WHEN TIMING MATTERS: INTERLEAVED CROSS-BLOCK POST-TRAINING QUANTIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Block-wise post-training quantization is making low bit-width compression of large language models increasingly feasible, especially under strict memory budgets. However, PTQ still lacks the global optimization view of quantization-aware training. While most recent PTQ work focuses on block-level improvements, cross-block variants use a moving window to capture inter-block dependencies, offering a step toward global behavior. We go further and propose Interleaved Cross-Block Quantization (ICBQ), a scheduling modification that interleaves short cross-block refinements with block-wise quantization so that errors are corrected before they propagate further, offering a new route toward more global optimization. The interleaved schedule naturally creates boundary overlap, which we analyze separately as a secondary effect. In the reported experiments, ICBQ reduces ternary-quantization perplexity relative to a matched Sequential CBQ baseline, yields finite perplexity in configurations where the baseline has severe degradation, and also transfers to 3-bit and 2-bit GPTQ.

1 INTRODUCTION

Deploying large language models (LLMs) at sub-two-bit precision is an important challenge in efficient ML systems. In principle, one could directly optimize the quantized network end to end through quantization-aware training (QAT), as in LLM-QAT and LittleBit (Liu et al., 2024b; Lee et al., 2025). Such global approaches provide the best performance, as all weights are quantized at the same time in a single optimization process, usually with the the model's end loss as a guiding optimization objective. At the scale of modern LLMs, however, such global optimization is rarely feasible. Once models exceed even a few billion parameters, storing the full computation graph, intermediate activations, and optimizer state on a single accelerator typically exceeds the available memory budgets (Ichikawa et al., 2026). Consequently, practical PTQ methods instead rely on layer- or block-local surrogate objectives (Frantar et al., 2022; Li et al., 2021; Shao et al., 2024; Bulat et al., 2024). These methods aim to reduce global quantization error by decomposing the problem into a sequence of local subproblems, whose peak memory cost is primarily governed by the reconstruction unit. For large-scale LLMs, block-wise PTQ is a practical route to reducing end-to-end output error within realistic hardware constraints. Nonetheless, a considerable performance gap remains between local PTQ approaches and global methods like QAT, motivating PTQ methods that are more global while retaining the low-memory advantage of block-wise optimization.

Within practical PTQ pipelines, layer-wise weight-only methods such as GPTQ (Frantar et al., 2022) and AWQ (Lin et al., 2024) recover much of the floating-point accuracy at three- to four-bit weight precision, while SmoothQuant (Xiao et al., 2023) targets W8A8 weight–activation PTQ by smoothing activation outliers. Subsequent methods strengthen local objectives in complementary ways: QEP (Arai & Ichikawa, 2025) explicitly propagates and compensates quantization errors; LoaQ (Lin & Wan, 2025) introduces output-matching factors within layer-wise PTQ; and LPCD (Ichikawa et al., 2025a) optimizes relaxed objectives over arbitrary submodules before projecting them with layer-wise quantizers. More recently, extreme low-bit methods such as OneBit (Xu et al., 2024), BitNet (Wang et al., 2023), DBF (Boža & Macko, 2026), and MDBF (Ichikawa et al., 2025b) have pushed quantization into increasingly aggressive precision regimes. Although these methods have led to significant leaps in performance, their common denominator is that they improve quantization within a block, which limits their ability to approach truly global optimization.

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.

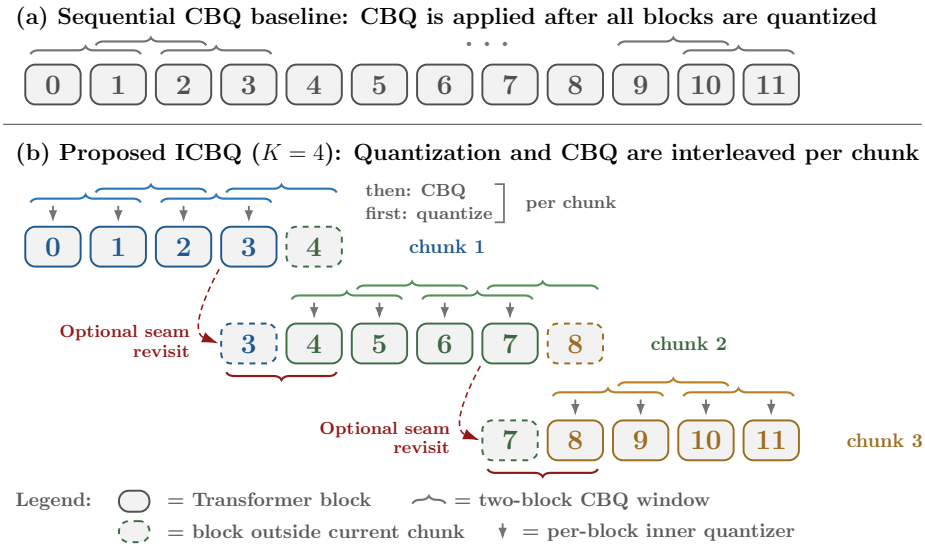


Figure 1: **ICBQ interleaves short CBQ refinements during progressive quantization** ($L = 12$, $K = 4$). **Panel a. Sequential CBQ** performs a single post-hoc sweep over adjacent block pairs. **Panel b. Interleaved CBQ** inserts a short CBQ pass after each chunk of newly quantized blocks and recomputes student and teacher activations from stored layer-0 inputs between chunks.

CBQ (Ding et al., 2025) introduced cross-block reconstruction to capture dependencies across multiple Transformer blocks. Refining two or more blocks at a time puts the focus on inter-layer behavior and is one step towards more global optimization in PTQ. Since two or more blocks are refined at once, the memory burden is higher than blockwise PTQ, while staying far below QAT's requirements, and comfortably within single-accelerator memory budgets. We adapt this principle to a fixed two-block, weight-only reconstruction objective, and refer to the resulting baseline as Sequential CBQ for brevity.

Our Sequential CBQ baseline applies each adjacent two-block window once in a left-to-right sweep after all blocks have been quantized. Residual errors introduced early in the pass can continue propagating through the remaining layers without another chance for correction, increasing final activation mismatch with depth. The effect can be pronounced under our ternary DBF setting. For example, Sequential CBQ produces very large perplexities for Llama-3-8B and Qwen3-8B in our setting. This suggests that the inter-block character of CBQ is not fully exploited by a single left-to-right sweep.

