay map as

$$D_t(p) = (1 - \eta_t\lambda_t)p.$$

For the scalar handoff, let $p' = s_p p$ with $s_A = c$, $s_B = 1/c$, and $c > 0$. This subsection first compares the decay map itself and then combines it with the gradient/state update that already satisfies commutation.

Corollary (Exact Transport Condition for Decoupled Decay)**.** For any factor parameter p, let the converted path use $\eta'_{p,t}$ and $\lambda'_{p,t}$. Then

$$D'_t(p') = T_c(D_t(p))$$

for every p if and only if

$$\eta'_{A,t}\lambda'_{A,t} = \eta_t\lambda_t, \qquad \eta'_{B,t}\lambda'_{B,t} = \eta_t\lambda_t.$$

Thus, with the plain-SGD or Momentum rates

$$\eta'_A = c^2\eta_t, \qquad \eta'_B = \eta_t/c^2,$$

use

$$\lambda'_A = \lambda_t/c^2, \qquad \lambda'_B = c^2\lambda_t;$$

with the AdamW rates

$$\eta'_A = c\eta_t, \qquad \eta'_B = \eta_t/c,$$

use

$$\lambda'_A = \lambda_t/c, \qquad \lambda'_B = c\lambda_t.$$

Together with the corresponding gradient/state transport, these conditions give complete one-step commutation.

Complete proof of the decoupled-decay transport corollary. First compare only the decoupled-decay term. Let the coordinate scale factors for A and B be $s_A = c$ and $s_B = 1/c$, so $p' = s_p p$. The reference coordinate after decay is

$$p^+_{\text{ref}} = (1 - \eta_t \lambda_t)p,$$

Transporting it to the new coordinates gives

$$T_c(p^+_{\text{ref}}) = s_p(1 - \eta_t \lambda_t)p.$$

The converted path first changes coordinates and then applies local decay, giving

$$p^+_{\text{new}} = (1 - \eta'_p \lambda'_p)s_p p.$$

For any nonzero p, the two expressions are equal if and only if

$$(1 - \eta'_p \lambda'_p)s_p = s_p(1 - \eta_t \lambda_t),$$

Since $s_p > 0$, this is equivalent to $\eta'_p \lambda'_p = \eta_t \lambda_t$. Taking $p = A$ and $p = B$ gives the decay-invariance condition above.

With the plain-SGD or Momentum learning-rate transport, $\eta'_A = c^2 \eta_t$ and $\eta'_B = \eta_t / c^2$, the product condition gives

$$\lambda'_A = \lambda_t / c^2, \qquad \lambda'_B = c^2 \lambda_t.$$

With AdamW's $\eta'_A = c\eta_t$ and $\eta'_B = \eta_t / c$, it gives

$$\lambda'_A = \lambda_t / c, \qquad \lambda'_B = c\lambda_t.$$

These conditions make the decay term commute together with the already established gradient/state update. □

A.8 Coordinate-wise AdamW State and Trajectory Restoration

Assumptions and notation. Consider the same scalar handoff $A' = cA$, $B' = B/c$ with $c > 0$. For each factor tensor $p \in \{A, B\}$, let m_p and v_p denote the first- and second-moment optimizer states, let $\beta_1, \beta_2 \in [0, 1)$, let $\epsilon > 0$, and let k be the current bias-correction step. The reference coordinate-wise AdamW recurrence is

$$\begin{aligned} m^+_p &= \beta_1 m_p + (1 - \beta_1) g_p, \\ v^+_p &= \beta_2 v_p + (1 - \beta_2) g_p \odot g_p, \\ p^+ &= (1 - \eta_t \lambda_t)p - \eta_t \frac{\widehat{m}^+_p}{\sqrt{\widehat{v}^+_p} + \epsilon}, \end{aligned}$$

where

$$\widehat{m}^+_p = \frac{m^+_p}{1 - \beta_1^k}, \qquad \widehat{v}^+_p = \frac{v^+_p}{1 - \beta_2^k}.$$

The two paths share k, β_1, β_2, the batch, and the random state.

Theorem (Exact Coordinate-wise AdamW Transport (corresponding to Theorem 4.10))**.** If the handoff transports

$$m'_A = m_A / c, \quad v'_A = v_A / c^2, \qquad m'_B = c m_B, \quad v'_B = c^2 v_B,$$

and synchronizes the update metadata

$$\begin{aligned} \eta'_A &= c\eta_t, \quad \epsilon'_A = \epsilon / c, \quad \lambda'_A = \lambda_t / c, \\ \eta'_B &= \eta_t / c, \quad \epsilon'_B = c\epsilon, \quad \lambda'_B = c\lambda_t, \end{aligned}$$

then the first-moment state, second-moment state, bias correction, normalized update, and decoupled decay satisfy commutation, with

$$(A')^+ = cA^+, \qquad (B')^+ = B^+ / c.$$

Lemma 4.7 therefore restores a continuation trajectory of arbitrary length.

Complete proof of Theorem 4.10. Let the reference state before the current step be m_A, v_A, m_B, v_B, and assume that handoff sets

$$m'_A = m_A/c, \quad v'_A = v_A/c^2, \qquad m'_B = cm_B, \quad v'_B = c^2 v_B.$$

By Lemma 4.3, $g'_A = g_A/c$ and $g'_B = cg_B$.

First handle the A-side first-moment state:

$$\begin{aligned}(m'_A)^+ &= \beta_1 m'_A + (1-\beta_1) g'_A \\ &= \beta_1 \frac{m_A}{c} + (1-\beta_1)\frac{g_A}{c} \\ &= \frac{m_A^+}{c}.\end{aligned}$$

The second-moment state satisfies

$$\begin{aligned}(v'_A)^+ &= \beta_2 \frac{v_A}{c^2} + (1-\beta_2)\left(\frac{g_A}{c}\right) \odot \left(\frac{g_A}{c}\right) \\ &= \frac{\beta_2 v_A + (1-\beta_2) g_A \odot g_A}{c^2} \\ &= \frac{v_A^+}{c^2}.\end{aligned}$$

The B side correspondingly satisfies

$$\begin{aligned}(m'_B)^+ &= \beta_1(cm_B) + (1-\beta_1)(cg_B) = cm_B^+, \\ (v'_B)^+ &= \beta_2(c^2 v_B) + (1-\beta_2)(cg_B) \odot (cg_B) = c^2 v_B^+.\end{aligned}$$

Let the current bias-correction step be k. Both paths share k, β_1, and β_2, so the denominators $d_1 = 1-\beta_1^k$ and $d_2 = 1-\beta_2^k$ are identical, and

$$\begin{aligned}\widehat{m}'_A{}^+ &= \widehat{m}_A^+/c, \quad \widehat{v}'_A{}^+ = \widehat{v}_A^+/c^2, \\ \widehat{m}'_B{}^+ &= c\widehat{m}_B^+, \quad \widehat{v}'_B{}^+ = c^2\widehat{v}_B^+.\end{aligned}$$

Since $c > 0$,

$$\sqrt{\widehat{v}'_A{}^+} = \frac{\sqrt{\widehat{v}_A^+}}{c}, \qquad \sqrt{\widehat{v}'_B{}^+} = c\sqrt{\widehat{v}_B^+}.$$

With $\epsilon'_A = \epsilon/c$, the normalized A-side update is unchanged:

$$\begin{aligned}\frac{\widehat{m}'_A{}^+}{\sqrt{\widehat{v}'_A{}^+} + \epsilon'_A} &= \frac{\widehat{m}_A^+/c}{\sqrt{\widehat{v}_A^+}/c + \epsilon/c} \\ &= \frac{\widehat{m}_A^+}{\sqrt{\widehat{v}_A^+} + \epsilon}.\end{aligned}$$

With $\epsilon'_B = c\epsilon$, the B side likewise satisfies

$$\frac{\widehat{m}'_B{}^+}{\sqrt{\widehat{v}'_B{}^+} + \epsilon'_B} = \frac{c\widehat{m}_B^+}{c\sqrt{\widehat{v}_B^+} + c\epsilon} = \frac{\widehat{m}_B^+}{\sqrt{\widehat{v}_B^+} + \epsilon}.$$

Finally compare the decay and normalized-gradient terms. From $\eta'_A = c\eta$, $\lambda'_A = \lambda/c$, and $\eta'_B = \eta/c$, $\lambda'_B = c\lambda$,

$$\eta'_A \lambda'_A = \eta'_B \lambda'_B = \eta\lambda.$$

Let the normalized reference gradients be h_A and h_B. Then

$$\begin{aligned}(A')^+ &= (1-\eta\lambda)cA - (c\eta)h_A \\ &= c\big((1-\eta\lambda)A - \eta h_A\big) = cA^+,\end{aligned}$$

and

$$\begin{aligned}(B')^+ &= (1-\eta\lambda)\frac{B}{c} - \frac{\eta}{c}h_B \\ &= \frac{(1-\eta\lambda)B - \eta h_B}{c} = \frac{B^+}{c}.\end{aligned}$$

Thus the m, v states, bias correction, epsilon, decay, and factor update all commute in one step; Lemma 4.7 then gives the complete continuation trajectory. □

A.9 Positive-Diagonal Closure, Main Theorem, and Multi-Module Composition

Assumptions and notation. Let $C = \mathrm{diag}(c_1, \ldots, c_r)$ with each $c_i > 0$, and define

$$A' = CA, \qquad B' = BC^{-1}.$$

Let a_i denote the ith row of A and b_i the ith column of B; then $a'_i = c_i a_i$ and $b'_i = b_i / c_i$. Each latent direction uses the scalar rules above, with its state, epsilon, learning rate, and decay transported in the AdamW case.

Corollary (Exact Positive-Diagonal Continuation Transport (corresponding to Corollary 4.11))**.** Under these conditions, the factor, optimizer state, and update metadata of each latent direction satisfy the corresponding scalar one-step commutation relation. Concatenating the relations for $i = 1, \ldots, r$ in block coordinates gives

$$T_C \circ U_t = U_t^{(C)} \circ T_C,$$

and Lemma 4.7 restores the multi-step continuation after a positive-diagonal handoff.

Theorem (LoRA-GACT Checkpoint Continuation Restoration (corresponding to Theorem 4.12))**.** Consider the registered continuation system defined in the main text and a positive scalar or positive-diagonal handoff with per-direction metadata. When the handoff transports factors, optimizer state, and update metadata according to the registered rule, for any continuation length

$$X'_t = \Phi(X_t), \qquad sB'_t A'_t = sB_t A_t, \qquad t_0 \le t \le T.$$

Under the same frozen base model, subsequent batches, random states, scheduler, and step counter, the model weights, responses, and downstream decisions determined by those responses agree step by step.

Corollary (Closure for Multi-Module LoRA (corresponding to Corollary 4.13))**.** Suppose that the model contains K LoRA modules with products $F_k = B_k A_k$. Each module may use a scalar or positive-diagonal handoff, and the registered optimizer recurrence separates by factor tensor or latent direction. If every module satisfies the LoRA-GACT one-step commutation relation, the joint transport

$$\Phi = \Phi_1 \times \cdots \times \Phi_K$$

satisfies the joint update commutation relation and restores the continuation trajectory, responses, and deterministic downstream decisions of the joint model.

Complete proof of Corollary 4.11. Let $C = \mathrm{diag}(c_1, \ldots, c_r)$ with $c_i > 0$. Denote the ith row of A by $a_i = A_{i,:}$ and the ith column of B by $b_i = B_{:,i}$. Then

$$A' = CA, \qquad B' = BC^{-1}, \qquad a'_i = c_i a_i, \qquad b'_i = b_i / c_i.$$

Since $BA = \sum_i b_i a_i$, each latent direction satisfies term by term

$$b'_i a'_i = (b_i / c_i)(c_i a_i) = b_i a_i,$$

Thus $B'A' = BA$. On the same input, the functional gradient $M = \partial \ell / \partial (BA)$ remains unchanged; the chain rule gives

$$g'_A = C^{-1} g_A, \qquad g'_B = g_B C,$$

namely

$$g'_{a_i} = g_{a_i} / c_i, \qquad g'_{b_i} = c_i g_{b_i}.$$

Plain SGD. The ith direction of the reference path is

$$a_i^+ = a_i - \eta g_{a_i}, \qquad b_i^+ = b_i - \eta g_{b_i}.$$

Let the converted path use direction-wise learning rates $\eta'_{a_i} = c_i^2 \eta$ and $\eta'_{b_i} = \eta / c_i^2$, then

$$\begin{aligned}
(a'_i)^+ &= c_i a_i - \eta'_{a_i} g'_{a_i} = c_i a_i - (c_i^2 \eta)(g_{a_i} / c_i) \\
&= c_i (a_i - \eta g_{a_i}) = c_i a_i^+, \\
(b'_i)^+ &= b_i / c_i - \eta'_{b_i} g'_{b_i} = b_i / c_i - (\eta / c_i^2)(c_i g_{b_i}) \\
&= \frac{b_i - \eta g_{b_i}}{c_i} = b_i^+ / c_i.
\end{aligned}$$

Thus the one-step plain-SGD update commutes in every direction.

Momentum SGD. Let the reference buffers be u_{a_i}, u_{b_i}, with recurrence

$$u_{a_i}^+ = \mu u_{a_i} + g_{a_i}, \qquad u_{b_i}^+ = \mu u_{b_i} + g_{b_i},$$
$$a_i^+ = a_i - \eta u_{a_i}^+, \qquad b_i^+ = b_i - \eta u_{b_i}^+.$$

Transport the buffers as

$$u'_{a_i} = u_{a_i}/c_i, \qquad u'_{b_i} = c_i u_{b_i}.$$

Then

$$(u'_{a_i})^+ = \mu(u_{a_i}/c_i) + g_{a_i}/c_i = u_{a_i}^+/c_i,$$
$$(u'_{b_i})^+ = \mu(c_i u_{b_i}) + c_i g_{b_i} = c_i u_{b_i}^+.$$

Substituting the direction-wise learning rates above gives

$$(a'_i)^+ = c_i a_i^+, \qquad (b'_i)^+ = b_i^+/c_i.$$

Coupled-L_2. Let the effective gradient entering the buffer be

$$q_{a_i} = g_{a_i} + \lambda a_i, \qquad q_{b_i} = g_{b_i} + \lambda b_i.$$

The converted path uses $\lambda'_{a_i} = \lambda/c_i^2$ and $\lambda'_{b_i} = c_i^2\lambda$. Term-by-term calculation gives

$$q'_{a_i} = g'_{a_i} + \lambda'_{a_i} a'_i = g_{a_i}/c_i + (\lambda/c_i^2)(c_i a_i) = q_{a_i}/c_i,$$
$$q'_{b_i} = g'_{b_i} + \lambda'_{b_i} b'_i = c_i g_{b_i} + (c_i^2\lambda)(b_i/c_i) = c_i q_{b_i}.$$

Therefore the Momentum buffer recurrence and factor update satisfy

$$(u'_{a_i})^+ = u_{a_i}^+/c_i, \qquad (u'_{b_i})^+ = c_i u_{b_i}^+, \qquad (a'_i)^+ = c_i a_i^+, \qquad (b'_i)^+ = b_i^+/c_i.$$

When $\mu = 0$, this result reduces directly to coupled-L_2 plain SGD.