We go one step beyond CBQ and propose a scheduling enhancement which interleaves short CBQ refinements with progressive quantization, as illustrated in Figure 1. Rather than changing the local quantization mechanism itself, this modification changes when cross-block refinement is applied: we divide the network into chunks and insert a brief cross-block refinement after each chunk, so that the current prefix is corrected before more blocks are processed. Because the change is temporal rather than structural, it preserves the calibration inputs used by existing block-wise PTQ pipelines and leverages CBQ's two-block objective to its fullest while keeping peak memory essentially unchanged. A natural consequence of chunked interleaving is a boundary pair at each chunk seam. We treat revisiting that boundary as a secondary mechanism and analyze its effect separately. We evaluate the new schedule with ternary DBF and GPTQ inner quantizers.

2 Preliminaries

Notation and inner quantizers. We consider a Transformer consisting of L blocks. The b-th fp16 teacher block is denoted by $f_b^\star$, where $b \in \{0, \dots, L-1\}$. Given a calibration set $\mathcal{X} = \{x^{(n)}\}_{n=1}^{N}$, the teacher activations are defined by

$$X_0^{\star,(n)} = x^{(n)}, \quad X_{b+1}^{\star,(n)} = f_b^\star\big(X_b^{\star,(n)}\big). \tag{1}$$

The corresponding quantized student block is written as $f_b^{\theta_b}$, and the student activations are denoted by $\hat{X}_b^{(n)}$. For each sample, an activation tensor has shape (T, d_{hidden}), where T is the sequence length and d_{hidden} is the hidden width. Thus, $X_b^{\star,(n)}, \hat{X}_b^{(n)} \in \mathbb{R}^{T \times d_{\text{hidden}}}$. Unless otherwise stated, all activation norms are Frobenius norms and are denoted by $\|\cdot\|_F$. We use N exclusively for the number of calibration samples, and reserve $S = S(L, K)$ for the number of chunk seams introduced in Section 3.

An *inner quantizer* is the low-bit routine used to initialize a single Transformer block before any cross-block refinement. We denote it by $f_b^{\theta_b} = Q(f_b^\star; \mathcal{Z}_b)$, where $\mathcal{Z}_b = \{Z_b^{(n)}\}_{n=1}^N$ is the set of calibration activations supplied to block b during quantization. The role of Q is local: it receives one fp16 block and produces one low-bit student block whose weights belong to a prescribed codebook. The block-wise and cross-block procedures below do not depend on the specific choice of Q.

Representative choices for Q include Hessian-aware quantizers, such as GPTQ (Frantar et al., 2022), as well as factorized quantizers designed for extreme low-bit settings, such as DBF (Boža & Macko, 2026) and MDBF (Ichikawa et al., 2025b). The original DBF represents a weight matrix using two binary sign matrices and diagonal scaling matrices. In our experiments, we use a ternary extension in which the factor entries take values in $\{-1, 0, +1\}$; unless otherwise stated, DBF below refers to this ternary implementation. Specifically, for a weight matrix $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, our parameterization is

$$\widehat{W} = D_a A D_m B D_b, \;\; D_a = \operatorname{diag}(a), \; D_m = \operatorname{diag}(m), \; D_b = \operatorname{diag}(b). \tag{2}$$

where $A \in \{-1, 0, +1\}^{d_{\text{out}} \times k}$ and $B \in \{-1, 0, +1\}^{k \times d_{\text{in}}}$ are ternary factors, $a \in \mathbb{R}_{\geq 0}^{d_{\text{out}}}$, $m \in \mathbb{R}_{\geq 0}^{k}$, and $b \in \mathbb{R}_{\geq 0}^{d_{\text{in}}}$ are nonnegative scaling vectors, and k denotes the intermediate dimension that determines the compression ratio. We use this ternary DBF implementation as the primary inner quantizer, while GPTQ serves as a complementary baseline for evaluating whether the proposed schedule generalizes beyond a single quantization method.

Block-wise post-training quantization and prefit. Post-training quantization compresses a pre-trained model using only a small calibration set, without retraining on the original training corpus. A natural but impractical objective would be to quantize all Transformer blocks jointly so that the final output of the student model matches that of the fp16 teacher model. For modern LLMs, this global objective is usually too expensive because it requires storing activations and optimization state across many blocks. Block-wise PTQ avoids this cost by decomposing the problem into local subproblems, quantizing one Transformer block at a time (Li et al., 2021; Shao et al., 2024; Frantar et al., 2022).

In a standard block-wise PTQ pipeline, the blocks are processed in depth order. At block b, the inner quantizer Q is applied to the fp16 block $f_b^\star$ using the calibration activations available at that depth:

$$f_b^{\theta_b} \leftarrow Q\big(f_b^\star; \mathcal{Z}_b\big). \tag{3}$$

After block b has been quantized, the student activation stream is advanced by

$$\hat{X}_{b+1}^{(n)} = f_b^{\theta_b}\big(\hat{X}_b^{(n)}\big), \;\; n = 1, \ldots, N. \tag{4}$$

Thus, later blocks are calibrated using activations that already include the effects of earlier quantized blocks. This is the key distinction between block-wise PTQ and purely layer-isolated quantization: the method still optimizes local blocks, but it propagates the student activations through the quantized prefix.

Before applying the inner quantizer, one may optionally perform a *prefit* step. Prefit is a high-precision reconstruction step that adjusts a floating-point copy of the block so that it better matches the fp16 teacher block on the calibration inputs. Given the calibration inputs $\mathcal{Z}_b$ for block b, the prefit solves

$$\phi_b^{\text{pre}} = \arg\min_{\phi_b} \frac{1}{N} \sum_{n=1}^{N} \Big\| f_b^{\phi_b}\big(Z_b^{(n)}\big) - f_b^\star\big(Z_b^{(n)}\big) \Big\|_F^2. \tag{5}$$

The resulting high-precision block $f_b^{\phi_b^{\text{pre}}}$ is then passed to the inner quantizer Q. Importantly, prefit itself does not specify a low-bit codebook and is therefore independent of the choice of Q; it only prepares a better floating-point starting point for the subsequent quantization step. In this work, we enable prefit only when the inner quantizer is ternary DBF.

Cross-block quantization. The limitation of the block-wise procedure above is that each block is judged mainly by how well it reconstructs its own fp16 output. However, in a Transformer, the output error of block i becomes the input error of block $i{+}1$. A locally small error at the output of one block may therefore be amplified or rotated by the next block, especially at very low precision. CBQ (Ding et al., 2025) addresses this issue through cross-block reconstruction over multiple blocks. To isolate the scheduling effect studied here, we use a simplified weight-only adaptation with a fixed window of two adjacent Transformer blocks, which keeps the memory footprint local.

For the adjacent window $(i, i{+}1)$, our two-block CBQ-style objective minimizes the reconstruction error at the output of the second block:

$$\min_{\theta_i,\theta_{i+1}} \mathcal{L}_{\mathrm{cbq}}(\theta_i,\theta_{i+1}) = \frac{1}{N}\sum_{n=1}^{N} \left\| f_{i+1}^{\theta_{i+1}}\big(f_i^{\theta_i}(\hat{X}_i^{(n)})\big) - Y_{i+2}^{(n),\star} \right\|_F^2, \tag{6}$$

where the fp16 two-block target is

$$Y_{i+2}^{(n),\star} = f_{i+1}^{\star}\big(f_i^{\star}(X_i^{\star,(n)})\big). \tag{7}$$

Here, $\hat{X}_i^{(n)}$ is the current student-side activation entering the window, while $X_i^{\star,(n)}$ is the corresponding teacher activation. The objective therefore asks the current student window to reproduce the fp16 teacher output after two consecutive blocks. In practice, we optimize Eq. (6) with a rollback safeguard: if the updated window does not reduce the reconstruction MSE, the parameters are restored to their previous values. The Sequential CBQ baseline used in this paper first initializes all blocks with the inner quantizer Q, and then performs a single left-to-right sweep over the adjacent windows $(0,1),(1,2),\ldots,(L{-}2,L{-}1)$, solving Eq. (6) once for each window and advancing the student activation stream after every step. Thus, throughout this paper, Sequential CBQ denotes this fixed two-block adaptation: Q provides the initial low-bit blocks, and the CBQ-style objective refines neighboring blocks independently of the choice of inner quantizer.

3 Interleaved Cross-Block Quantization

3.1 Why interleaving? The limits of one-pass CBQ

Cross-block variants use a moving window to capture inter-block dependencies, but prior approaches still apply that window in a single left-to-right sweep, so early errors can propagate before correction. A simple way to view this is through the block-wise recurrence

$$\hat{X}_{b+1}^{(n)} = f_b^{\theta_b}\big(\hat{X}_b^{(n)}\big), \tag{8}$$

which shows that the student prefix at depth $b+1$ is already determined by all earlier quantized blocks. In a one-pass schedule, once a block has been processed, its mismatch is carried forward through the remaining layers unless a later pass corrects it. CBQ is therefore already more global than block-isolated PTQ, but a single sweep still conditions later refinements on a prefix that has already accumulated quantization drift, which limits how fully the cross-block objective can act as a global correction. ICBQ addresses this by partitioning the depth axis into consecutive chunks of K blocks and inserting one short interleaved CBQ sweep after each chunk. After each chunk-level refinement, activation re-roll recomputes student and teacher streams from the saved layer-0 inputs so the next quantization/refinement step sees a coherent state. A natural consequence of chunked interleaving is a boundary pair at each chunk seam. We treat revisiting that boundary as a secondary mechanism and analyze its effect separately.

3.2 Outer driver and interleaving schedule

The outer driver is the orchestrator of ICBQ: it decides when each standard block-wise PTQ operation runs and in what order. It sweeps blocks left-to-right ($b = 0,\ldots,L{-}1$), keeps a chunk-start pointer b_{cs}, and for each block applies the usual local quantization step (with optional prefit), then advances the student and teacher streams by one layer through f_b^{Q} and $f_b^{\star}$. This follows the standard block-wise PTQ progression from Section 2, but adds explicit chunk-level control over when interleaved cross-block refinement is triggered.

The chunk closes whenever $b+1-b_{\text{cs}} \geq K$ or $b = L-1$, i.e., once the current chunk reaches size K (or the model ends). At closure, the orchestrator launches a short interleaved CBQ call (Algorithm 2 in Appendix A) on the window range

$$[w_{\min}, w_{\max}) = \big[\max(0, b_{\text{cs}} - 1),\ \min(b{+}1, L{-}1)\big), \tag{9}$$

solving (6) on each pair $(i, i{+}1)$ with $i = w_{\min}, \ldots, w_{\max}{-}1$ (the upper limit avoids the non-existent pair $(L{-}1, L)$). After this local refinement, the driver re-rolls activations over the processed prefix and advances to the next chunk; the exact inner-loop and re-roll mechanics are detailed in Section 3.3. It then sets $b_{\text{cs}} \leftarrow b + 1$ and continues to the next chunk. In short, the outer driver preserves block-by-block PTQ while interleaving per-chunk CBQ: quantize progressively, refine at chunk boundaries, re-roll, repeat.

3.3 Activation re-roll for coherent interleaved refinement

In ICBQ, the pair updates themselves follow the same regular CBQ subproblem from Section 2. However, they are interleaved after each chunk during ongoing block-wise quantization. Because interleaved refinement updates earlier blocks while later chunks are still pending, cached student and teacher activations can become stale. Re-roll restores coherence by recomputing both streams from saved layer-0 inputs: locally to the window start during the within-chunk sweep, and through the processed prefix $0, \ldots, b$ at chunk end before the next outer step. This mechanism is encoded by Assumption B.4 and used by Theorem 4.1.

Concretely, given the window range $[w_{\min}, w_{\max})$ from the outer driver, the inner call performs a local block-wise pair sweep. It first recomputes the student and teacher streams to the window start (through blocks $0, \ldots, w_{\min}{-}1$), then iterates over pair indices $i = w_{\min}, \ldots, w_{\max}{-}1$: it recomputes the teacher target $Y^{\star}_{i+2}$ from $f^{\star}_i, f^{\star}_{i+1}$, updates (θ_i, θ_{i+1}) by minimizing the two-block loss (6), and advances the student/teacher streams one layer via $f_i^{\theta_i}$ and $f_i^{\star}$. Pseudocode is given in Appendix A.

3.4 Boundary overlap as a secondary effect

In ICBQ, chunked interleaving naturally creates boundary interfaces between consecutive chunks. Since these boundaries are where quantization error is handed from one chunk to the next, we distinguish the minimal interleaved schedule from an optional seam-revisit variant. Writing $b_{\text{cs}} = (c{-}1)K$ for chunks indexed by $c = 1, 2, \ldots$ (with the final chunk truncated at the last valid pair index), the window range (9) becomes

$$\Big[\max(0, (c{-}1)K{-}1),\ \min(cK, L{-}1)\Big). \tag{10}$$

In the minimal interleaved schedule, the inner sweep stops one pair earlier, while the seam-revisit variant extends the sweep by one trailing pair so the boundary pair is revisited across adjacent chunks. We analyze whether that extra revisit gives any benefit beyond the minimal interleaved schedule in Section 4 and in the experiments.