Coordinate-wise AdamW. We now fully expand the AdamW recurrence for one latent direction. Each element of a_i and b_i maintains a first-moment state m and a second-moment state v:

$$m_{a_i}^+ = \beta_1 m_{a_i} + (1-\beta_1) g_{a_i},$$
$$v_{a_i}^+ = \beta_2 v_{a_i} + (1-\beta_2) g_{a_i} \odot g_{a_i},$$
$$m_{b_i}^+ = \beta_1 m_{b_i} + (1-\beta_1) g_{b_i},$$
$$v_{b_i}^+ = \beta_2 v_{b_i} + (1-\beta_2) g_{b_i} \odot g_{b_i}.$$

For the converted path, direction-wise transport is

$$m'_{a_i} = m_{a_i}/c_i, \quad v'_{a_i} = v_{a_i}/c_i^2, \qquad m'_{b_i} = c_i m_{b_i}, \quad v'_{b_i} = c_i^2 v_{b_i}.$$

Therefore,

$$(m'_{a_i})^+ = m_{a_i}^+/c_i, \qquad (v'_{a_i})^+ = v_{a_i}^+/c_i^2,$$
$$(m'_{b_i})^+ = c_i m_{b_i}^+, \qquad (v'_{b_i})^+ = c_i^2 v_{b_i}^+.$$

Let the current bias-correction step be k. Because the two paths share k, β_1, β_2,

$$\widehat{m}'^{+}_{a_i} = \widehat{m}^+_{a_i}/c_i, \quad \widehat{v}'^{+}_{a_i} = \widehat{v}^+_{a_i}/c_i^2, \quad \widehat{m}'^{+}_{b_i} = c_i \widehat{m}^+_{b_i}, \quad \widehat{v}'^{+}_{b_i} = c_i^2 \widehat{v}^+_{b_i}.$$

Set the direction-wise numerical metadata to

$$\eta'_{a_i} = c_i\eta, \quad \epsilon'_{a_i} = \epsilon/c_i, \quad \lambda'_{a_i} = \lambda/c_i,$$
$$\eta'_{b_i} = \eta/c_i, \quad \epsilon'_{b_i} = c_i\epsilon, \quad \lambda'_{b_i} = c_i\lambda.$$

Since $c_i > 0$, the normalized updates satisfy

$$h'_{a_i} := \frac{\widehat{m}'^{+}_{a_i}}{\sqrt{\widehat{v}'^{+}_{a_i}} + \epsilon'_{a_i}} = \frac{\widehat{m}^+_{a_i}/c_i}{\sqrt{\widehat{v}^+_{a_i}}/c_i + \epsilon/c_i} = h_{a_i},$$
$$h'_{b_i} := \frac{\widehat{m}'^{+}_{b_i}}{\sqrt{\widehat{v}'^{+}_{b_i}} + \epsilon'_{b_i}} = \frac{c_i\widehat{m}^+_{b_i}}{c_i\sqrt{\widehat{v}^+_{b_i}} + c_i\epsilon} = h_{b_i}.$$

Finally, the AdamW parameter update is

$$a_i^+ = (1-\eta\lambda)a_i - \eta h_{a_i}, \qquad b_i^+ = (1-\eta\lambda)b_i - \eta h_{b_i}.$$

The converted path gives

$$\begin{aligned}(a_i')^+ &= (1-\eta'_{a_i}\lambda'_{a_i})c_i a_i - \eta'_{a_i} h'_{a_i} \\ &= c_i(1-\eta\lambda)a_i - c_i\eta h_{a_i} = c_i a_i^+, \\ (b_i')^+ &= (1-\eta'_{b_i}\lambda'_{b_i})b_i/c_i - \eta'_{b_i} h'_{b_i} \\ &= \frac{(1-\eta\lambda)b_i - \eta h_{b_i}}{c_i} = b_i^+/c_i.\end{aligned}$$

This proves simultaneous commutation of m, v, bias correction, ϵ, decay, and the parameter update for one direction. AdamW is elementwise, so latent directions do not mix; concatenating $i = 1, \ldots, r$ gives the matrix-level commutation relation. The derivation also gives the protocol requirement: when the c_i differ, ϵ, learning rate, and decay must be transported direction by direction; retaining them as shared scalars generally breaks the denominator identity above.

Concatenation and the multi-step conclusion. For every i, the four recurrences above satisfy

$$T_{c_i} \circ U_{i,t} = U_{i,t}^{(c_i)} \circ T_{c_i}.$$

Concatenating the factors, state, and metadata of all latent directions in block coordinates gives

$$T_C \circ U_t = U_t^{(C)} \circ T_C.$$

Applying Lemma 4.7 yields multi-step continuation restoration for the positive-diagonal handoff. □

Complete proof of Theorem 4.12. First consider a scalar handoff. Plain-SGD one-step commutation follows from Theorem 4.8; commutation with linear momentum state follows from Theorem 4.9; the complete commutation derivations for coupled L_2, decoupled decay, and AdamW are given in the corresponding subsections of this appendix. Thus, for every registered optimizer recurrence, a joint map Φ_c over factors, state, and update metadata satisfies

$$\Phi_c \circ U_t = U_t^{(c)} \circ \Phi_c.$$

For a positive-diagonal handoff, Corollary 4.11 gives the same one-step commutation in every latent direction. Since C is positive diagonal, all directions can be concatenated in block coordinates to form a joint map Φ_C satisfying

$$\Phi_C \circ U_t = U_t^{(C)} \circ \Phi_C.$$

The handoff construction itself satisfies $X'_{t_0} = \Phi(X_{t_0})$. Lemma 4.7 therefore gives, for every $t_0 \le t \le T$,

$$X'_t = \Phi(X_t).$$

Taking the factor part of Φ, the scalar case gives $B'_t A'_t = B_t A_t$, while the positive-diagonal case gives $(B_t C^{-1})(C A_t) = B_t A_t$. Multiplying by the fixed LoRA scaling s gives $s B'_t A'_t = s B_t A_t$.

The frozen base model is identical on both paths, and the continuation system uses the same subsequent batches, random states, scheduler sequence, and step counter. Thus the complete model weights and forward responses match at every step, as do all deterministic decisions determined by those responses. This completes the proof from one-step update commutation to checkpoint continuation restoration of arbitrary length. □

Complete proof of Corollary 4.13. Write the K LoRA modules in the model as

$$F_k = B_k A_k, \qquad k = 1, \ldots, K.$$

Choose an independent scalar or positive-diagonal handoff for each module, and let Φ_k denote transport of that module together with its optimizer state and update metadata. Since each module product is preserved,

$$F'_k = F_k \qquad \text{for all } k,$$

each modified layer weight and therefore the full-model forward output are unchanged at handoff.

With the same continuation batches and random states, the loss is the same function of all module products. Hence the functional gradient $M_{k,t} = \partial \ell_t / \partial F_{k,t}$ is the same on both paths, and its factor gradients transform covariantly under the scalar or diagonal rule for that module. Because the assumed optimizer recurrence separates by factor tensor or latent direction, every k satisfies

$$\Phi_k \circ U_{k,t} = U'_{k,t} \circ \Phi_k.$$

Let the joint state transport be $\Phi = \Phi_1 \times \cdots \times \Phi_K$. The one-step relation holds for all modules simultaneously, so the joint state satisfies

$$\Phi \circ U_t = U'_t \circ \Phi.$$

The joint handoff satisfies $X'_{t_0} = \Phi(X_{t_0})$. Lemma 4.7 then shows inductively that every module retains its corresponding transformed reference state at every subsequent step. Each $F_{k,t}$ therefore matches the reference, so the model weights, responses, and deterministic downstream decisions also match the reference. □

A.10 Finite-Precision Error and Multi-Step Residuals

The exchange relations above give exact equalities in exact arithmetic. An actual checkpoint handoff uses finite-precision scaling, matrix multiplication, optimizer-state serialization, and forward computation, so the theoretical exchange and implementation error must be measured separately. This section defines the errors, their propagation, and the corresponding experimental gates.

Local exchange residual. Write a continuation state as

$$X_t = (A_t, B_t, S_t, \Omega_t),$$

where S_t denotes the optimizer state and Ω_t denotes update metadata such as the learning rate, ϵ, decay, scheduler, and step counter. For the continuous tensor and scalar components of X_t, fix a product norm, for example the square root of the sum of squared Frobenius and Euclidean component norms; every state norm below refers to this fixed norm. Discrete fields such as the step counter and scheduler index are compared for exact equality and are not embedded in the continuous norm. Let T be the exact handoff transport, U_t the reference optimizer recurrence, and U'_t the recurrence after transformation. The preceding results establish

$$T \circ U_t = U'_t \circ T.$$

In a floating-point implementation, denote the implemented transport and updates by $\widehat{T}$, $\widehat{U}_t$, and $\widehat{U}'_t$, respectively. Given the reference implementation state $\widehat{X}_t$, define the local exchange residual at step t by

$$\rho_t = \left\| \widehat{U}'_t\Big(\widehat{T}(\widehat{X}_t)\Big) - \widehat{T}\Big(\widehat{U}_t(\widehat{X}_t)\Big) \right\|.$$

This quantity includes factor scaling, state scaling, numerical metadata, and the rounding error of one update; it does not mix trajectory error accumulated in later steps into the single-step measurement.

Let the state error between the transformed path and the transported reference path be

$$e_t = \left\| \widehat{X}'_t - \widehat{T}(\widehat{X}_t) \right\|.$$

If the finite-precision update after transformation satisfies the Lipschitz condition in the current state neighborhood

$$\left\| \widehat{U}'_t(X) - \widehat{U}'_t(Y) \right\| \le L_t \|X - Y\|,$$

the triangle inequality gives

$$\begin{aligned} e_{t+1} &= \left\| \widehat{U}'_t(\widehat{X}'_t) - \widehat{T}(\widehat{U}_t(\widehat{X}_t)) \right\| \\ &\le \left\| \widehat{U}'_t(\widehat{X}'_t) - \widehat{U}'_t(\widehat{T}(\widehat{X}_t)) \right\| + \rho_t \\ &\le L_t e_t + \rho_t. \end{aligned}$$

Recursing yields

$$e_T \le \left(\prod_{j=t_0}^{T-1} L_j \right) e_{t_0} + \sum_{k=t_0}^{T-1} \rho_k \prod_{j=k+1}^{T-1} L_j.$$

Thus, the initial handoff error, the single-step exchange residual, and the local amplification of the optimizer update jointly determine the final trajectory error. The theoretical exchange relation of LoRA-GACT attributes the algorithmic residual to implementation precision. The reported experiments measure handoff and trajectory errors; ρ_t can additionally be evaluated by the paired one-step computation in its definition.

Rounding term for product computation. Let u be the floating-point unit roundoff and let r be the inner dimension of the matrix multiplication, with $ru < 1$. Under the standard rounding model and assuming no overflow or underflow, define

$$\gamma_r = \frac{ru}{1 - ru}.$$

Under the standard componentwise rounding model, if $\widehat{F}' = \mathrm{fl}(\widehat{B}'\widehat{A}')$, then

$$\left\| \widehat{F}' - \widehat{B}'\widehat{A}' \right\|_F \le \gamma_r \|\widehat{B}'\|_F \|\widehat{A}'\|_F.$$

Let the exact transformed factors be $B' = BC^{-1}$ and $A' = CA$, and let

$$E_A = \widehat{A}' - A', \qquad E_B = \widehat{B}' - B'.$$

Because $B'A' = BA$, we have

$$\|\widehat{F}' - BA\|_F \le \gamma_r \|\widehat{B}'\|_F \|\widehat{A}'\|_F + \|E_B\|_F \|A'\|_F + \|B'\|_F \|E_A\|_F + \|E_B\|_F \|E_A\|_F.$$

For a scalar handoff, $\|A'\|_F = c\|A\|_F$ and $\|B'\|_F = \|B\|_F/c$; for a positive-diagonal handoff, the corresponding quantities are $\|CA\|_F$ and $\|BC^{-1}\|_F$. This expression shows that checking the numerical errors of the two factor files alone does not establish function equivalence: BA and the logits on fixed inputs must also be checked directly.

Optimizer-state and metadata errors. For each state tensor s, define the serialization round-trip error by

$$\varepsilon_{\mathrm{state}}(s) = \frac{\|\widehat{\mathrm{load}}(\widehat{\mathrm{save}}(s)) - s\|_F}{\max(1, \|s\|_F)}.$$

For factors, state, and metadata at handoff, record

$$\varepsilon^{\Delta W}_{\mathrm{handoff}}, \quad \varepsilon_{\mathrm{logit}}, \quad \varepsilon_{\mathrm{state}}, \quad \varepsilon_{\mathrm{step}},$$

where the last term is a discrete consistency check for the step counter, scheduler position, and bias-correction index. Continuous-tensor errors enter ρ_t; inconsistency in the step, batch order, or bias-correction index is a continuation-system mismatch and is marked separately as a protocol error rather than treated as floating-point rounding error.

Experimental checks and layered gates. The reported records provide the following layered checks: among 705 real RoBERTa gauge records, the maximum relative error of ΔW is 7.66×10^{-16} and the maximum relative error of fixed-input outputs is 8.91×10^{-16}; the maximum recorded CPU-FP32 state round-trip error is zero; and the maximum absolute difference in the initial output response for I3/I4 initialization is 2.7907×10^{-6}. Across the continuation matrices for Qwen3, Qwen2.5, Llama, and RoBERTa, the final d_W of LoRA-GACT remains around 10^{-8}, while held-out output response and decision disagreement are both zero. The main text uses a 10^{-6} handoff gate and a 10^{-3} trajectory-restoration gate; the error recurrence in this section shows that these gates correspond to initial transport correctness and multi-step continuation correctness, respectively, rather than to the same numerical check.

When an implementation uses a different dtype, fused kernel, asynchronous communication, or cross-library transpose/scaling convention, the four errors above should be recorded again. The theoretical proofs provide the exact transport identity, while the experimental gates verify whether a given implementation preserves an acceptable finite-precision approximation of that identity.

Table 3: Complete baseline and common-reference details for scalar continuation. The first block is measured against each method's uninterrupted reference, and the second block against the listed target reference; values are 3-seed means. d_W, response, and dis denote relative functional-update distance, output response, and decision disagreement, respectively; dis is reported as a proportion, – marks an inapplicable protocol field, and state is recorded for the fixed model and rank.

Task	Arm / method	Target	Restored	d_W	Response	Dis	Trials/seed	State MiB
Own-reference baseline (target = each method's own optimizer)								
QQP	raw	own	0/3	0.0477	0.5239	–	0	0
QQP	global-LR selected	own	0/3	0.0468	0.5321	–	6	0
QQP	LoRA+	own	0/3	0.1077	1.0559	–	0	0
QQP	LoRA-RITE (own optimizer)	own	3/3	9.46×10^{-9}	0	–	0	12.28
QQP	LoRA-GACT	own	3/3	9.59×10^{-9}	0	–	0	0
RTE	raw	own	0/3	0.0294	0.4061	–	0	0
RTE	global-LR selected	own	0/3	0.0364	0.5243	–	6	0
RTE	LoRA+	own	0/3	0.0661	0.8164	–	0	0
RTE	LoRA-RITE (own optimizer)	own	3/3	8.85×10^{-9}	0	–	0	12.28
RTE	LoRA-GACT	own	3/3	9.13×10^{-9}	0	–	0	0
Common-reference comparison								
QQP	raw gauge	plain SGD	0/3	0.04765	0.52391	0.05581	–	0
QQP	LoRA-GACT	plain SGD	3/3	9.59×10^{-9}	0	0	–	0
QQP	RITE-after-SGD	plain SGD	0/3	7.4137	9.6868	0.48471	–	12.28
QQP	RITE-after-SGD+gauge	plain SGD	0/3	7.4137	9.6868	0.48471	–	12.28
QQP	RITE-from-start	RITE	3/3	9.29×10^{-9}	0	0	–	12.28
RTE	raw gauge	plain SGD	0/3	0.02939	0.40613	0.04556	–	0
RTE	LoRA-GACT	plain SGD	3/3	9.13×10^{-9}	0	0	–	0
RTE	RITE-after-SGD	plain SGD	0/3	8.1985	6.8901	0.40048	–	12.28
RTE	RITE-after-SGD+gauge	plain SGD	0/3	8.1985	6.8901	0.40048	–	12.28
RTE	RITE-from-start	RITE	3/3	9.54×10^{-9}	0	0	–	12.28

B Scalar Continuation Evidence

B.1 Complete Baseline and Common-Reference Details

This subsection records scalar-handoff evidence in three layers: reference accounting, per-seed records, and a task-level summary aligned with the main table. Together, the three tables check whether a function-equivalent handoff preserves subsequent continuation and whether LoRA-GACT restoration remains stable across tasks and seeds.