Peak-memory footprint. ICBQ retains a frozen depth-0 calibration snapshot throughout progressive quantization. Each such bank has size

$$\Delta\text{mem}_{\text{ICBQ}} = N \cdot T \cdot d_{\text{hidden}} \cdot \text{bytes}(\text{dtype}), \tag{11}$$

which depends only on the calibration workload, not on the model depth L or the chunk size K. Since each chunk's CBQ call finishes before the post-chunk re-roll, ICBQ never holds two active CBQ windows concurrently. Thus,

$$\text{Peak}_{\text{ICBQ}} \leq \text{Peak}_{\text{CBQ}} + \Delta\text{mem}_{\text{ICBQ}}, \tag{12}$$

where Peak_{CBQ} is the peak of the standalone CBQ phase and $\Delta\text{mem}_{\text{ICBQ}}$ is an L, K-independent additive constant. At $T = 2048$ in fp16, one bank is $\approx 4.3\,\text{GB}$ with $N = 256$ and $d_{\text{hidden}} = 4096$, so the extra footprint remains well below the ternary DBQ peak reported in Appendix C.3. Thus, ICBQ introduces no depth- or chunk-dependent memory growth.

4 THEORY: SEAM-AMORTIZED ERROR CONTRACTION

The interleaving schedule naturally creates boundary seams. This section analyzes the effect of adding a seam revisit at those boundaries, using Sequential CBQ as the bound-level reference; the no-revisit interleaved schedule is evaluated later as an empirical ablation. The interpretation is intentionally worst-case: it compares upper bounds, not guaranteed realized perplexity on every run. Full assumptions and proofs are in Appendix B. The tracked quantity is the boundary mismatch passed from one block pair to the next: $m_i \coloneqq \|\hat{X}_{i+1} - X^\star_{i+1}\|_F$. The reader can interpret this as "how much activation error is handed from pair $(i{-}1, i)$ to pair $(i, i{+}1)$." We denote by m_i^{out} the mismatch after the CBQ call on $(i, i{+}1)$, with $m_{-1}^{\text{out}} = 0$ from the shared start $\hat{X}_0 = X^\star_0$.

Base behavior away from seams. Under Assumptions B.1–B.4 (Appendix B), each regular two-block CBQ step has a damp-and-inject form: it contracts part of incoming mismatch (through $\tau \coloneqq \gamma\rho < 1$) and adds local approximation noise (ϵ_{sub}). Formally, Lemma B.9 (Appendix B.4) gives

$$m_i^{\text{out}} \le \tau m_{i-1}^{\text{out}} + (1{+}\gamma)\epsilon_{\text{sub}}, \;\; m_{-1}^{\text{out}} \coloneqq 0, \tag{13}$$

for every non-seam pair in either schedule.

Why seams are special. At a seam $(cK{-}1, cK)$, re-rolling resets the local context from stored inputs, and that seam pair is effectively revisited across adjacent chunks. Intuitively, non-seam pairs get one cleanup pass, while seams get an extra cleanup pass on carried mismatch. This yields

$$m_i^{\text{out}} \le \gamma\tau m_{i-1}^{\text{out}} + (1{+}\gamma{+}\gamma^2)\epsilon_{\text{sub}} \;\text{ at a seam} \tag{14}$$

(Lemma B.11, Appendix B.6). Chaining Eq. (13) with one seam recurrence per seam across the depth gives the theorem below: each seam contributes one additional multiplicative contraction factor in the bound.

Theorem 4.1 (Seam-amortized error contraction; informal, formal: Theorem B.5 in Appendix B.2, proved in Appendix B)**.** *Let e_L^{seq} and e_L^{int} denote the depth-L activation mismatches of Sequential CBQ and one-pass ICBQ with chunk size K, on the same calibration set, the same inner quantizer Q, and the same CBQ optimizer budget. Write $\mathcal{B}_L^{\text{seq}}$ for the depth-L upper bound obtained by unrolling Eq.* (13) *alone (Appendix B.8). Under Assumptions B.1–B.4 and $\tau < 1$, for every $L \ge 2$ and $1 \le K \le L$,*

$$\mathbb{E}\big\|e_L^{\text{int}}\big\|_F \le \gamma^{S(L,K)}\mathcal{B}_L^{\text{seq}} + C(L, K, \rho, \gamma)\epsilon_{\text{sub}}, \;\; S(L, K) = \lceil L/K \rceil - 1, \tag{15}$$

where $C(L, K, \rho, \gamma) \le 1 + (1{+}\gamma{+}\gamma^2)\rho/(1 - \tau)$ uniformly in L, K*, and the per-seam factor γ is tight at the recurrence level (Proposition B.20).*

Corollary 4.2 (Geometric decay in the number of seams; informal, formal: Corollary B.8 in Appendix B.2)**.** *Define the* amortization gain $G(L, K) \coloneqq \mathcal{B}_L^{\text{seq}}/\mathcal{B}_L^{\text{int}}$*, where $\mathcal{B}_L^{\text{int}}$ is the right-hand side of Eq.* (15)*. If for some $\delta \ge 0$ the residual term satisfies $C\epsilon_{\text{sub}} \le \delta\gamma^{S(L,K)}\mathcal{B}_L^{\text{seq}}$, then $G(L, K) \ge \gamma^{-S(L,K)}/(1{+}\delta)$. Thus, when additive noise is not dominant, per-seam shrink factors compound geometrically: the ICBQ-over-sequential bound ratio is exponential in the number of seams.*

Together, Theorem 4.1 and Corollary 4.2 show that seams add an extra contraction factor to the bound, so smaller K means more seams and more opportunities for bound-level shrinkage using the same calibration data. Equation (15) remains a bound-vs-bound comparison against the sequential *upper bound* $\mathcal{B}_L^{\text{seq}}$, not necessarily the realized sequential error (Remark B.7). In Section 5, we compare the seam-revisit variant against the no-revisit interleaved schedule to isolate its empirical effect.

5 EXPERIMENTS

Our experiments assess whether ICBQ improves perplexity relative to Sequential CBQ and whether the schedule transfers from one weight-only inner quantizer to another. We then isolate seam-revisit and chunk-size effects on the same model set (Section 5.4, Tables 3–4). **Formatting note.** In all tables of this section, **bold** marks the strongest value in each comparison group.

Table 1: **Ternary DBF results on seven base models.** ICBQ has lower PPL than Sequential CBQ in every reported PPL comparison and higher or equal average zero-shot accuracy in six of seven cases. FP model values are included for reference.

	FP model			No-Ref. (ternary DBF only)			Seq. CBQ			**ICBQ (ours)**		
Metric	PPL (Wiki-2)	PPL (C4)	Avg. ZS	PPL (Wiki-2)	PPL (C4)	Avg. ZS	PPL (Wiki-2)	PPL (C4)	Avg. ZS	PPL (Wiki-2)	PPL (C4)	Avg. ZS
Standard models												
Mistral-7B	4.69	7.99	0.7128	276.31	143.22	0.3684	36.5	31.1	0.4586	**13.8**	**16.8**	**0.5404**
Llama-2-7B	4.86	7.27	0.6668	26.6	17.71	0.4856	10.7	13.6	0.5330	**9.37**	**12.5**	**0.5533**
Llama-2-13B	4.35	6.73	0.6924	9.92	11.89	0.5960	7.94	10.48	**0.6503**	**7.36**	**10.09**	0.6501
Deep and non-standard-attention models												
Llama-3.2-3B	9.62	15.63	0.6320	66.94	61.44	0.4482	42.6	40.63	0.4706	**37.6**	**37.01**	**0.4857**
Llama-3-8B	7.37	12.24	0.6977	154.43	73.82	0.4715	42.95	48.98	0.5026	**34.97**	**42.32**	**0.5215**
Qwen3-8B	8.58	13.93	0.6946	$>$ 1e4	$>$ 1e4	0.3495	$>$ 1e4	5752.9	0.4032	**25.86**	**29.9**	**0.5513**
Qwen3-14B	7.58	12.48	0.7292	907.71	484.59	0.3529	61.9	44.82	**0.6871**	**19.29**	**24.67**	**0.6871**

5.1 SETUP

Experimental setup. We evaluate seven publicly released base models spanning 3B to 14B parameters. Every experiment is *weight-only*; activations remain in fp16, and no rotation or activation quantizer is applied. We evaluate ternary DBF (1.58-bit) with prefit enabled, and GPTQ at W3 (three bits, per channel) and W2g128 (two bits, group size 128) with prefit disabled. Unless otherwise stated, the chunk size is $K = 4$ and the seam revisit option is activated, while the no-revisit variant is reported separately in the seam ablation. We report C4-calibrated results in the main text, as C4 calibration was more stable than WikiText-2 calibration for some non-standard-attention architectures in our setting (Appendix C.2). Our direct benchmark is the matched Sequential CBQ baseline defined above, based on the fixed two-block adaptation of CBQ (Ding et al., 2025); we additionally report No-Refinement (inner quantizer only). During CBQ/ICBQ refinement, ternary DBF runs update the scaling vectors and ternary factors (the factors through STE), while GPTQ runs update scales/zero-points and integer weights (integer weights through STE). Additional setup details are listed in Appendix C.1.

Metrics. We report perplexity (PPL, ↓) on C4 test splits, using a standard autoregressive language-model evaluation pipeline. We also report average zero-shot accuracy (↑) over 7 standard benchmarks. Benchmark details and full per-task zero-shot values are provided in Appendix C.1 and Appendix D.

5.2 ICBQ IMPROVES PERPLEXITY OVER SEQUENTIAL CBQ

Table 1 shows lower PPL for ICBQ than for Sequential CBQ in every reported comparison across standard and deep/non-standard models. Qwen3-8B illustrates the difference: ICBQ yields finite PPL where Sequential CBQ has divergence or very large PPL. The C4 average zero-shot metric follows the same overall pattern; the exception is Llama-2-13B, where Sequential CBQ exceeds ICBQ by 0.0002.

5.3 SCHEDULE TRANSFER TO GPTQ W3 AND W2G128

Table 2 reports GPTQ-only, Sequential CBQ, and ICBQ ($T_{\text{pre}} = 0$, full-model quant) on the seven-model set. ICBQ lowers C4 PPL versus Sequential CBQ on every W3 and W2g128 cell in the table. At W2g128, ICBQ attains the lowest C4 PPL on five of seven models; at W3, GPTQ-only is best on several 7B–14B models, so the interleaved schedule can win over sequential CBQ even when blockwise CBQ does not improve on GPTQ-only. Appendix D, Table 7, shows a similar average trend with task-level exceptions. This suggests that the schedule effect and CBQ-style refinement in general are most reliable in the lower-bit regime, while W3 remains more mixed.

Table 2: **The contraction transfers to GPTQ on the seven-model set.** PPL on C4 (↓), weight-only GPTQ inner quantizer, fp16 activations, no progressive block prefit ($T_{\text{pre}} = 0$), and ICBQ with seam revisit. Within each bit-width block, lower is better.

	W3 (per-channel)			W2g128 (group size 128)		
Model	GPTQ only	Seq. CBQ	ICBQ (ours)	GPTQ only	Seq. CBQ	ICBQ (ours)
Mistral-7B	**10.26**	10.37	10.32	18.08	14.67	**14.20**
Llama-2-7B	9.91	9.80	**9.75**	17.90	13.76	**13.54**
Llama-2-13B	**8.09**	8.19	8.18	11.80	10.92	**10.82**
Llama-3.2-3B	31.56	23.55	**23.14**	108.11	45.52	**41.56**
Llama-3-8B	38.60	24.04	**23.52**	111.39	44.68	**38.68**
Qwen3-8B	**18.38**	21.32	21.09	35.18	29.22	**28.29**
Qwen3-14B	**15.12**	17.54	17.51	**20.76**	22.53	22.21

Table 3: **Effect of seam revisiting on DBF interleaved CBQ** (seven-model set, ternary DBF, C4 calibration). *Seam revisit* is the standard ICBQ schedule (chunk-boundary pairs optimized twice); in the *no revisit* setting each adjacent pair is visited exactly once.

	C4 PPL (↓)		Avg. zero-shot (↑)	
Base model	Seam revisit	No revisit	Seam revisit	No revisit
Mistral-7B	**16.8**	16.95	0.5404	**0.5442**
Llama-2-7B	12.5	**12.44**	0.5533	**0.5540**
Llama-2-13B	10.09	**10.05**	**0.6501**	0.6218
Llama-3.2-3B	37.01	**36.75**	**0.4857**	0.4853
Llama-3-8B	**42.32**	43.10	0.5215	**0.5217**
Qwen3-8B	**29.90**	30.02	**0.5513**	0.5434
Qwen3-14B	24.67	**24.51**	**0.6871**	0.6087

5.4 Schedule ablations

Returning to ternary DBF, we isolate two schedule-level factors on the same seven-model set used throughout the main experiments: seam revisit and chunk size K. We also study the effect of prefit in Appendix D.