Reference accounting. Table 3 separates own-reference and common-reference evaluations. The former uses each method's uninterrupted continuation as the target and asks whether the method restores its own optimization process; the latter fixes plain-SGD continuation as the common target and asks whether a switch across optimizer/reference systems preserves that continuation. The table retains response, decision disagreement, restoration count, trials per seed, and state MiB so that behavior, pass counts, search cost, and state overhead can be checked together.

The own-reference block shows that raw, global-LR, and LoRA+ retain measurable trajectory error; LoRA-RITE and LoRA-GACT restore their respective optimizer references. The common-reference block separates the target systems: LoRA-GACT restores the original plain-SGD target, RITE-from-start restores the RITE target, and switching from plain SGD to RITE retains a measurable continuation gap. The target column explicitly identifies the continuation system for both blocks.

Per-seed results. Table 4 expands every QQP and RTE seed and records held-out loss, d_W, response, and decision disagreement. The task-level means are aggregated from these paired records, and each global-LR winner is selected independently within a seed before aggregation.

The per-seed records agree with the aggregate accounting: LoRA-GACT places d_W and response inside the restoration gate for all three seeds on both tasks, while raw and global-LR retain trajectory and output deviations.

Task-level summary. Table 5 gives the final three-seed task-level means and the complete numerical source for the main scalar result. The table reports held-out loss together with the three restoration checks—d_W, response, and dis; the trials-per-seed column records the extra cost of global learning-rate search.

Table 4: Qwen3-1.7B scalar experiment per-seed results. All reported arms are measured against the uninterrupted reference for the same seed.

Task	Seed	Arm	Held-out loss	d_W	Response / dis
QQP	42	raw	17.7499	0.04319	0.69182 / 0.06193
QQP	42	global-LR	16.6557	0.04187	0.53425 / 0.05505
QQP	42	LoRA-GACT	17.2821	1.19×10^{-8}	0 / 0
QQP	43	raw	17.0311	0.05093	0.48389 / 0.06193
QQP	43	global-LR	16.1985	0.05022	0.57472 / 0.05505
QQP	43	LoRA-GACT	16.5794	7.07×10^{-9}	0 / 0
QQP	44	raw	17.6321	0.04884	0.39603 / 0.05046
QQP	44	global-LR	16.6778	0.04826	0.48720 / 0.06422
QQP	44	LoRA-GACT	17.2377	9.79×10^{-9}	0 / 0
RTE	42	raw	14.5280	0.03581	0.40705 / 0.04317
RTE	42	global-LR	14.4381	0.03587	0.40058 / 0.00719
RTE	42	LoRA-GACT	14.4673	9.90×10^{-9}	0 / 0
RTE	43	raw	14.5615	0.02926	0.40894 / 0.05036
RTE	43	global-LR	14.3869	0.03780	0.66061 / 0.12230
RTE	43	LoRA-GACT	14.6273	9.30×10^{-9}	0 / 0
RTE	44	raw	14.4097	0.02311	0.40241 / 0.04317
RTE	44	global-LR	14.0588	0.03545	0.51158 / 0.04317
RTE	44	LoRA-GACT	14.3928	8.18×10^{-9}	0 / 0

Table 5: Task-level 3-seed means for the scalar gauge. For global-LR, the winner is selected from six candidates per seed on the selection split before aggregation; d_W, response, and dis denote final relative functional-update distance, held-out output response error, and decision disagreement rate, respectively, with dis reported as a proportion.

Task	Method	Held-out loss	d_W	Response	Dis	Trials per seed
QQP	raw	17.4711	0.04765	0.52391	0.05581	0
QQP	global-LR selected	16.5097	0.04678	0.53206	0.05887	6
QQP	LoRA-GACT	17.0329	9.59×10^{-9}	0	0	0
RTE	raw	14.4997	0.02939	0.40613	0.04556	0
RTE	global-LR selected	14.2946	0.03637	0.52426	0.05755	6
RTE	LoRA-GACT	14.4958	9.13×10^{-9}	0	0	0

The summary shows that global-LR can change held-out loss without closing the trajectory, response, and decision checks, whereas LoRA-GACT restores all three errors to numerical precision with zero search on both tasks. Appendix B therefore closes the evidence chain from reference definition to seed-level records and task-level conclusions.

C Optimizer-Specific Transport Evidence

Building on the scalar evidence in Appendix B, this section adds optimizer state, decay, and the positive-diagonal rule. It organizes the evidence as a state-transport table, a rule-level figure, and a trajectory-level figure. The table provides cell-level terminal values, while the two composite figures retain time trajectories, task responses, and uncertainty for each rule.

State transport. Table 6 tests transport with nonzero optimizer state. Raw retains the state in the original coordinates, factor-only changes the factors while retaining the original state, and LoRA-GACT synchronizes the factor update rule with the corresponding state transport. The four rows cover QQP/RTE under Momentum and AdamW.

Table 6: Nonzero optimizer-state transport. Values are 3-seed task-level means, with each row corresponding to one task–optimizer cell.

Task	Optimizer	Raw d_W	Factor-only d_W	LoRA-GACT d_W	Raw response	Factor-only response	LoRA-GACT response
QQP	Momentum	0.187897	0.087090	8.98×10^{-9}	2.513722	0.161197	0
QQP	AdamW	0.684218	0.269769	8.25×10^{-9}	0.747083	0.131964	0
RTE	Momentum	0.111279	0.111391	9.47×10^{-9}	2.971925	0.376724	0
RTE	AdamW	0.759332	0.343763	9.08×10^{-9}	0.548491	0.093466	0

Across the four task–optimizer cells, factor-only retains residual d_W and response, whereas LoRA-GACT reduces d_W to approximately 10^{-8} and response to zero. The result makes state transport a component of the continuation rule rather than an implementation detail.

Rule-level comparison. Figure 4 uses aligned axes to present five registered rules: Plain SGD, Momentum SGD, coupled-L_2 Momentum, AdamW, and the positive-diagonal extension. Each row represents one rule or extension, and each column fixes QQP/d_W, QQP/response, RTE/d_W, or RTE/response; curves and endpoints retain seed trajectories, means, and uncertainty so that method differences are directly comparable across rules.

The rule-level comparison follows one common pattern: raw, factor-only, and shared-decay conditions retain deviation, while LoRA-GACT tracks the reference; the positive-diagonal row additionally checks direction-wise update coefficients and state transport. This figure provides the cross-rule comparison.

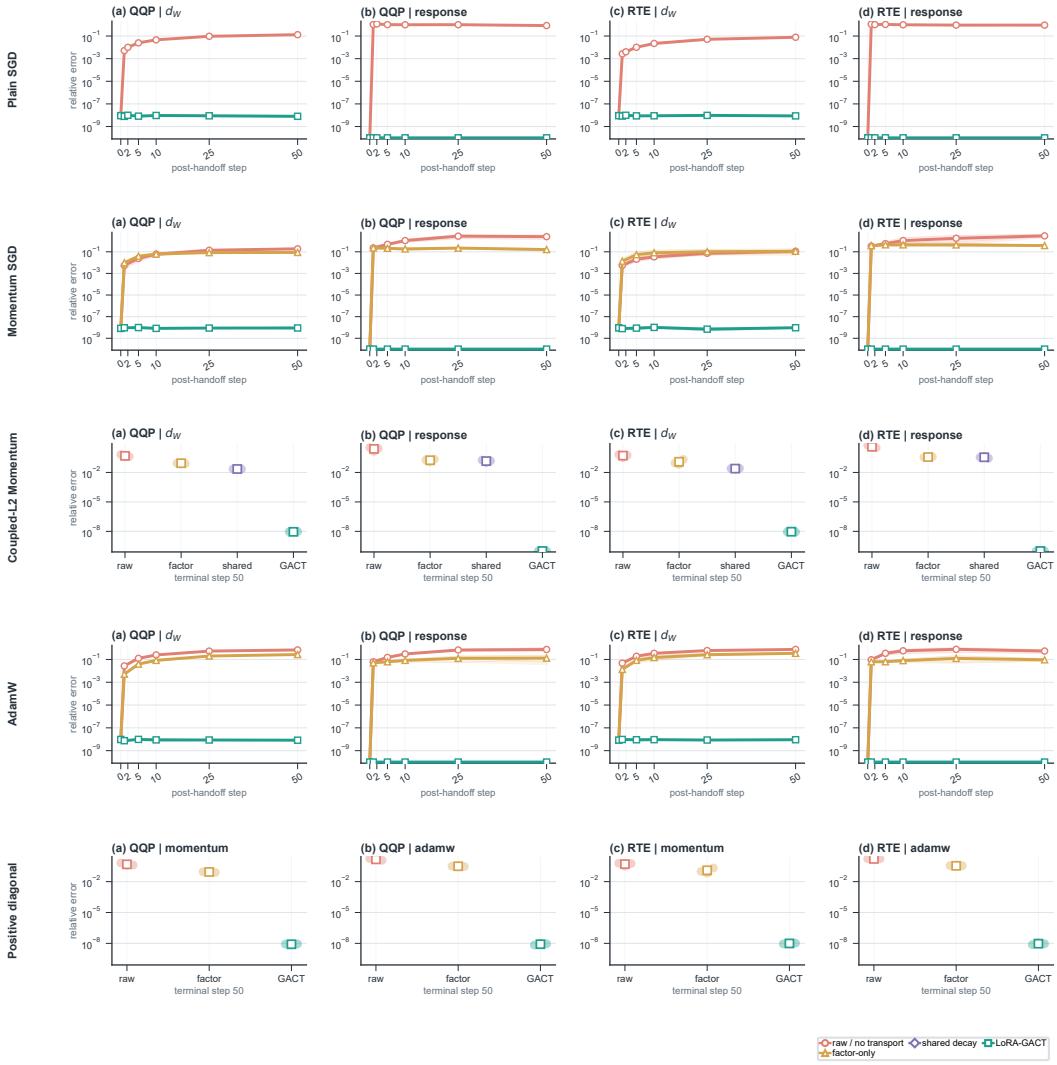

Figure 4: Optimizer-rule comparison for plain SGD, Momentum SGD, coupled-L_2 Momentum, AdamW, and the positive-diagonal extension. In each row, panels (a)–(d) report QQP/d_W, QQP/output response, RTE/d_W, and RTE/output response. Solid curves are seed means, faint curves are individual seeds, translucent bands are mean $\pm$ seed standard deviation, and endpoint-only rows show seed points with mean whiskers. Shaded restoration regions correspond to $d_W < 10^{-6}$ in distance panels and response $< 10^{-3}$ in response panels.

Trajectory-level comparison. Figure 5 places nonzero Momentum SGD, coupled-L_2, coordinate-wise AdamW, decoupled decay, and positive-diagonal conditions in one trajectory view. The four columns record relative ΔW error, held-out response, train loss, and the optimizer driver; the first four rows use complete time series, and the final row retains the terminal comparison for the positive-diagonal experiment.

This figure provides the longitudinal comparison: the first four rows show complete post-handoff trajectories under state, decay, and optimizer-driver changes, while the final row adds the terminal positive-diagonal behavior. Together with Table 6, Appendix C covers single-step rules, nonzero state, and multi-step trajectories.

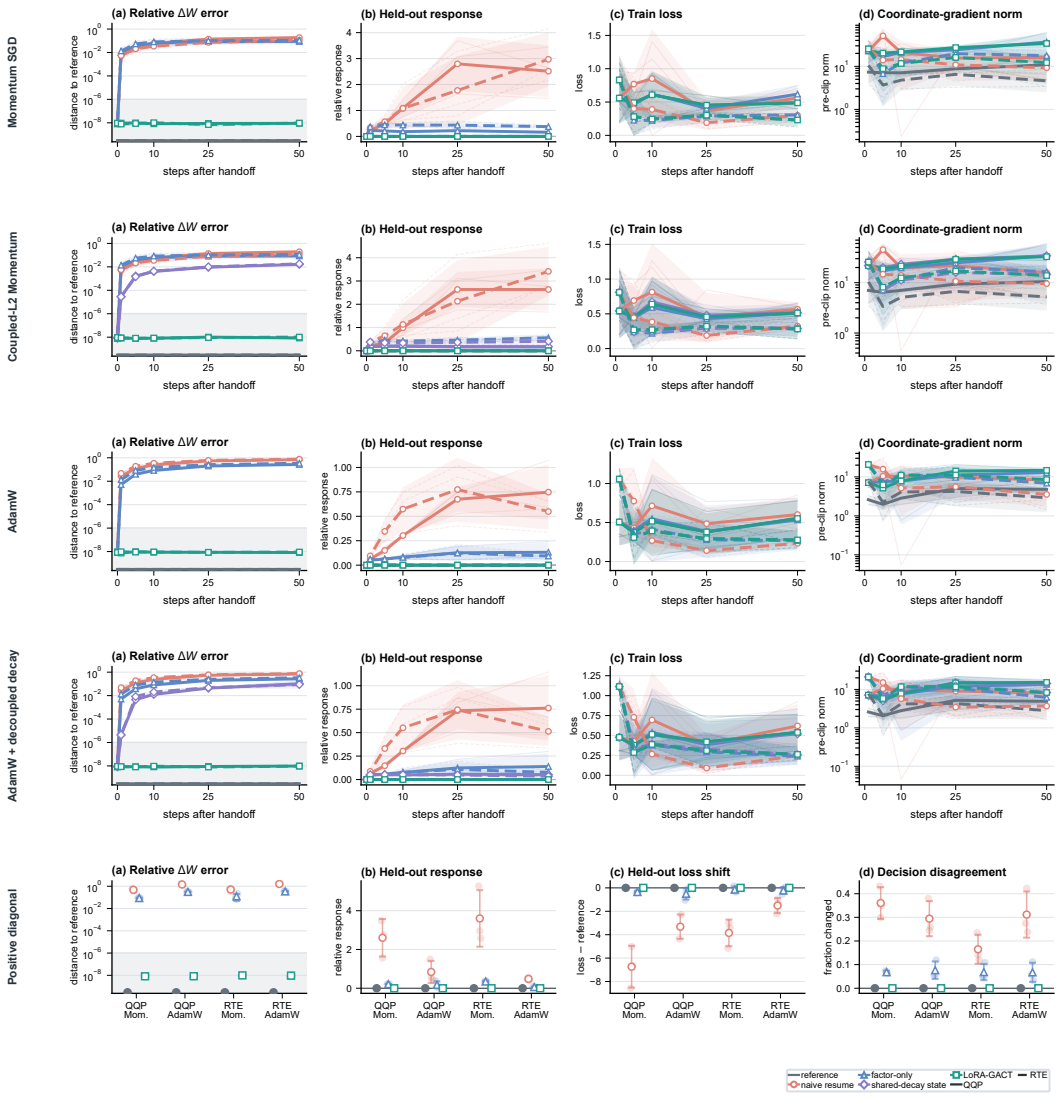

Figure 5: Composite trajectory family corresponding to Figures 13–17. Rows report nonzero Momentum SGD, coupled-L_2 Momentum SGD, coordinate-wise AdamW, AdamW with decoupled decay, and the positive-diagonal extension. Panels (a)–(d) retain the unified views of relative ΔW error, held-out response, train loss, and the optimizer driver; dense rows use solid curves for seed means, faint curves for individual seeds, and translucent bands for mean $\pm$ sample SD, while the endpoint-only row reports seed observations with mean $\pm$ SD whiskers. Shaded restoration regions mark $d_W < 10^{-6}$ where applicable.

D LoRA Factor-Coordinate Analysis: Mechanism Evidence

This section summarizes the complete protocol and key values of the LoRA factor-coordinate identifiability analysis. The analysis evaluates factor-level interpretations through identifiability, initialization robustness, functional relevance, and intervention validity, thereby locating the coordinate-transport issue in checkpoint continuation.