Effect of seam revisit. We ablate the optional seam-revisit variant against the minimal interleaved schedule on the seven-model ternary DBF setting. The effect is small and model-dependent: seam revisit slightly improves PPL on some models, but no-revisit is better on others, and the same mixed pattern appears for average zero-shot accuracy. This suggests that seam revisiting is a secondary mechanism rather than the main source of the ICBQ gains. Appendix Table 8 shows the same ablation for GPTQ-interleaved CBQ ($T_{\text{pre}} = 0$): seam revisit helps C4 PPL more often, with similarly small margins.

Effect of chunk size. Chunk size is the primary schedule knob: it controls how often ICBQ interleaves cross-block refinement, and in the seam-revisit variant it also increases the number of seams. Table 4 reports PPL on C4 while varying $K \in \{2, 4, 8, L\}$. Empirically, PPL improves monotonically as K decreases.

6 Related Work

We organize prior PTQ work by the axis each method *refines*—block unit, objective, weight code, or basis—and position ICBQ as a *schedule*-level refinement under a fixed pair-level objective.

Refining the block unit and objective. BRECQ (Li et al., 2021), OmniQuant (Shao et al., 2024), and QLLM (Liu et al., 2024a) replace layer-wise with block-wise reconstruction. CBQ (Ding et al., 2025) uses cross-block reconstruction to capture dependencies across multiple blocks; our direct baseline instantiates this principle with fixed adjacent two-block windows. A complementary

Table 4: **Effect of chunk size K on the seven-model set** (PPL on C4 (↓), ternary DBF). $K = L$ is the no-interleaving sequential baseline.

Base model	$K = L$ (Seq.)	$K = 8$	$K = 4$	$K = 2$
Mistral-7B	31.1	18.91	16.8	**14.91**
Llama-2-7B	13.6	12.80	12.5	**12.21**
Llama-2-13B	10.48	10.17	10.09	**9.96**
Llama-3.2-3B	40.63	37.77	37.01	**35.33**
Llama-3-8B	48.98	43.26	42.32	**39.94**
Qwen3-8B	5752.9	33.29	29.90	**28.81**
Qwen3-14B	44.82	25.47	24.67	**23.58**

thread sharpens the local *objective* or accounts for cross-layer effects: QEP (Arai & Ichikawa, 2025) explicitly propagates quantization errors to compensate for their accumulation; LoaQ (Lin & Wan, 2025) introduces output-matching factors to improve output-level approximation within layer-wise PTQ; and LPCD (Ichikawa et al., 2025a) optimizes relaxed objectives over arbitrary submodules and projects the solutions with layer-wise quantizers. ICBQ changes the *schedule* while retaining a fixed pair-level objective; it may be combined with compatible inner objectives.

Refining the weight code and basis. BitNet (Wang et al., 2023), the 1.58-bit line of Ma et al. (2024), DBF (Boža & Macko, 2026), MDBF (Ichikawa et al., 2025b), and OneBit (Xu et al., 2024) sharpen binary or ternary weight representations; ICBQ is agnostic to this axis and uses the selected method as its inner quantizer. QuIP (Chee et al., 2023) applies incoherence processing for low-bit parameter quantization. Rotation-based methods such as QuaRot (Ashkboos et al., 2024) and SpinQuant (Liu et al., 2025) facilitate low-bit weight–activation inference by reducing outlier effects; these transformations target the quantization basis and may be combined with ICBQ as preprocessing.

Schedule theory beyond quantization. The *order* in which subproblems are solved controls propagated error, as in asynchronous block-coordinate descent, multigrid V/W-cycles, and parallel quasi-quantum annealing (Ichikawa & Arai, 2025). ICBQ applies this idea to block-wise PTQ by interleaving short refinements with progressive quantization.

7 Conclusion

Block-wise PTQ has mainly focused on what happens *inside* the reconstruction window; ICBQ also considers *when* the window is applied. By interleaving short cross-block refinements with progressive quantization, ICBQ brings block-wise PTQ a step closer to more global optimization while preserving the inner objective and calibration-data regime. In the reported ternary DBF experiments, ICBQ has lower PPL than Sequential CBQ on the seven main models; it also yields finite PPL for Qwen3-8B and Llama-3-8B configurations with severe Sequential-CBQ degradation. The GPTQ results show that the schedule can also be used with W3 and W2g128. These results suggest that schedule design is a useful additional consideration for low-bit weight-only PTQ, and its benefit seems more pronounced at lower bit-widths.

Limitations. Even with interleaving, the method remains block-local and does not fully capture truly global end-to-end output error. Bridging this gap—both algorithmically and theoretically—is a key direction for future work. In addition, our experiments focus on language models. Whether the same schedule-level gains transfer to other families (e.g., vision models) remains open.

Broader impacts. Potential positive impact: ICBQ lowers the memory and energy footprint of deploying large language models at sub-two-bit precision, which enables wider access to LLM inference on commodity hardware and reduces the carbon cost of inference-heavy workloads. Potential negative impact: a cheaper deployment pipeline could lower the cost barrier for malicious or low-quality deployment of LLMs; however, ICBQ is a pure quantization schedule that does not alter model behavior beyond reconstruction error, so the risk profile matches that of existing PTQ methods such as GPTQ (Frantar et al., 2022) and CBQ (Ding et al., 2025). We view the net impact as positive within the standard PTQ risk envelope.

AI USE STATEMENT

Generative AI tools were used in this work for research ideation, including brainstorming and refining research ideas and providing feedback on aspects of the methodology and experimental design. Generative AI tools were also used to assist with drafting and editing portions of the manuscript for clarity and organization. All methodological choices, implementations, experiments, analyses, and interpretations of the results were reviewed and validated by the authors. The authors reviewed all AI-assisted content for technical correctness and consistency with the reported results and take full responsibility for the final content of this work.

ETHICS STATEMENT

This work does not involve human subjects, personal data, or other sensitive data. The authors have read and followed the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

We provide the implementation details, experimental settings, and supplementary material necessary to reproduce the reported results. Code and additional resources are included in the supplementary material.

REFERENCES

Yamato Arai and Yuma Ichikawa. Quantization error propagation: Revisiting layer-wise post-training quantization. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2504.09629.

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L. Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. QuaRot: Outlier-free 4-Bit inference in rotated LLMs. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3180. URL https://papers.nips.cc/paper_files/paper/2024/hash/b5b939436789f76f08b9d0da5e81af7c-Abstract-Conference.html.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439, 2020. doi: 10.1609/aaai.v34i05.6239. URL https://ojs.aaai.org/index.php/AAAI/article/view/6239.

Vladimír Boža and Vladimír Macko. Addition is almost all you need: Compressing large language models with double binary factorization. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=k5kUKoewdQ.

Adrian Bulat, Yassine Ouali, and Georgios Tzimiropoulos. QBB: Quantization with binary bases for LLMs. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0105. URL https://papers.nips.cc/paper_files/paper/2024/hash/05b69cc4c8ff6e24c5de1ecd27223d37-Abstract-Conference.html.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher M. De Sa. QuIP: 2-bit quantization of large language models with guarantees. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0196. URL https://papers.nips.cc/paper_files/paper/2023/hash/0df38cd13520747e1e64e5b123a78ef8-Abstract-Conference.html.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1300. URL https://aclanthology.org/N19-1300/.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. URL https://arxiv.org/abs/1803.05457.

Xin Ding, Xiaoyu Liu, Zhijun Tu, Yun Zhang, Wei Li, Jie Hu, Hanting Chen, Yehui Tang, Zhiwei Xiong, Baoqun Yin, and Yunhe Wang. CBQ: Cross-block quantization for large language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=eW4yh6HKz4.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022. doi: 10.48550/arXiv.2210.17323. URL https://arxiv.org/abs/2210.17323.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation. Zenodo, 2021. URL https://zenodo.org/records/5371629.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Yuma Ichikawa and Yamato Arai. Optimization by parallel quasi-quantum annealing with gradient-based sampling. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9EfBeXaXf0.

Yuma Ichikawa, Yudai Fujimoto, and Akira Sakai. LPCD: Unified framework from layer-wise to submodule quantization. *arXiv preprint arXiv:2512.01546*, 2025a. URL https://arxiv.org/abs/2512.01546.

Yuma Ichikawa, Yoshihiko Fujisawa, Yudai Fujimoto, Akira Sakai, and Katsuki Fujisawa. More than bits: Multi-envelope double binary factorization for extreme quantization. *arXiv preprint arXiv:2512.24545*, 2025b. URL https://arxiv.org/abs/2512.24545.

Yuma Ichikawa, Keiji Kimura, Akihiro Yoshida, Yudai Fujimoto, Hiroki Tokura, Yamato Arai, Yoshiyuki Ishii, Yusei Kawakami, Genki Shikada, Achille Jacquemond, Yoshihiko Fujisawa, Katsuki Fujisawa, Takumi Honda, and Akira Sakai. OneComp: One-line revolution for generative AI model compression. *arXiv preprint arXiv:2603.28845*, 2026. URL https://arxiv.org/abs/2603.28845.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

Banseok Lee, Dongkyu Kim, Youngcheon You, and Youngmin Kim. LittleBit: Ultra low-bit quantization via latent factorization. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/a917c1d57088897beba47f96b4fd7f3c-Abstract-Conference.html.

Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. BRECQ: Pushing the limit of post-training quantization by block reconstruction. In *International Conference on Learning Representations*, 2021.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *Proceedings of Machine Learning and Systems*, volume 6, 2024. URL https://proceedings.mlsys.org/paper_files/paper/2024/hash/42a452cbafa9dd64e9ba4aa95cc1ef21-Abstract-Conference.html.

Li Lin and Xiaojun Wan. LoaQ: Layer-wise output approximation quantization. *arXiv preprint arXiv:2509.06297*, 2025. URL https://arxiv.org/abs/2509.06297.

Jing Liu, Ruihao Gong, Xiuying Wei, Zhiwei Dong, Jianfei Cai, and Bohan Zhuang. QLLM: Accurate and efficient low-bitwidth quantization for large language models. In *International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=FIplmUWdm3.

Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. LLM-QAT: Data-free quantization aware training for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 467–484. Association for Computational Linguistics, 2024b. doi: 10.18653/v1/2024.findings-acl.26. URL https://aclanthology.org/2024.findings-acl.26/.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. SpinQuant: LLM quantization with learned rotations. In *International Conference on Learning Representations*, 2025.

Shuming Ma, Hongyu Wang, Lingxiao Ma, Lei Wang, Wenhui Wang, Shaohan Huang, Li Dong, Ruiping Wang, Jilong Xue, and Furu Wei. The era of 1-Bit LLMs: All large language models are in 1.58 bits. *arXiv preprint arXiv:2402.17764*, 2024. URL https://arxiv.org/abs/2402.17764.

Meta Llama Team. Llama 3.2 model card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md, 2024. Accessed: 2026-05-01.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260/.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021. doi: 10.1145/3474381.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. OmniQuant: Omnidirectionally calibrated quantization for large language models. In *International Conference on Learning Representations*, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. URL https://arxiv.org/abs/2307.09288.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Haoran Wang, Lingxiao Ma, Fan Yang, Ruiping Wang, Yi Wu, and Furu Wei. BitNet: Scaling 1-Bit transformers for large language models. *arXiv preprint arXiv:2310.11453*, 2023. URL https://arxiv.org/abs/2310.11453.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38087–38099. PMLR, 2023. URL https://proceedings.mlr.press/v202/xiao23c.html.

Yuzhuang Xu, Xu Han, Zonghan Yang, Shuo Wang, Qingfu Zhu, Zhiyuan Liu, Weidong Liu, and Wanxiang Che. OneBit: Towards extremely low-bit large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2122. URL https://papers.nips.cc/paper_files/paper/2024/hash/7a7a3f53faafc0161be0fcb57e5fa078-Abstract-Conference.html.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

APPENDIX CONTENTS

- Appendix A: algorithm and schedule details.
- Appendix B: formal assumptions, theorem statements, proofs, and related theory diagnostics.
- Appendix C: experimental and implementation details.
- Appendix D: supplementary empirical results.