D.1 Analysis Object and Four Identifiability Conditions

For any positive scalar c, $A' = cA$ and $B' = B/c$ preserve $B'A' = BA$; direction-wise reciprocal rescaling with a positive diagonal matrix also preserves the product. Consequently, uncontrolled factor norms, directions, subspaces, and single-factor transplant results are properties of raw factor coordinates. The factor-coordinate analysis therefore treats factor-level semantic attribution as a testable claim subject to four conditions:

1. **Identifiability:** the conclusion remains stable under gauge transformations that preserve ΔW and model outputs;
2. **Initialization independence:** the conclusion persists across which side is zero, initial scales, and initial functional responses;
3. **Functional relevance:** the structure is reproducible on held-out tasks, fixed-input output responses, or other function-space objects and remains stable beyond raw parameters;
4. **Intervention validity:** factor-level operations have a consistent definition under equivalent parameterizations and produce reproducible behavioral consequences.

Together, these four conditions determine the evidential strength of a factor-semantic interpretation, and the analysis records raw coordinates, functional objects, and behavior trajectories separately; results satisfying the conditions enter the subsequent continuation-mechanism analysis.

D.2 Complete Analysis Protocol

The identifiability check samples multiple function-equivalent reparameterizations of trained adapters and records raw-factor changes, ΔW error, and fixed-input output error. The initialization check compares I1–I4: I1 uses $A \neq 0, B = 0$, I2 uses $A = 0, B \neq 0$, I3 uses nonzero factors with the native factor-scale imbalance, and I4 applies reciprocal gauge rescaling while matching I3 in ΔW and initial output response up to the reported numerical construction error. The functional-semantics check fits task probes on construction seeds and compares raw factor features, normalized factors, the ΔW singular spectrum, functional sketches, and fixed prompt-bank responses on held-out seeds. The intervention-validity check replaces donor A or B under raw, norm-matched, canonical, and canonical-norm conditions, while applying reciprocal rescaling to the donor that preserves its complete ΔW.

The main analysis uses RoBERTa-base with rank 8 and $\alpha = 16$ across SST-2, MRPC, QNLI, QQP, and RTE; initialization controls use MRPC and QNLI. The four conditions for each task–seed share the backbone, data subset, and batch order. I3 sets the initial $\|\Delta W\|_F / \|W_0\|_F$ of each module to 10^{-3}, while I4 uses the same random tensors with reciprocal rescaling. The main AdamW protocol uses learning rate 2×10^{-4}, weight decay 0, batch size 4, and checkpoints at steps $0, 1, 2, 5, 10, 25, 50, 100, 200, 500$; ordinary SGD and first-order gauge-compensated SGD provide optimizer controls. All I3/I4 comparisons use task–seed pairs as the statistical unit and report bootstrap 95% intervals.

D.3 Gauge Identifiability Results

The 47 real RoBERTa adapters produce 705 gauge-transformation records. The mean relative change in raw factors is approximately 2.94, while the maximum relative errors in ΔW and fixed-input outputs are 7.66×10^{-16} and 8.91×10^{-16}, respectively. The 16 BERT adapters produce 240 records, with a mean raw-factor change of 2.648 and a maximum ΔW error of 1.75×10^{-15}. Thus, one functional update corresponds to factor coordinates with substantially different numerical structures, while functional objects provide a stable anchor for cross-gauge comparison.

Table 7: I1–I4 initialization results from the factor-coordinate analysis.

Condition	Initial construction	Step 0 $\log_{10}(\|g_A\| / \|g_B\|)$	Initial response RMS	500-step accuracy (%)
I1	$A \neq 0, B = 0$	-30.880	0	63.68 ± 7.11
I2	$A = 0, B \neq 0$	30.872	0	63.75 ± 7.00
I3	Both nonzero, native scale	0.983	0.010171	64.32 ± 7.30
I4	Both nonzero, balanced gauge	-0.008	0.010171	64.36 ± 7.36

Table 8: I3/I4 optimizer-boundary control summary. The maximum absolute initial output-response difference is 2.7907×10^{-6}; all three optimizers share the step-0 I3/I4 gradient ratio. Accuracy differences are I4–I3 in percentage points, with 95% bootstrap CIs in brackets; MRPC/QNLI columns report task means, and the last column reports the mean gradient-ratio gap from step 1 to step 500.

Optimizer	Step 0 I3 / I4	Accuracy difference mean [95% CI]	MRPC Δ accuracy	QNLI Δ accuracy	Gradient-ratio gap step 1 $\rightarrow$ 500
AdamW	$+0.983/-0.008$	$+0.045\,[-1.103, +1.383]$	$+0.637$	-0.547	$0.868 \rightarrow 0.078$
Ordinary SGD	$+0.983/-0.008$	$+0.483\,[-0.613, +2.202]$	-0.441	$+1.406$	$0.446 \rightarrow 0.171$
Compensated SGD	$+0.983/-0.008$	$+0.771\,[-0.569, +2.959]$	-0.490	$+2.031$	$0.991 \rightarrow 0.914$

D.4 I1–I4 Initialization and Dynamics

I1 and I2 produce the zero-side gradient reversal: $\|g_A\|$ is zero at step 0 for I1, while $\|g_B\|$ is zero for I2. The maximum absolute initial output-response difference between I3 and I4 is 2.7907×10^{-6}, whereas the mean step-0 $\log_{10}(\|g_A\| / \|g_B\|)$ is $+0.9831$ and -0.0082, respectively. The complete summary follows.

After 500 steps, the I4–I3 accuracy results for all three optimizers are summarized in Table 8; all three 95% bootstrap intervals cross zero, and MRPC and QNLI have opposite directions. The AdamW 500-step loss difference has mean -0.000922 with interval $[-0.009111, +0.008872]$.

The finite I1/I2 log-ratio values use the numerical stabilizer 10^{-30} added to each gradient norm before taking the ratio; they are display values rather than finite mathematical logarithms of zero.

Both zero-side initialization and reciprocal gauge rescaling consistently rewrite raw gradient dynamics, while endpoint differences after functional matching vary with task and optimizer. This result shows that factor asymmetry depends on optimization conditions and representation coordinates, making update-rule transport after checkpoint handoff the direct target of LoRA-GACT.

D.5 Functional Objects and Single-Factor Intervention Results

On the RoBERTa held-out task-identification protocol, the ΔW singular spectrum reaches 0.95, the functional sketch reaches 0.90, and fixed-input output response reaches 1.00; the raw-factor probe changes across equivalent gauges. Output response remains at 0.90–1.00 under norm, margin, and prompt-bank controls, with prediction consistency of 1.00 across 60 end-to-end gauge checks. In the BERT factor-level replication, spectral features and the functional sketch reach 0.938 and 0.813, respectively, and remain unchanged across 15 gauges.

Four tasks, five held-out seeds, and 12 directed task pairs form 240 primary hybrids and 360 donor-gauge counterfactuals. After applying a canonical transformation that preserves the donor's complete adapter function, raw-A donor hybrids change task assignment in 62.8% of cases and raw-B donors in 64.4%. The mean B-minus-A cosine under the canonical-norm condition is 0.044; the seed bootstrap is positive, while the task-pair bootstrap interval is $[-0.015, 0.107]$, the cross-bootstrap interval is $[-0.033, 0.124]$, and 7 of 12 task pairs are positive. The normalized results support function-space objects as cross-task comparison anchors and bind factor-role interpretation to the representation and task conditions.

D.6 From Coordinate Analysis to LoRA-GACT

The factor-coordinate analysis establishes the mechanistic starting point for LoRA-GACT: raw factor statistics depend jointly on representation coordinates, gradient roles can be rewritten by zero

Table 9: Complete task–seed endpoint records from the factor-coordinate analysis. Each row is one complete initialization-analysis run; I1–I4 denote the four initialization conditions. Gradient norms are measured at step 0, parameter norms at step 500, and accuracy and loss from the final evaluation of the same run. The table expands two tasks, five seeds, and four conditions into 40 records; effective ranks, responses, and complete checkpoint trajectories are retained in the corresponding record.

Task	Seed	Condition	$\|g_A\|_{F,t=0}$	$\|g_B\|_{F,t=0}$	$\log_{10}(\|g_A\|_F/\|g_B\|_F)_{t=0}$	Accuracy	Loss	$\|A\|_{500}$	$\|B\|_{500}$
MRPC	101	I1	0.000	4.567	-30.660	68.38	0.5891	8.305	1.686
MRPC	101	I2	4.448	0.000	30.648	68.38	0.6035	1.649	8.228
MRPC	101	I3	1.912	0.200	0.979	68.87	0.5918	1.880	4.334
MRPC	101	I4	0.610	0.629	-0.013	69.85	0.5935	2.556	2.429
MRPC	102	I1	0.000	6.738	-30.829	68.63	0.6061	8.271	1.595
MRPC	102	I2	7.026	0.000	30.847	68.38	0.6072	1.791	8.260
MRPC	102	I3	3.057	0.296	1.014	68.87	0.6120	2.005	4.360
MRPC	102	I4	0.978	0.927	0.024	69.36	0.6034	2.556	2.395
MRPC	103	I1	0.000	6.819	-30.834	69.36	0.6027	8.309	1.691
MRPC	103	I2	5.911	0.000	30.772	68.38	0.6424	1.595	8.228
MRPC	103	I3	2.556	0.301	0.929	68.38	0.6336	1.929	4.325
MRPC	103	I4	0.816	0.943	-0.063	68.63	0.6171	2.530	2.372
MRPC	104	I1	0.000	9.307	-30.969	73.04	0.5848	8.301	1.676
MRPC	104	I2	9.501	0.000	30.978	72.06	0.5862	1.831	8.261
MRPC	104	I3	4.094	0.408	1.001	72.30	0.5752	2.176	4.427
MRPC	104	I4	1.308	1.278	0.010	74.02	0.5737	2.474	2.332
MRPC	105	I1	0.000	5.412	-30.733	69.85	0.5861	8.210	1.548
MRPC	105	I2	5.685	0.000	30.755	71.08	0.5874	1.599	8.193
MRPC	105	I3	2.419	0.237	1.009	70.59	0.5811	1.791	4.299
MRPC	105	I4	0.772	0.741	0.018	70.34	0.5859	2.323	2.201
QNLI	101	I1	0.000	5.397	-30.732	57.62	0.6733	8.286	1.715
QNLI	101	I2	4.949	0.000	30.695	52.15	0.7161	1.713	8.250
QNLI	101	I3	2.131	0.237	0.955	59.38	0.6461	1.826	4.316
QNLI	101	I4	0.680	0.742	-0.038	56.84	0.6803	2.170	2.135
QNLI	102	I1	0.000	5.646	-30.752	54.88	0.6656	8.221	1.596
QNLI	102	I2	5.439	0.000	30.736	55.47	0.6675	1.552	8.198
QNLI	102	I3	2.332	0.250	0.969	54.10	0.6761	1.636	4.250
QNLI	102	I4	0.746	0.781	-0.020	52.34	0.6777	2.097	2.114
QNLI	103	I1	0.000	24.989	-31.398	61.33	0.6479	8.243	1.671
QNLI	103	I2	23.867	0.000	31.378	60.94	0.6554	1.489	8.209
QNLI	103	I3	10.293	1.096	0.973	64.84	0.6443	1.744	4.319
QNLI	103	I4	3.287	3.433	-0.019	63.09	0.6445	2.118	2.137
QNLI	104	I1	0.000	4.430	-30.646	61.13	0.6419	8.322	1.862
QNLI	104	I2	4.269	0.000	30.630	63.48	0.6237	1.776	8.241
QNLI	104	I3	1.808	0.190	0.979	65.23	0.6308	1.994	4.363
QNLI	104	I4	0.578	0.594	-0.012	63.87	0.6282	2.383	2.337
QNLI	105	I1	0.000	17.654	-31.247	52.54	0.6753	8.186	1.507
QNLI	105	I2	18.957	0.000	31.278	57.23	0.6534	1.682	8.230
QNLI	105	I3	8.123	0.770	1.023	50.59	0.6864	1.812	4.324
QNLI	105	I4	2.594	2.409	0.032	55.27	0.6641	2.105	2.178

Table 10: Per-seed dynamics for function-matched I3/I4. The two conditions use the same task, seed, random tensors, and training budget; I4 matches the initial ΔW and output response at initialization up to the reported numerical construction error. Δd_W, Δ response, and Δ loss are defined as I4 minus I3; step 1 records the first post-update difference and step 500 records the full-training difference. The paired trajectories are the recorded I3/I4 trajectory pairs.

Task	Seed	Initial response diff.	Step 1 gradient-ratio diff.	Step 1 Δd_W	Step 1 response	Step 500 Δd_W	Step 500 response	Step 500 Δ loss
MRPC	101	1.70×10^{-7}	0.8713	-0.02376	0.01224	0.13455	0.33152	-0.01207
MRPC	102	5.70×10^{-7}	0.7858	-0.02377	0.00534	-0.04965	0.38922	-0.00874
MRPC	103	2.79×10^{-6}	0.9156	-0.02375	0.00393	0.12377	-1.04296	-0.02049
MRPC	104	4.02×10^{-7}	0.9230	-0.02377	0.01145	-0.34447	-1.60807	-0.00393
MRPC	105	6.50×10^{-7}	0.9197	-0.02389	0.00642	0.03818	0.02161	0.01082
QNLI	101	1.21×10^{-7}	0.8658	-0.02388	0.00859	-0.25806	-0.37173	0.04761
QNLI	102	2.89×10^{-7}	0.8899	-0.02384	0.00225	-0.09225	0.39075	-0.00344
QNLI	103	1.12×10^{-6}	0.8505	-0.02376	0.01288	-0.20259	0.28562	0.01269
QNLI	104	2.13×10^{-7}	0.7904	-0.02379	0.01012	-0.21936	-0.08167	-0.01085
QNLI	105	2.67×10^{-7}	0.8682	-0.02384	0.00548	-0.19291	0.62803	-0.02667

initialization and equivalent scale parameterizations, task structure is more stable in ΔW and output response, and single-factor intervention depends on an explicit gauge definition. LoRA-GACT converts this mechanism into a checkpoint-continuation problem with measurable drift and an explicit

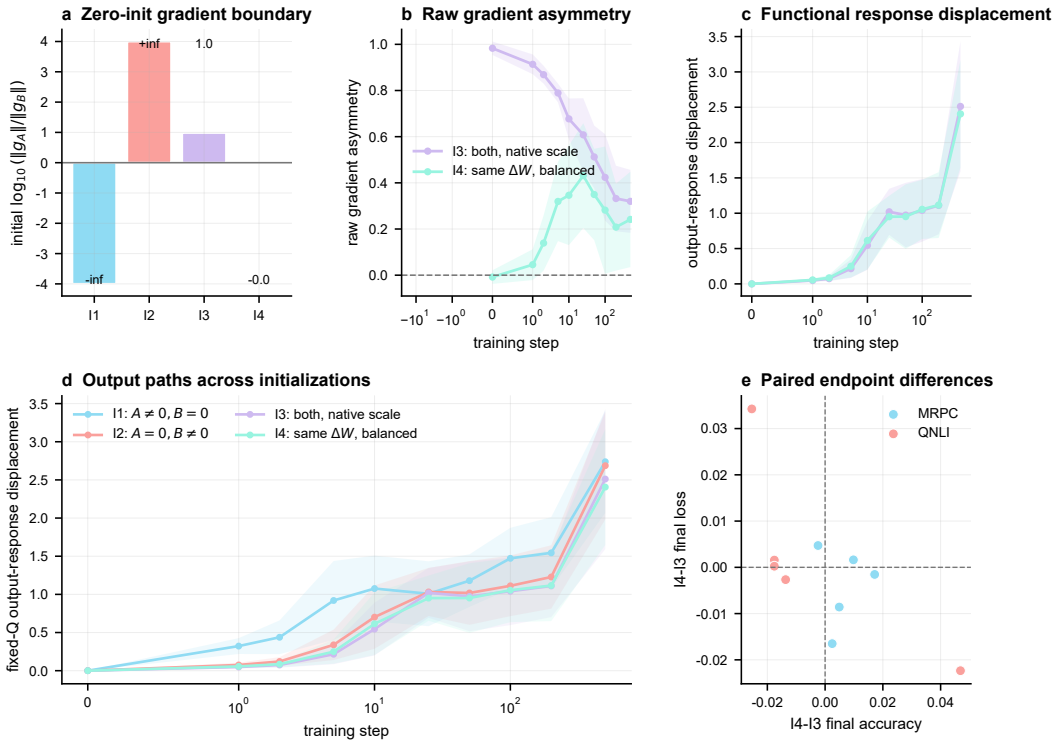

Figure 6: I1–I4 initialization audit on real models. I1/I2 show that the zero-initialized side determines the first-step gradient boundary; I3/I4 match the initial functional response while retaining different raw-gradient and response dynamics. Paired endpoints are recorded by task–seed, and functional responses together with endpoint differences separate coordinate dynamics from downstream behavior.

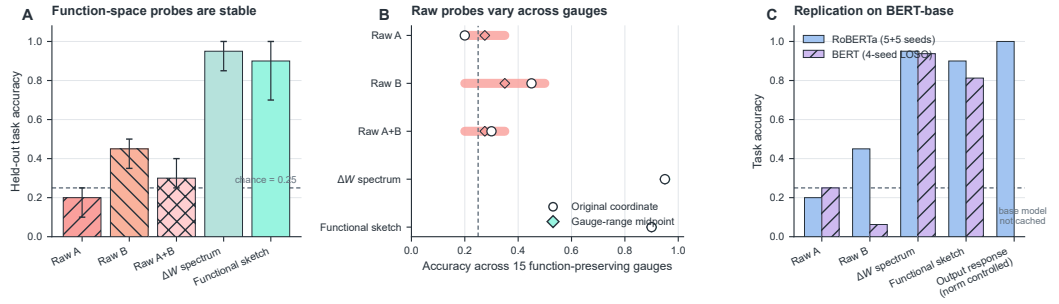

Figure 7: Stability of task structure in functional objects. The left plot compares held-out probes, the center plot shows the raw-probe range across 15 function-equivalent gauges, and the right plot reports the BERT-base replication; the ΔW singular spectrum, functional sketch, and output response remain stable under the test protocol, while the raw-factor probe changes with coordinates.

restoration protocol, executing coordinate and state transport under specified gauge, optimizer, state, and finite-precision conditions.

E Reproducibility Summary

All results are generated from machine-readable records containing the protocol version, model and data fingerprints, seed, optimizer, gauge, checkpoint state, complete record counts, and gate decisions. The experimental assets are organized into five groups: scalar and optimizer-transport

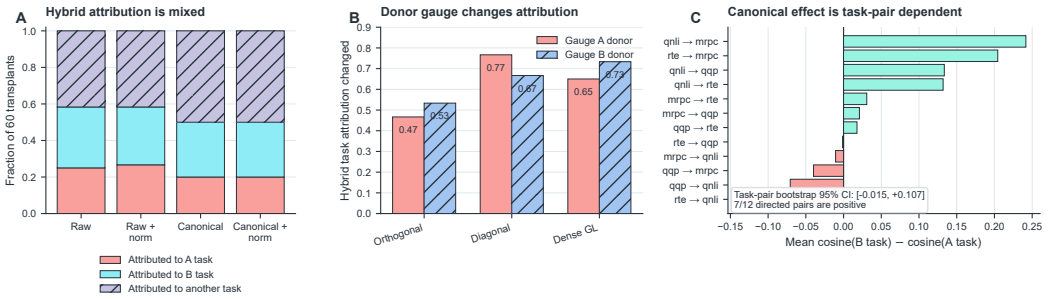

Figure 8: Donor-gauge single-factor intervention audit. Different donor gauges change raw-hybrid task assignments; task-pair effects remain directionally heterogeneous under the canonical-norm condition. Bootstrap intervals and directed-pair distributions retain the task-level variation.

validation, positive-diagonal extension, cross-task and cross-architecture generalization, recommendation ranking and checkpoint exchange, and baseline comparison. The release package contains the package-level implementation, summaries, manifests, and documentation. Full-model experiments and the complete Qwen/RoBERTa records are provided in the supplementary material rather than being duplicated in the release package. The main text reports the configurations and summaries directly tied to the argument; unified fields and execution order are given in Appendix F.

For readability, the main text uses method and experiment-object names. For example, "common-reference trajectory comparison" denotes the paired experiment between the plain-SGD target and RITE's own uninterrupted reference, while "scalar experiment" denotes the scalar-gauge continuation matrix. These names correspond one-to-one with the protocol definitions in the appendix and identify the relevant entry points through the protocol version during reproduction.

F LoRA-GACT Reproduction Protocol and Evidence Ledger

This appendix organizes the mathematical rules, experimental objects, and result records of LoRA-GACT into an executable reproduction contract. It has five layers: reproduction objects and runtime environment, protocol execution and evaluation contract, machine records and complete results, exchange interoperability and mechanism validation, and reproduction entry points and applicable protocols. Every layer returns to the same paired record and is indexed by protocol version, model and data fingerprints, seed, optimizer, gauge, checkpoint state, and gate results; tables, figures, logs, and code entry points are therefore traceable to the same underlying records.

F.1 Reproduction Objects and Runtime Environment

The main experiments use an isolated CUDA Python environment and record runtime versions, model fingerprints, data fingerprints, and random seeds. Table 11 fixes the numerical conventions; the subsequent protocol, result, and interoperability records follow the same fingerprint chain.

The versions in the environment table are jointly confirmed by the main matrix and the independent `loralib` run; every JSON result also records the runtime Python, PyTorch, CUDA, model, and data fingerprints. This recording system distinguishes reproducing the same protocol in the same environment from reproducing it on another software stack; cross-stack results are recorded as independent interoperability evidence.

Models, Modules, and Tensor Layouts. We consistently use $A \in \mathbb{R}^{r\times n}$, $B \in \mathbb{R}^{m\times r}$, and $\Delta W = sBA$. For each injected linear module, A has the latent dimension along its rows and B has the latent dimension along its columns. A scalar handoff applies only

$$A' = cA, \qquad B' = B/c,$$

and does not modify $s = \alpha/r$. The main matrix uses $r = 8$, $\alpha = 16$, and `q_proj/v_proj`, so the scalar experiments use $s = 2$. The encoder replication maps the same logic to query/value

Table 11: LoRA-GACT runtime environment and precision conventions.

Item	Configuration and handling
Runtime	Windows; experiment scripts run in offline mode
Python	Isolated CUDA Python environment
CUDA/PyTorch	CUDA 12.6; verified third-party environment: `torch 2.6.0+cu126`
Transformers/PEFT	Verified versions: `transformers 5.14.1` and `peft 0.20.0`
Independent LoRA implementation	Microsoft `loralib 0.1.2`, used as the third-party scalar-handoff validation unit
Base-model precision	Frozen base uses BF16; base parameters remain fixed during handoff
LoRA-factor precision	A, B, and transport computations use FP32; factors are moved to CPU FP32 before disk serialization
Gradients and state	Gradients are computed at model-parameter precision and converted to FP32 CPU summaries or tensors before recording and transport
Offline constraint	Runs set Hugging Face model, dataset, and tokenizer offline flags; the main matrix uses cached models and data
Randomness	Seeds are recorded for model construction, data ordering, candidate ranking, and evaluation batches; the main matrix uses 42, 43, and 44

modules and reads factor dimensions from each module's input and output dimensions rather than hard-coding a model-specific hidden size.

The models include Qwen3-1.7B, RoBERTa-base, Qwen2.5-1.5B, and Llama-3.2-1B. The extension covers different base-model scales and decoder/encoder families to test the same gauge-induced transport across architectures. Base parameters are frozen for every model; only LoRA A/B are updated, and each module retains its own scaling, parameter shapes, and module name.

For factor-parameter mapping, the experiment runner uses stable keys of the form

`module_name::A` and `module_name::B`

to map model modules to optimizer state. During save and load, the module-key set, A/B shapes, and parameter count are checked before state is bound to parameter objects in the new model. Records that pass all structural checks enter the method-metric summary.

The unified experiment pipeline uses a fixed causal-probe protocol: task examples are converted to a common yes/no verbalizer and processed with a shared probe structure. MRPC and QNLI I1–I4 comparisons use the same prompts, label mapping, and batch order. LoRA-GACT causal continuation reuses the same prompt loader, while the RoBERTa side records outputs through the encoder classification interface. Factor-coordinate metrics enter the mechanism analysis in Appendix D, and continuation metrics enter the main RQ matrix.

Data and Task Evaluation. NLP causal view. GLUE data are read from the Arrow cache or official TSV files. The loader retains text, labels, and data identifiers for each task and converts tasks into a unified causal binary view: MRPC uses paraphrase, QNLI/RTE/MNLI use entailment as positive, QQP uses duplicate question, SST-2 uses positive sentiment, CoLA uses grammatical, and STS-B uses a score of at least 3.0 as high similarity. The STS-B threshold forms a fixed functional probe, while the original GLUE regression score remains the independent task definition.

Training indices are generated by `numpy RandomState(seed).permutation` and then truncated to the specified train limit. The unified generalization matrix uses train limit 2048, evaluation limit 512, batch size 4, and fixed-input batch $Q = 32$; scalar experiments and baselines use validation limit 872, split into selection and held-out halves. The selection half serves global-LR candidate selection, and the held-out half is used for final reporting after candidate selection. Every arm uses the same indices, collator, padding rules, subsequent batch order, and evaluation inputs.

Recommendation-ranking view. For Beauty, Sports, and Toys, downstream evaluation uses one held-out positive item and 29 negative candidates per user. Candidate sets, user order, ranking batches, tokenization, and evaluation seeds are shared across all arms and seeds, and all candidates are evaluated on the same set. We record ranking loss, HR@5, HR@10, NDCG@5, NDCG@10, MRR, prediction-ranking hashes, and hash changes relative to the uninterrupted reference. The recommendation matrix contains three datasets crossed with AdamW and Momentum SGD, three seeds per cell, and six cells in total; the Beauty zero-decay supplementary control is summarized separately, while the weight-decay protocol is reported as an independent statistic.

Table 12: Reproduction scale for tasks and downstream evaluations.

Task family	Data and model	Recording rule
Main NLP matrix	Qwen3-1.7B; CoLA/MNLI/MRPC/ QNLI/QQP/RTE/ SST-2/STS-B	8 tasks × 2 optimizers × 3 seeds; complete arms and pairwise rows per cell
Encoder replication	RoBERTa-base; 4 GLUE tasks	8 cells; 3 seeds per cell
Cross-decoder	Qwen2.5-1.5B, Llama-3.2-1B; QQP/RTE	4 cells per model; 3 seeds per cell
Recommendation downstream	Qwen3-1.7B; Beauty/Sports/Toys	6 cells; one positive + 29 negatives per user
Factor-coordinate analysis	RoBERTa-base/BERT; MRPC/QNLI and task pairs	Identifiability and dynamics evidence; restoration means use the corresponding continuation matrix

F.2 Protocol Execution and Evaluation Contract

This subsection defines the execution objects, pairing relation, and evaluation contract of the main experiments. The main matrix centers on Qwen3-1.7B, injects LoRA into decoder `q_proj/v_proj`, and extends to query/value modules in RoBERTa-base and decoder modules in Qwen2.5-1.5B and Llama-3.2-1B. The NLP and recommendation-ranking matrices use the same checkpoint, data order, evaluation inputs, and continuation gate.

The scalar handoff applies $c = 4$ at step 50, uses the I3 nonzero native-scale initialization, and uses I4 as the converted arm after matching the initial ΔW and output response of I3 up to the reported numerical construction error. The positive-diagonal extension uses the specified

$$C = \mathrm{diag}(0.25, 0.5, 1, 2, 4, 8, 0.5, 4).$$

The factors, optimizer state, epsilon, decay, and learning-rate metadata are synchronized according to the corresponding transport rule. I1/I2 and the complete initialization, factor-coordinate, and gradient analyses are included in Appendix D.

The fair comparison includes raw continuation, six-candidate global-LR selection, LoRA+ (Hayou et al., 2024), LoRA-RITE (Yen et al., 2025), and LoRA-GACT. Raw retains the original global learning rate; global-LR selects candidates only on the selection split; LoRA+ uses its preregistered B-over-A learning-rate ratio; LoRA-RITE runs its invariant optimizer from the start of training and is paired with RITE's own uninterrupted reference; and LoRA-GACT uses factor update, state, decay, and numerical transport matched to the original optimizer recurrence. LoRA-GACT converted arms are derived from an existing warm checkpoint to test preservation of the original optimizer continuation, while the RITE-own-reference arm is derived from a RITE warm run. Each paired arm shares the corresponding seed, handoff point, subsequent batches, random states, and held-out evaluation inputs. Candidate records, state bytes, execution order, and schema fields are specified in the following subsections.

Execution order of a complete LoRA-GACT run. Each main task–seed–optimizer cell follows the sequence below. The uninterrupted reference and converted arms are derived from the same CPU-FP32 snapshot of a warm checkpoint, so their differences arise from the handoff and subsequent optimizer protocol.

1. Fix the model snapshot, tokenizer, data source, training indices, evaluation indices, seed, and maximum sequence length; compute and record data/model fingerprints.
2. Construct I3 with both A and B nonzero and run the I3 warm-up. The scalar experiment warms up from step 0 to 50; the factor-coordinate analysis additionally records checkpoints at 0, 1, 2, 5, 10, 25, 50, 100, 200, and 500.
3. At step 50, save A/B, ΔW, fixed-input logits, base logits, optimizer state, optimizer step, scheduler metadata, learning rate, decay, epsilon, and data-iterator metadata.
4. Establish the uninterrupted I3 continuation for the reference; reconstruct each converted arm from the same snapshot and apply the I4 gauge without changing s.
5. Before any subsequent update, check the handoff factor shapes, relative ΔW error, fixed-input output response, and logits error; start continuation after all checks pass.

6. Construct optimizer parameter groups for each arm. Raw uses the same global learning rate; factor-only applies factor learning-rate transport; full LoRA-GACT applies factor learning-rate, state, epsilon, and decay transport together.
7. Recreate the train loader for each arm with the same seed-fixed generator and subsequent batch sequence. Each global-LR trial starts independently from the same converted checkpoint and initial state.
8. At each checkpoint, save A/B, ΔW, outputs, loss, predictions, decision hash, gradient summaries, peak memory, and wall-clock time; run selection, held-out, and downstream evaluations at the endpoint.
9. Pair every converted arm with its uninterrupted reference and run the gate evaluator; the summary script reads only complete JSON records that pass schema checks.

The scalar-experiment timeline is

$$\text{I3 step 0–50} \longrightarrow \text{I4 handoff at 50} \longrightarrow \text{continuation step 50–100.}$$

Thus, "matching output at the handoff point" and "matching output at step 100" are independent checks. Continuation records retain final task scores, the handoff snapshot, and the reference pair, which together determine continuation correctness.

Transport Rules, State, and Serialization. Plain SGD and Momentum SGD. Under the scalar gauge, gradients along A rows scale by $1/c$ and gradients along B columns by c. LoRA-GACT therefore uses the following rule for plain SGD and Momentum SGD:

$$\eta'_A = c^2\eta, \qquad \eta'_B = \eta/c^2, \qquad u'_A = u_A/c, \qquad u'_B = cu_B.$$

In the scalar experiment with $c = 4$ and $\eta = 2 \times 10^{-4}$, the I4 factor learning rates are $\eta'_A = 3.2 \times 10^{-3}$ and $\eta'_B = 1.25 \times 10^{-5}$. With zero optimizer state, state transport writes zero buffers; when handoff occurs after warm-up, the momentum buffers produced during warm-up are first written to the CPU snapshot and then transformed by the rule above.

AdamW. In the coordinate-wise AdamW runner, the first- and second-moment states and epsilon for A/B use

$$\begin{aligned} m'_A &= m_A/c, \quad v'_A = v_A/c^2, \quad \eta'_A = c\eta, \quad \epsilon'_A = \epsilon/c, \\ m'_B &= cm_B, \quad v'_B = c^2 v_B, \quad \eta'_B = \eta/c, \quad \epsilon'_B = c\epsilon. \end{aligned}$$

Here m is `exp_avg` and v is `exp_avg_sq`; the step and bias-correction counters remain unchanged. The theorem in the main text proves stepwise commutation in exact arithmetic; the FP32-factor, BF16-base implementation measures finite-precision residuals with the continuation gate. Applying the SGD $c^2/1/c^2$ rule directly to AdamW serves as a preregistered control, while LoRA-GACT uses AdamW's own state and update rule.

Two weight-decay recurrences. Momentum SGD uses the coupled-L_2 form of PyTorch SGD: it first forms $g_p + \lambda_p p$ and then writes the result to the momentum buffer. Its exact transport is:

$$\lambda'_A = \lambda/c^2, \qquad \lambda'_B = c^2\lambda.$$

The complete proof is given in Appendix A. AdamW uses decoupled decay; for each factor group, transport preserves the decay product for the same parameter coordinate:

$$\eta'_A\lambda'_A = \eta\lambda, \qquad \eta'_B\lambda'_B = \eta\lambda.$$

The corresponding AdamW values are therefore $\lambda'_A = \lambda/c$ and $\lambda'_B = c\lambda$. The shared-decay control transforms only factors and state while retaining the same two decay hyperparameters, isolating decay as an independent error source. The four-cell result at $\lambda = 0.01$ and the supplementary sweep over $\{0.001, 0.005, 0.01, 0.05\}$ are recorded separately.

Positive-diagonal gauge. Let $C = \mathrm{diag}(c_1, \ldots, c_r)$. The ith row of A and ith column of B use the scalar rules with c_i and $1/c_i$, respectively:

$$\begin{aligned} \eta'_{A,i} &= c_i^2\eta, \quad \eta'_{B,i} = \eta/c_i^2, \quad u'_{A,i} = u_{A,i}/c_i, \quad u'_{B,:,i} = c_i u_{B,:,i}, \\ \eta'_{A,i} &= c_i\eta, \quad \eta'_{B,i} = \eta/c_i, \quad \epsilon'_{A,i} = \epsilon/c_i, \qquad \epsilon'_{B,i} = c_i\epsilon \end{aligned}$$

Table 13: Arm definitions in the LoRA-GACT main matrix.

Arm	Post-handoff rule	Question answered by the arm
Continuous reference	I3 original-coordinate continuation with the original optimizer and batches	Target trajectory for the handoff comparison
Converted raw	I4, shared original global LR for A/B, state retained unchanged	How much ordinary save/rescale-and-resume deviates
Converted global-LR	I4, preregistered global-LR grid per seed	Whether empirical search can replace coordinate transport
Converted factor-only	I4, A/B learning-rate transport without nonzero state/decay transport	Whether changing learning rates alone suffices
Converted full LoRA-GACT	I4, optimizer-specific transport of LR, state, epsilon, and decay	Whether the complete protocol restores the reference
LoRA+	Paper-fixed B-over-A ratio and its own reference	Whether a fixed asymmetric ratio restores the current gauge
LoRA-RITE	Invariant optimizer from the start, paired with RITE's own uninterrupted reference	Whether its own optimizer preserves gauge invariance

The first line above is used for plain SGD and Momentum SGD, and the second for AdamW. The experimental diagonal matrix is $\mathrm{diag}(0.25, 0.5, 1, 2, 4, 8, 0.5, 4)$. The diagonal-state run also transforms m/v row by row and column by column with the same first- and second-order scales; compressing C to a single geometric-mean scalar serves as a control.

State serialization format. Each LoRA parameter stores its associated state, while the base model is referenced as an external frozen component. Momentum state contains `momentum_buffer`; AdamW state contains `exp_avg`, `exp_avg_sq`, and `step`. State is converted to CPU FP32 before serialization with `torch.save`, loaded with `weights_only=True`, and checked for parameter keys, field sets, tensor count, file size, SHA-256, and maximum round-trip absolute error. The maximum recorded round-trip error is zero; this result corresponds to the serialization loop used in the experiments, while the extended seven-check scalar interoperability result is recorded as an independent validation unit.

Arm Definitions, Comparisons, and Fairness. "Better training" and "reference restoration" are treated as separate objectives, and each method is split into an independent uninterrupted reference and converted continuation. Table 13 gives the fixed semantics of the main comparison.

The main global-LR grid is $\{2\times10^{-5}, 5\times10^{-5}, 10^{-4}, 2\times10^{-4}, 5\times10^{-4}, 10^{-3}\}$. Each candidate starts from the same converted checkpoint and subsequent data order; the candidate with the lowest selection-half loss is marked selected, while trajectories and held-out results for all candidates are retained. A single global scalar supplies one scale degree of freedom, whereas the $1/c$ change in the A gradient and the c change in the B gradient require two reciprocal update coefficients; selected loss and trajectory restoration are therefore reported separately.

The main LoRA+ comparison uses a matched plain-SGD reproduction, so the difference lies primarily in the factor learning-rate law; a paper-native AdamW check is listed as a secondary result. LoRA-RITE uses its public invariant-optimizer implementation and accumulates its additional state from a RITE continuation. The main matrix reports restoration relative to each method's own reference, while the common-reference block uses the plain-SGD target: LoRA-GACT restores that continuation, whereas both arms that switch from plain SGD to RITE retain a measurable continuation gap. The current matrix also shows lower plain-SGD state cost and fewer search trials for LoRA-GACT, which can be applied at an existing checkpoint boundary.

Metrics, Statistical Units, and Restoration Gates. For each task–seed–optimizer pair, we record the final functional-update distance between the converted arm and the uninterrupted reference:

$$d_W = \frac{\left\|\Delta W^{\mathrm{cont}} - \Delta W^{\mathrm{ref}}\right\|_F}{\max(\left\|\Delta W^{\mathrm{ref}}\right\|_F, 10^{-12})},$$

as well as held-out output response, logit distance, loss gap, and decision disagreement. Output response is the RMS on fixed evaluation inputs with $Q = 32$; recommendation ranking metrics are computed on fixed users and candidate sets. Binary predictions also retain a SHA-256 decision hash, detecting cases where the mean is close but a small number of critical decisions flip.

The main statistical unit is the task–seed–optimizer paired cell. Multiple gauge versions of one adapter are pooled within the same task–seed pair, and multiple checkpoints from one seed are pooled within that seed unit. Summaries report the mean, sample standard deviation, worst seed, worst candidate, and the number of passing units; task heterogeneity is reported by task strata.

The gates are defined as follows:

1. **M0 handoff match:** relative ΔW error, fixed-input logit error, and output response at conversion are each at most 10^{-6}.
2. **M1 problem exists:** in at least one task, at least two of three seeds exceed 10^{-3} in final raw-I4 d_W or output response, or exceed 1% decision disagreement. Tasks meeting this threshold enter the continuation-problem matrix.
3. **M2 restoration:** LoRA-GACT simultaneously satisfies $d_W \leq 10^{-3}$, output response $\leq 10^{-3}$, and decision disagreement $\leq 1\%$ on the main experimental seeds; downstream cells additionally require loss/main-metric difference $\leq 10^{-3}$ and matching decision hashes when available.
4. **M3 practical comparison:** LoRA-GACT uses zero candidate trials and reports all raw, global-LR, LoRA+, and LoRA-RITE controls. The trajectory gate is required for a restoration label, and global-LR task loss is registered separately from the trajectory result.

M0, M1, M2, and M3 are joint criteria, with task loss, d_W, output gates, and decision gates registered separately. For example, the incorrect $c^2/1/c^2$ AdamW control has lower QQP loss than the reference but $d_W = 2.64114$, which exceeds the restoration threshold; the LoRA-GACT label is determined by the complete joint criteria.

Resampling and Probe Leakage Control. The factor-coordinate analysis and LoRA-GACT continuation experiments separate construction, selection, and held-out roles. For task probes, linear or fixed probes over raw/canonical/functional features are fitted on construction seeds and evaluated on held-out seeds for direction, accuracy, and functional consistency; gauge variants of the same seed belong to one evaluation group, and probe training and evaluation are separated by seed group.

Reported intervals use 10,000 bootstrap resamples. Main paired intervals resample task–seed units; task heterogeneity uses task-pair bootstrap; when both seed and task variation are present, cross-resampling and seed bootstrap are both reported. Label flips, margins, and prompt-bank results are interpreted on complete seed/task pairs, and central claims use complete paired evidence.

For baselines, LoRA+, LoRA-RITE, and LoRA-GACT are each paired with the corresponding uninterrupted reference; the LoRA-RITE reference comes from running RITE from the start, whereas the LoRA-GACT reference retains the original optimizer recurrence. Downstream differences relative to the common raw reference are marked with a separate comparison target in the table. For recommendation ranking, the user is the evaluation unit, and each user's 29 negative candidates form the same candidate set.

F.3 Machine Ledger, Evidence Index, and Complete Results

Each experiment run writes the fields below. The summary evaluator first checks schema and record completeness, then generates the summary; per-seed files, the manifest, and figure indexes preserve the trace from the original run to the paper values.

Result files use three layers: summary, per-seed JSON/CSV, and integrity manifest. The summary aggregates complete records; per-seed files retain all main-experiment arms; the manifest records expected records, actual records, pairwise rows, and gate decisions. Reported means are generated from records that pass schema and integrity checks.

Reproduction Package Assets. The release documentation defines the release-package contents and package-level entry points, and states the boundary between bundled assets and the supplementary material. The reproducibility manifest records protocol versions, record counts, and evidence mappings. Bundled results are mapped to their implementation entry points, summaries, and raw result records; full-model experiments and complete records are indexed in the supplementary material.

Table 14: Minimum field schema for the LoRA-GACT experiment ledger.

Field group	Concrete fields	Purpose
Identity	run id, protocol version, task, model, module list, seed, arm, optimizer	Distinguish protocol versions and paired units
Provenance	model identifier/fingerprint, tokenizer fingerprint, dataset identifier, split, data fingerprint	Verify model and data consistency across runs
Initialization	rank, alpha, s, I3/I4, gauge matrix/scale, factor shapes, initialization summary	Reconstruct the handoff and check BA
Optimizer	base LR, A/B LR, betas, momentum, epsilon, weight decay, decay law, scheduler	Reproduce coordinate transport
State	state field names, step, tensor count, dtype, CPU file, bytes, SHA-256, round-trip error	Verify nonzero state transport
Data	train/eval limit, batch size, evaluation Q, max length, index hash, batch-order hash	Ensure reference and converted arms use the same inputs
Function	ΔW norm, handoff error, fixed-input logits, output response, final d_W, checkpoint rows	Evaluate trajectory restoration
Task behavior	loss, accuracy/F1/MCC or HR/NDCG/MRR, decision vector, decision hash	Evaluate downstream propagation without replacing functional metrics
Resources	peak CUDA memory, wall-clock, trainable parameter count, trial count	Compare LoRA-GACT search/state cost
State machine	start/end step, completed checkpoints, gate booleans, schema status	Check training-process and result integrity

Table 15: Detailed summary of the behavior-propagation logs. First post-handoff values correspond to the $t = 50 \to 51$ update after functional matching; repeated-handoff decision disagreement is reported as a proportion and corresponds to consecutive exchanges at steps 50, 100, and 150.

Behavior metric	Raw	Factor-only	LoRA-GACT
Mean first post-handoff output response	0.180	0.157	0
Repeated-handoff decision disagreement range (proportion)	0.24–0.50	–	0

Complete Matrices, Figures, and Log Evidence. The completed matrices are read in the order of process logs, behavior propagation, rule comparisons, and generalization. Each result first identifies the recorded object and then the question checked by its figure or table; complete JSON/CSV records are indexed by the package manifest and the corresponding implementation entry point.

Process Logs and Behavioral Propagation. The repeated-handoff experiment performs three consecutive positive-diagonal rescalings at steps 50, 100, and 150 and records dense trajectories for every task–seed–optimizer cell. Figure 9 directly plots the saved checkpoint records by task–optimizer row: QQP/AdamW, QQP/Momentum SGD, RTE/AdamW, and RTE/Momentum SGD. The first two columns show d_W and output response across the full training timeline, while the last two zoom into the first handoff window (steps 45–65). An observation is recorded every five training steps, with the boundaries before and after handoff retained; LoRA+ is included, and LoRA-RITE is added to the AdamW row as its own-optimizer baseline. Each converted arm is measured against the corresponding uninterrupted reference; the LoRA-RITE curve uses RITE's own uninterrupted reference and is not an original AdamW transport arm, showing cumulative repeated-handoff stress, the first post-handoff split, and LoRA-GACT restoration together.

Behavior Logs and Record Scale. In the summary logs for the first post-handoff update, mean output response is 0.180 for raw resume, 0.157 for factor-only, and 0 for LoRA-GACT; after three consecutive handoffs, raw decision disagreement reaches 0.24–0.50 as a proportion, while LoRA-GACT remains at 0. The Qwen generalization matrix contains 16 task–optimizer cells, 192 arm records, and 864 pairwise rows; the cross-decoder matrix contains eight cells and 96 arm records; the recommendation matrix contains six cells. These counts describe the ledger-record level, while statistical inference remains at the task–seed–optimizer paired-cell level.

Common-reference trajectory comparison. The QQP/RTE common-reference experiment matrix contains two tasks, three seeds, and seven arms, with 21 complete records per task. Starting from the same plain-SGD warm checkpoint and switching to the plain-SGD target, raw-gauge mean

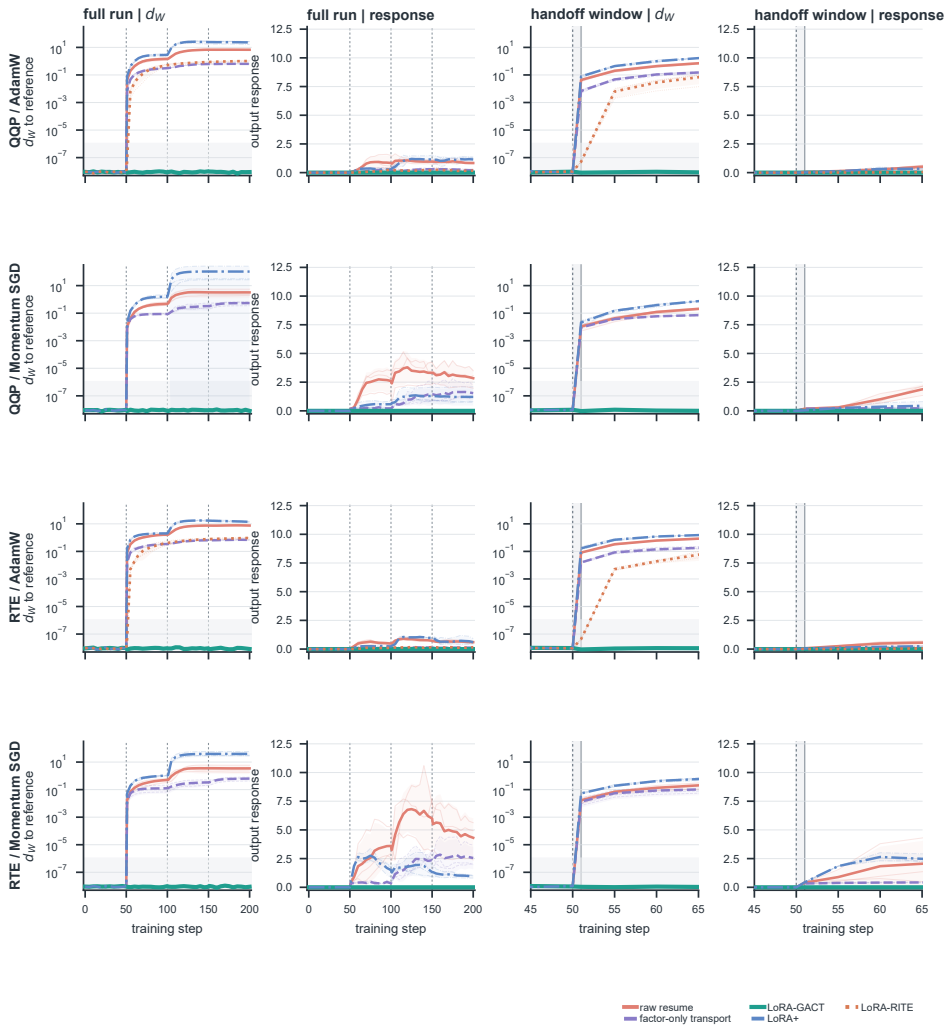

Figure 9: Dense repeated-handoff process logs (3 seeds). The four rows are QQP/AdamW, QQP/Momentum SGD, RTE/AdamW, and RTE/Momentum SGD; the first two columns show d_W and output response over the full trajectory, and the last two show local views of the first handoff window (steps 45–65). An observation is recorded every five training steps, retaining the $50/51$ boundary; colors and line styles identify methods, faint lines are individual seeds, bold lines are three-seed means, and shaded bands show sample SD together with the restoration regions $d_W < 10^{-6}$ or response $< 10^{-3}$. LoRA-RITE appears only in the AdamW rows, with its curve measured against RITE's own uninterrupted reference rather than the original AdamW continuation. Each converted arm is measured against the corresponding uninterrupted reference; $t = 50 \to 51$ is the first update after functional matching, and the full view shows accumulated deviation after later handoffs.

d_W is $0.04765/0.02939$ for QQP/RTE, while LoRA-GACT gives $9.59 \times 10^{-9}/9.13 \times 10^{-9}$. The two common-target arms that switch from this plain-SGD checkpoint to RITE give $7.4137/8.1985$ and remain separated from the original plain-SGD continuation. All six primary LoRA-GACT cells meet the restoration gate. The RITE-from-start arm, evaluated against RITE's own uninterrupted reference, gives $d_W = 9.29 \times 10^{-9}/9.54 \times 10^{-9}$, and all six cells meet the gate on its own target. The first result validates optimizer-preserving handoff for an existing plain-SGD checkpoint; the second validates own-optimizer invariance after using RITE from the start. They answer different continuation questions.

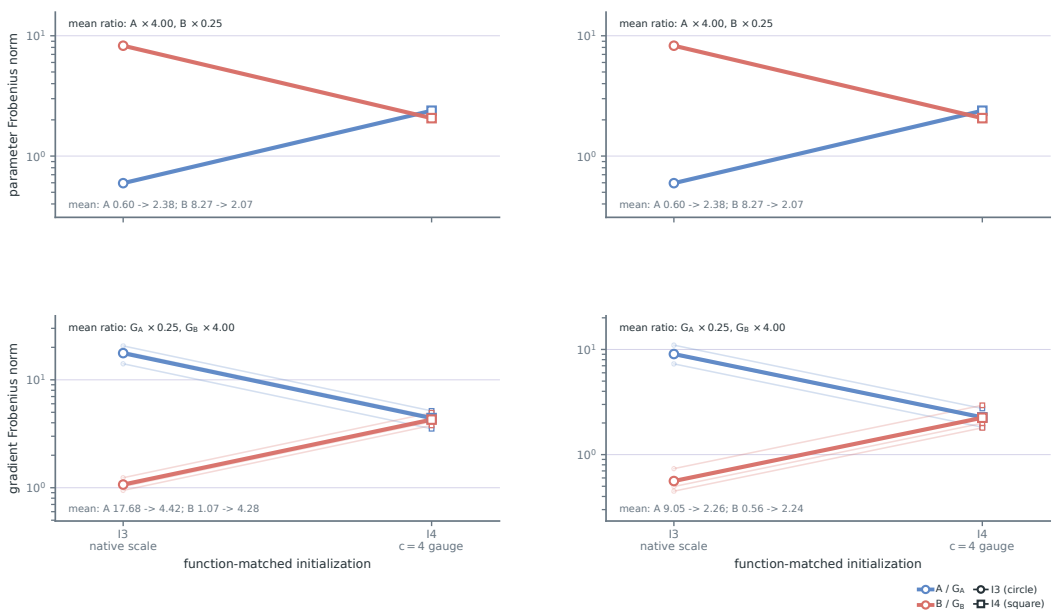

Figure 10: Parameter- and gradient-norm comparison at initialization. The top row shows checkpoint parameter norms $\|A\|_F, \|B\|_F$, and the bottom row shows factor-gradient norms $\|g_A\|_F, \|g_B\|_F$ on the same next batch; columns correspond to QQP and RTE. Thin lines connect I3/I4 paired records for each seed, bold lines are three-seed means, error bars are sample SD, circles denote I3, and squares denote I4. The protocol is plain SGD at step 0 with $c = 4$; the parameter mean ratios are $A \times 4.00, B \times 0.25$, and the gradient mean ratios are $G_A \times 0.25, G_B \times 4.00$. The figure directly checks reciprocal parameter scaling and gradient covariance.

Scalar experiment. For Qwen3-1.7B on QQP/RTE with three seeds, the $c = 4$ handoff forms six paired cells. Raw mean d_W is 0.04765/0.02939 for QQP/RTE; after selecting a winner from six candidates for each seed, global-LR selected held-out loss/d_W/response/dis is QQP $16.5097/0.04678/0.53206/0.05887$ and RTE $14.2946/0.03637/0.52426/0.05755$; LoRA-GACT gives $d_W = 9.59 \times 10^{-9}/9.13 \times 10^{-9}$ with zero response and dis. The QQP seed-42 sweep over $c \in \{0.25, 0.5, 2, 8\}$ is a single-seed scale exploration, while the multi-seed $c = 4$ matrix supports the main conclusions.

Optimizer and state. For zero-state Momentum SGD, mean raw/LoRA-GACT d_W is $0.34283/8.98 \times 10^{-9}$; the coordinate-wise AdamW result is $0.78108/8.25 \times 10^{-9}$; the c^2 control that applies the SGD law to AdamW gives 2.64114. In the four QQP/RTE × Momentum SGD/AdamW cells with nonzero state, raw and factor-only both exceed the restoration threshold, while all 12 seed–task–optimizer records for full LoRA-GACT reach approximately 10^{-8} with zero response and decision disagreement.

Weight decay. At $\lambda = 0.01$, the final d_W of the shared full-state control is 0.016402 for QQP Momentum SGD, 0.103644 for QQP AdamW, 0.018103 for RTE Momentum SGD, and 0.088777 for RTE AdamW; gauge-aware full LoRA-GACT gives 8.53×10^{-9}, 9.75×10^{-9}, 1.01×10^{-8}, and 9.56×10^{-9}, respectively. This control shows that factor and state transport are synchronized with the decay-induced rule.

Positive-diagonal state. Across the four diagonal-plus-nonzero-state cells, mean raw/factor-only/full LoRA-GACT d_W is $0.480711/0.086942/8.20 \times 10^{-9}$ for QQP Momentum SGD, $1.456020/0.310734/8.31 \times 10^{-9}$ for QQP AdamW, $0.503070/0.125753/9.84 \times 10^{-9}$ for RTE Momentum SGD, and $1.606223/0.344831/9.34 \times 10^{-9}$ for RTE AdamW. All full LoRA-GACT cells have zero response and decision disagreement. The result supports latent-direction-wise transport, with the corresponding rule transmitted for each diagonal gauge.

Table 16: LoRA-GACT scalar-scale exploration on QQP. $c \in \{0.25, 0.5, 2, 8\}$ are exploratory runs with seed 42; $c = 4$ is the three-seed main protocol.

c	raw d_W	raw response	LoRA-GACT d_W	LoRA-GACT response
0.25	0.18507	1.36608	$\sim 1.2 \times 10^{-8}$	0
0.5	0.06842	0.77075	$\sim 1.2 \times 10^{-8}$	0
2	0.03089	0.47835	$\sim 1.2 \times 10^{-8}$	0
4	0.04765	0.52391	9.59×10^{-9}	0
8	0.05741	1.36375	$\sim 1.2 \times 10^{-8}$	0

Table 17: SST-2 downstream scale stress test. Each scale uses three seeds; values are the mean $\pm$ sample standard deviation of the I4-versus-I3 accuracy-fraction shift at step 200 (multiply by 100 to obtain percentage points). LoRA-GACT has zero shift at every scale.

c	raw Δ accuracy	LoRA-GACT Δ accuracy
0.25	$+0.16399 \pm 0.14420$	0
0.5	$+0.14029 \pm 0.06657$	0
1	0.00000 ± 0.00000	0
2	-0.09824 ± 0.03348	0
4	-0.09442 ± 0.01368	0
8	-0.07720 ± 0.02640	0

Model, task, and recommendation generalization. All 16 Qwen3 cells, eight RoBERTa cells, and eight cross-model nonzero-state cells for Qwen2.5/Llama satisfy the complete gate. The cross-cell mean raw d_W for Qwen3 is 0.573286, compared with the mean of the 16 cell-level three-seed means, approximately 8.63×10^{-9}, for LoRA-GACT; the maximum full LoRA-GACT d_W in the cross-model matrix is 9.95×10^{-9} for Qwen2.5 and 1.19×10^{-8} for Llama. The Beauty, Sports, and Toys recommendation matrix contains six cells; all LoRA-GACT ranking hash changes are zero, and paired HR/NDCG/MRR differences from the reference are zero. These results show continuation deviations reaching ranking decisions, with task scores serving as downstream behavior-propagation metrics.

Detailed Main-Text Generalization and Scale Results. This subsection provides the task-, model-, and scale-level evidence compressed in the main text. It retains complete values, figures, and recording definitions, so every main-text summary can be traced through these labels to the original matrices.

Formal $c = 8$ downstream stress tests. This run uses Qwen3-1.7B, SST-2, plain SGD, three seeds, and 200 continuation steps; I3 is the reference initialization and I4 is $(8A, B/8)$. Raw-I3 has loss, accuracy, F1, and macro-F1 of 15.6824 ± 0.1414, 0.5138 ± 0.0209, 0.5816 ± 0.0293, and 0.5002 ± 0.0166, respectively. Raw-I4 gives 15.3458 ± 0.0866, 0.4365 ± 0.0076, 0.5117 ± 0.0164, and 0.4226 ± 0.0041, while LoRA-GACT I4 returns to the corresponding raw-I3 values. Raw-I4 decision disagreement is 8.26%, 11.81%, and 11.12% for the three seeds, whereas LoRA-GACT is zero for all seeds; paired d_W is $0.111970/0.109242/0.104638$ for raw-I4 and $8.18 \times 10^{-9}/1.01 \times 10^{-8}/9.52 \times 10^{-9}$ for LoRA-GACT.

The same protocol on QQP gives raw-I4 versus raw-I3 $d_W = 0.267427 \pm 0.005562$, output response 0.517443 ± 0.039419, and decision disagreement $6.5367\% \pm 0.7162\%$. The corresponding changes in loss, accuracy, F1, and macro-F1 are -0.878028 ± 0.150453, -0.014908 ± 0.027403, -0.020914 ± 0.016081, and -0.016124 ± 0.023962. LoRA-GACT has per-seed d_W of $1.028 \times 10^{-8}/8.023 \times 10^{-9}/8.458 \times 10^{-9}$, with zero output, decision, and downstream metric changes. Together with SST-2, this result shows that scalar-handoff continuation drift reaches downstream behavior and that LoRA-GACT restores the corresponding uninterrupted reference.

Beauty zero-decay SGD supplementary control. This independent protocol uses plain SGD and zero weight decay on Qwen Beauty ranking, comparing the uninterrupted reference, raw-I4, and LoRA-GACT after scalar handoff. Raw-I4 produces measurable trajectory and output deviations, while LoRA-GACT restores both to numerical precision; paired NDCG@10 records the downstream behavior propagation.

Table 18: Beauty zero-decay SGD supplementary control. d_W and output response are computed relative to the uninterrupted reference; NDCG@10 is the downstream ranking metric.

Metric	Reference	Raw-I4	LoRA-GACT
d_W	–	0.035369	8.28×10^{-9}
output response	–	1.019283	0
NDCG@10	0.173251	0.172912	0.173251

Table 19: Eight-task Qwen3-1.7B generalization. Raw d_W and raw response are three-seed means for each task–optimizer cell; the LoRA-GACT d_W columns report the three-seed maximum for each cell, and response is zero in all 16 cells. The three-seed maxima for the raw metrics are listed below for reference.

Task	AdamW			Momentum SGD		
	raw d_W	raw response	LoRA-GACT d_W	raw d_W	raw response	LoRA-GACT d_W
CoLA	0.863287	0.188624	9.09×10^{-9}	0.251083	0.442727	1.01×10^{-8}
MNLI	0.842874	0.117598	9.77×10^{-9}	0.262925	0.496834	8.36×10^{-9}
MRPC	0.832255	0.139273	1.09×10^{-8}	0.333025	0.315181	8.37×10^{-9}
QNLI	0.853402	0.125976	8.42×10^{-9}	0.307565	0.894469	8.85×10^{-9}
QQP	0.815205	0.090637	1.05×10^{-8}	0.375862	0.282301	9.79×10^{-9}
RTE	0.842166	0.074759	1.04×10^{-8}	0.266325	0.558822	9.63×10^{-9}
SST-2	0.844649	0.132273	8.76×10^{-9}	0.336466	0.661146	8.72×10^{-9}
STS-B	0.861246	0.256971	8.27×10^{-9}	0.284248	0.345008	8.78×10^{-9}

Three-seed maxima (task order: CoLA, MNLI, MRPC, QNLI, QQP, RTE, SST-2, STS-B): For AdamW, raw $(d_W, \text{response})$ are $(0.881889, 0.231515)$, $(0.850085, 0.175331)$, $(0.843768, 0.172471)$, $(0.878245, 0.172548)$, $(0.827563, 0.113117)$, $(0.852730, 0.110766)$, $(0.859974, 0.227946)$, and $(0.864757, 0.274271)$, respectively; for Momentum SGD, they are $(0.253815, 0.502870)$, $(0.271264, 0.567453)$, $(0.336903, 0.440651)$, $(0.312474, 1.143942)$, $(0.380690, 0.330355)$, $(0.282626, 0.686495)$, $(0.338279, 1.047040)$, and $(0.300056, 0.424408)$.

In the cross-model validation, Qwen2.5 raw d_W has a cell-level mean range from 0.3812 to 0.9911, factor-only has a cell-level mean range from 0.0555 to 0.3319, and the maximum LoRA-GACT d_W over all records is 9.95×10^{-9}; the corresponding Llama cell-level mean ranges are 0.2428–0.9289 and 0.0581–0.3228, with a maximum LoRA-GACT d_W of 1.19×10^{-8}. Raw and factor-only retain measurable deviations in all eight cells, while LoRA-GACT held-out response and decision disagreement are zero.

F.4 Real Toolchain Problem and Restoration

PEFT-SVD reconstruction. This experiment connects the existing Qwen/GLUE runner to PEFT 0.20.0's `add_weighted_adapter(combination_type="svd")` path. At step 25, each seed performs weighted-SVD reconstruction, disk export, and reload, after which an independent continuation arm trains to step 50. Reference, naive, and LoRA-GACT use the same subsequent batches; LoRA-GACT computes the full-matrix coordinate map from the pre- and post-SVD factors and transports the associated update rule and Momentum buffer. All three configurations use seeds 42, 43, and 44; the intermediate points in Figure 12 and Table 21 come from actual checkpoints.

Fresh-process resume with nonzero state. Table 23 summarizes QQP, RTE, and SST-2 with Momentum SGD and AdamW, three seeds, and three handoff scales, totaling 162 records. Every arm loads from disk into a new Python process with nonzero optimizer state at handoff. Raw continuation changes the trajectory and decision in every task–optimizer–scale cell; LoRA-GACT transports the momentum buffer or Adam m/v, step, epsilon, and factor-wise learning rates, restoring the reference in all 54/54 terminal records with zero search trials.

Coupled rules and optimizer implementations. Global clipping and global normalization share statistics across factors, so full transport synchronizes the corresponding rule; AdamW requires the implementation-consistent factor-wise epsilon, state, and decay handling. Table 24 compares raw, a partial/local control, and full LoRA-GACT on QQP/RTE with three seeds. Full transport restores the reference under global operations and fused or unfused AdamW.

Table 20: Independent cross-decoder model validation. Each model contains four cells from QQP/RTE crossed with AdamW and Momentum SGD; values are maxima over all records.

Model	Max raw d_W	Max raw response	Max LoRA-GACT d_W	Max LoRA-GACT response
Qwen2.5-1.5B	1.1212	3.7888	9.95×10^{-9}	0
Llama-3.2-1B	0.9703	2.0900	1.19×10^{-8}	0

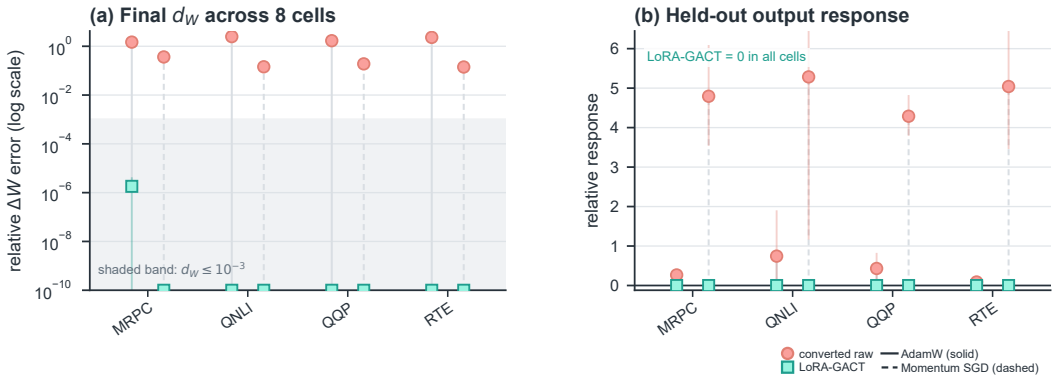

Figure 11: RoBERTa-base encoder replication. The two plots cover MRPC, QNLI, QQP, and RTE crossed with AdamW and Momentum SGD, totaling eight task–optimizer cells with three seeds per cell; circles denote converted raw, squares denote LoRA-GACT, solid/dashed lines denote AdamW and Momentum SGD, and light vertical lines show the seed min–max. The left plot reports terminal d_W on a log scale, and the right reports held-out output response; the shaded band marks the 10^{-3} restoration gate, LoRA-GACT has zero response in all cells, and exact zero is shown at a 10^{-10} display floor in the left plot.

General non-diagonal transformations. For an invertible full-matrix G, plain-SGD transport uses matrix-valued factor update coefficients so that the transformation and update satisfy the same commutation relation. The deterministic FP64 matrix experiment covers five seeds and mild, strong, and anisotropic non-diagonal G, restoring all 15/15 dense-G endpoints to numerical precision. Real PEFT-SVD reconstruction further validates the transport in nine task–seed–optimizer cells, including Momentum SGD state transport. Table 25 summarizes both levels of evidence.

F.5 Exchange, Interoperability, and Mechanism Validation

LoRA-GACT software implementation. The `lora-gact` package exposes the scalar handoff, optimizer-state transport, named-state save/load, and PEFT factor discovery in callable Python APIs. The v0.1.0 minimal smoke test runs in PyTorch 2.6.0+cu126 and covers Momentum SGD and AdamW continuation, named optimizer-state file round-trip, positive-diagonal product identity, and PEFT-style factor collection; the result is retained as a recorded smoke-test summary. The algorithmic comparison below uses global-LR, LoRA+, and LoRA-RITE as common baselines.

PEFT/Transformers path. Before handoff, the experimental pipeline copies LoRA factors from model parameters to CPU FP32 and saves them as a SafeTensors file or an equivalent CPU-FP32 factor snapshot; after loading into a new model, it checks base logits, factor tensors, ΔW, and fixed-input outputs before any optimizer update. For three QQP/RTE seeds, factor round-trip, base-logit round-trip, and handoff-output errors are all zero; after disk exchange, raw remains separated while LoRA-GACT returns to approximately 9×10^{-9}.

Microsoft loralib. The independent loralib 0.1.2 run replaces Qwen3 `q_proj/v_proj` with `loralib.Linear` and saves the library's LoRA state dict; QQP/RTE use seeds 42 and 43. Native save/load preserves continuation, while the converted raw arm has mean d_W of 0.0852/0.0500 for QQP/RTE and response of 0.5583/0.5095; scalar LoRA-GACT gives d_W of approximately $8.51 \times 10^{-9}/9.25 \times 10^{-9}$ with zero response and decision disagreement.

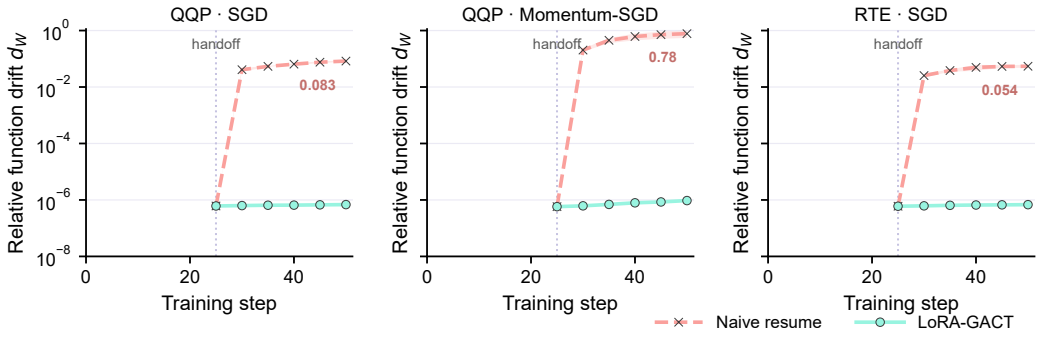

Figure 12: Function-space trajectory after a real PEFT weighted-SVD handoff. All three configurations hand off at step 25; curves show three-seed means and sample standard deviations. Naive resume departs from the uninterrupted reference, while LoRA-GACT preserves the reference continuation after transporting the full-matrix update rule and associated state.

Table 21: Stepwise relative functional-update distance after the real PEFT-SVD handoff (3-seed mean $\pm$ sample standard deviation). Step 25 is the function-matched handoff; later points are checkpoints from resumed training.

Task	Optimizer	Arm	Step 25	Step 30	Step 40	Step 50
QQP	SGD	naive	$6.12\times10^{-7}\pm2.81\times10^{-9}$	0.0411 ± 0.0060	0.0646 ± 0.0039	0.0829 ± 0.0036
QQP	SGD	LoRA-GACT	$6.12\times10^{-7}\pm2.81\times10^{-9}$	$6.32\times10^{-7}\pm3.51\times10^{-9}$	$6.55\times10^{-7}\pm6.02\times10^{-9}$	$6.84\times10^{-7}\pm1.14\times10^{-8}$
RTE	SGD	naive	$5.99\times10^{-7}\pm7.74\times10^{-9}$	0.0254 ± 0.0035	0.0495 ± 0.0120	0.0543 ± 0.0068
RTE	SGD	LoRA-GACT	$5.99\times10^{-7}\pm7.74\times10^{-9}$	$6.18\times10^{-7}\pm6.21\times10^{-9}$	$6.61\times10^{-7}\pm1.83\times10^{-8}$	$6.81\times10^{-7}\pm8.91\times10^{-9}$
QQP	Momentum SGD	naive	$5.80\times10^{-7}\pm2.49\times10^{-9}$	0.2033 ± 0.0239	0.6148 ± 0.1639	0.7768 ± 0.2284
QQP	Momentum SGD	LoRA-GACT	$5.80\times10^{-7}\pm2.49\times10^{-9}$	$6.18\times10^{-7}\pm1.25\times10^{-8}$	$7.91\times10^{-7}\pm3.57\times10^{-8}$	$9.50\times10^{-7}\pm7.81\times10^{-8}$

The extended cross-library protocol contains seven checks: export from library A, loading in library B, reconstruction of BA in library B, reconstruction of fixed-input logits in library B, optimizer-state restoration in library B, continuation on the same batches in library B, and dtype, transpose, scaling, and parameter-name mappings across A/B. The supplementary validation reports all seven checks as satisfied for the specified scalar handoff across the tested file-exchange and independent-implementation paths. This establishes scalar-path checkpoint exchange under the recorded implementations; it does not claim arbitrary library-A-to-library-B optimizer-state interoperability, non-scalar gauge support, or optimizers outside the specified protocol.

Factor-Coordinate Analysis Implementation and Supplementary Evidence. The factor-coordinate analysis forms the first stage of the LoRA-GACT argument. Under function-equivalent controls, it quantifies the coordinate dependence of raw-factor statistics, gradient roles, and single-factor interventions, thereby identifying the representation–optimization breakpoint at handoff. The continuation experiments test this breakpoint through the same data and output pipeline and measure coordinate/state transport by reference-trajectory restoration.

Gate 1: Identifiability. Sample reciprocal scalar and positive-diagonal gauges on trained adapters, requiring substantial raw-factor changes while ΔW, fixed-input outputs, and logits remain unchanged. When a raw probe changes across equivalent gauges, the reported object switches to a functional object; raw norm, direction, and effective rank are named separately as coordinate diagnostics.

Gate 2: Initialization confounding. I1 uses $A \neq 0, B = 0$, I2 uses $A = 0, B \neq 0$, I3 uses nonzero factors at native scale, and I4 uses a reciprocal gauge whose initial ΔW and output response match I3 up to the reported numerical construction error. Each task–seed uses the same random tensors, data order, and training budget; the step-0 state, the first post-update state, and the complete 500-step trajectory are recorded. The zero side reverses the gradient boundary and changes the dynamics; a fixed semantic role is tested jointly across initialization and function-matched conditions.

Table 22: Terminal function, output, and decision results after PEFT-SVD handoff. Values are 3-seed mean ± sample standard deviation; the decision-disagreement column reports disagreement with the uninterrupted reference.

Task	Optimizer	Arm	d_W	Output response	Decision disagreement
QQP	SGD	naive	0.08295 ± 0.00365	2.119 ± 1.164	0.0252 ± 0.0060
QQP	SGD	LoRA-GACT	$(6.84 \pm 0.11)\times10^{-7}$	$(4.87 \pm 0.68)\times10^{-5}$	0
RTE	SGD	naive	0.05429 ± 0.00682	0.559 ± 0.077	0
RTE	SGD	LoRA-GACT	$(6.81 \pm 0.09)\times10^{-7}$	$(6.11 \pm 2.24)\times10^{-5}$	0
QQP	Momentum SGD	naive	0.77680 ± 0.22838	4.781 ± 1.047	0.2031 ± 0.2572
QQP	Momentum SGD	LoRA-GACT	$(9.50 \pm 0.78)\times10^{-7}$	$(1.42 \pm 0.20)\times10^{-5}$	0

Table 23: End-to-end restoration with nonzero optimizer state. Each c contains 3 tasks × 2 optimizers × 3 seeds × 3 arms, totaling 54 records. The range spans 18 raw terminal records; the LoRA-GACT column reports the maximum.

c	Raw d_W range	Raw decision range	Max LoRA-GACT d_W	Restored
2	0.1863–1.0567	2.98%–23.05%	1.077×10^{-8}	18/18
4	0.9110–3.9545	12.96%–70.87%	1.077×10^{-8}	18/18
8	3.6611–9.3079	4.36%–78.33%	1.077×10^{-8}	18/18

Gate 3: Functional relevance. Fit probes on construction seeds and check raw factor features, normalized factors, the ΔW singular spectrum, functional sketches, and fixed prompt-bank outputs on held-out seeds. Report norm, margin, prompt-bank, and label-mapping results together; central claims use complete paired evidence.

Gate 4: Intervention validity. Repeat donor-A or donor-B replacement under raw, norm-matched, canonical, and canonical-norm conditions, while applying reciprocal rescaling to the donor that preserves its complete ΔW. The canonical condition tests directional consistency over multiple task pairs and held-out seeds, and donor-gauge counterfactuals directly measure the gauge identifiability of raw-factor intervention.

The factor-coordinate analysis summarizes 47 RoBERTa adapters/705 gauge records, 16 BERT adapters/240 records, I1–I4 gradient dynamics, functional sketches, and donor-gauge hybrid results in Appendix D. Gate results are stratified by reporting object: ΔW, singular spectrum, fixed output, or complete behavior trajectory; the minimum report also specifies the gauge, initialization, seed unit, and functional metric.

F.6 Reproduction Entry Points and Minimal Run Checklist

The release package is sufficient for its bundled package-level checks. To reproduce the full model matrices, follow the experiment entry points documented in the supplementary material. In either case, run the schema check before the main matrix; every rerun with a changed protocol uses a new output name and retains the previous result files.

Reproduction proceeds in four stages. (i) Verify that the bundled experiment plan and the supplementary protocol specification use the same protocol version. (ii) Run the package's checkpoint-exchange entry point to verify model, data, and CPU-FP32 factor round-trip. (iii) Launch the common-reference QQP/RTE matrix from the documented run entry point. (iv) For each arm, the documented entry points cover optimizer-state transport, positive-diagonal state transport, recommendation ranking, baselines, and cross-library interoperability; all result files are indexed by the manifest described in Appendix F.3.

Applicable Protocol. From the implementation perspective, the minimum reproducible object of LoRA-GACT is the ordered protocol below, containing the gauge, factor layout, optimizer law, state, numerical rules, and batch metadata:

$$\begin{aligned}&\text{known gauge} + \text{factor layout} + \text{optimizer update law}\\ &\quad + \text{state/epsilon/decay transport} + \text{same batches and step metadata}\\ &\qquad \Longrightarrow \text{verifiable continuation restoration.}\end{aligned}$$

Table 24: Restoration under extended update rules (3-seed mean d_W). The middle column omits the coupled statistic or uses a simplified c and $1/c$ AdamW control; full uses the registered recurrence.

Rule	Task	Implementation	Raw	Partial/control	Full LoRA-GACT	Restored
global clipping	QQP	manual SGD	0.04736	0.00798	9.48×10^{-9}	3/3
global clipping	RTE	manual SGD	0.02817	0.00442	1.01×10^{-8}	3/3
global normalization	QQP	manual SGD	0.00905	0.00350	9.44×10^{-9}	3/3
global normalization	RTE	manual SGD	0.00617	0.00105	9.48×10^{-9}	3/3
AdamW	QQP	fused	0.75134	0.21677	8.91×10^{-9}	3/3
AdamW	QQP	unfused	0.75542	0.23023	8.78×10^{-9}	3/3
AdamW	RTE	fused	0.77800	0.20023	8.78×10^{-9}	3/3
AdamW	RTE	unfused	0.77657	0.23603	9.62×10^{-9}	3/3

Table 25: Full-matrix plain-SGD transport and real SVD reconstruction. Errors are measured against the uninterrupted reference.

Evidence	Cells	Max handoff error	Max final d_W	Restored
FP64 non-diagonal G	15	1.49×10^{-15}	1.49×10^{-15}	15/15
PEFT-SVD, SGD	6	6.10×10^{-7}	6.98×10^{-7}	6/6
PEFT-SVD, Momentum SGD	3	5.81×10^{-7}	1.04×10^{-6}	3/3

Under the specified scalar and positive-diagonal gauges and model, data, precision, and optimizer conditions, LoRA-GACT restores the original optimizer's function-space continuation and measured downstream decisions through zero-search coordinate/state transport; the LoRA-RITE own-reference result evaluates continuation after using RITE from the start. Save/load and the independent loralib implementation provide supplementary checkpoint-exchange and scalar interoperability validation.