NOTATION

Symbol	Meaning
L	Number of transformer blocks, indexed by $b \in \{0, \dots, L-1\}$.
K	Chunk size used by ICBQ.
$S(L, K)$	Number of chunk seams, $S(L, K) = \lceil L/K \rceil - 1$.
$cK-1$	Pair index of the c-th seam, corresponding to boundary pair $(cK-1, cK)$.
$f_b^\star$	The fp16 teacher block at depth b.
$f_b^{\theta_b}$	The current student block at depth b with parameter state θ_b.
Q	Inner one-block weight-only quantizer, instantiated as DBF or GPTQ.
$X_b^\star$	Teacher activation at depth b on the calibration sample under discussion.
$\hat{X}_b$	Student activation at depth b.
e_b	Depth-b activation mismatch, $e_b = \hat{X}_b - X_b^\star$.
m_i	Midpoint residual of pair $(i, i+1)$, $m_i = \|\hat{X}_{i+1} - X_{i+1}^\star\|_F$.
u_i, v_i	Post-pair midpoint residuals for Sequential CBQ and ICBQ, respectively.
$\mathcal{T}_b$	Finite set of teacher and student activations at depth b visited by either schedule.
Θ_b	Finite set of parameter states assumed by block b during either schedule.
ρ	Uniform teacher-block Lipschitz constant on the visited activation sets.
γ	Per-sample midpoint contraction factor certified for each actual CBQ call.
ϵ_{sub}	Common residual envelope covering local representation error and CBQ fixed-point error.
τ	Shorthand $\tau = \gamma\rho$; the proof assumes $\tau < 1$.
$R, \widetilde{R}$	Recurrence forcing constants, $R = (1+\gamma)\epsilon_{\text{sub}}$ and $\widetilde{R} = (1+\gamma+\gamma^2)\epsilon_{\text{sub}}$.
$s(i)$	Number of seam indices at most i.
$\mathcal{U}_i, \mathcal{V}_i$	Deterministic upper bounds on sequential and interleaved post-pair midpoint residuals.
$\mathcal{B}_L^{\text{seq}}, \mathcal{B}_L^{\text{int}}$	Sequential and interleaved depth-L error upper bounds.
$C(L, K, \rho, \gamma)$	Additive residual constant in the formal ICBQ depth-L bound.

A Algorithm and Schedule Details

We record the full ICBQ schedule as two algorithms. The outer driver (Algorithm 1) implements the progressive weight-only quantization loop: for each block it optionally runs prefit, applies the inner quantizer Q, advances the activation streams, and at every chunk boundary calls the inner refinement followed by an end-of-chunk re-roll. The inner refinement (Algorithm 2) implements the interleaved CBQ call: given the window range $[w_{\min}, w_{\max})$ produced by the outer driver it starts with an in-pass re-roll through blocks $0, \ldots, w_{\min}-1$ and then sweeps through the valid pair indices, running at each pair a rollback-safeguarded Adam update on the plain-MSE two-block CBQ loss $\mathcal{L}_{\mathrm{cbq}}$ followed by a one-layer advance of the student and teacher streams. Here $\mathcal{L}_{\mathrm{pre}}$ denotes the one-block DBF prefit loss, and $\mathcal{L}_{\mathrm{cbq}}$ denotes the plain-MSE two-block reconstruction loss used in our CBQ-style adaptation.

Algorithm 1 Interleaved Cross-Block Quantization (outer progressive driver).

Require: calibration batches $\mathcal{X}$, teacher blocks $\{f_b^\star\}_{b=0}^{L-1}$, inner quantizer Q, chunk size K, prefit steps T_{pre}, inner-quantizer budget T_Q.

Save stored inputs $\{X_0^{(s)}\}_s$.
Initialize $\hat{X}_0 \leftarrow X_0^\star \leftarrow \{X_0^{(s)}\}_s$, $b_{\mathrm{cs}} \leftarrow 0$.
for $b = 0, \ldots, L-1$ **do**
  if $Q = \mathrm{DBF}$ **then**
    Prefit: run T_{pre} AdamW steps on $\mathcal{L}_{\mathrm{pre}}(\theta_b)$ in float32.
  end if
  Quantize: run the inner quantizer Q with budget T_Q on block b.
  Advance: $\hat{X}_{b+1} \leftarrow f_b^Q(\hat{X}_b)$; $X_{b+1}^\star \leftarrow f_b^\star(X_b^\star)$.
  if $b + 1 - b_{\mathrm{cs}} \geq K$ **or** $b = L-1$ **then**
    $w_{\min} \leftarrow \max(0, b_{\mathrm{cs}} - 1)$, $w_{\max} \leftarrow \min(b+1, L-1)$
    Call Algorithm 2 with $[w_{\min}, w_{\max})$.
    Re-roll: recompute $\hat{X}_{0\to b+1}$ and $X_{0\to b+1}^\star$ from stored inputs.
    $b_{\mathrm{cs}} \leftarrow b + 1$.
  end if
end for
return quantized model $\{f_b^Q\}_b$ with refined parameters.

Algorithm 2 Interleaved CBQ refinement on a window range $[w_{\min}, w_{\max})$.

Require: Student state $\{f_b^{\theta_b}\}_{b \leq w_{\max}}$, teacher state $\{f_b^\star\}_{b \leq w_{\max}}$, stored calibration inputs $\{X_0^{(s)}\}_s$.

Re-roll: advance student and teacher copies in lockstep through blocks $0, \ldots, w_{\min}-1$ from the stored inputs.
for $i = w_{\min}, \ldots, w_{\max} - 1$ **do**
  Compute teacher targets $\{Y_{i+2}^{(s),\star}\}_s$ by passing the current teacher state through $f_i^\star, f_{i+1}^\star$.
  Minimize the two-block CBQ loss $\mathcal{L}_{\mathrm{cbq}}$ on (θ_i, θ_{i+1}) with Adam; rollback on non-improvement.
  Advance student and teacher by one layer (through $f_i^{\theta_i}$ and $f_i^\star$).
end for

B Formal Theory and Proofs

This appendix collects the three structural hypotheses (Assumptions B.1, B.2, and B.4), two auxiliary remarks, and the formal theorem and corollary statements used by the detailed proof below. These formal statements correspond to the informal seam-amortized contraction theorem and amortization-gain corollary in the main paper.

B.1 Assumptions

Preamble. Fix a calibration set $\mathcal{X} = \{x^{(n)}\}_{n=1}^{N}$ and a chunk size K. Let $\mathcal{S}_{\mathrm{seq}}$ denote the fixed two-block Sequential CBQ adaptation defined in Section 2, and let $\mathcal{S}_{\mathrm{int}}$ denote ICBQ with this chunk size; both schedules are applied to the same calibration set, use the same fp16 teacher blocks $\{f_b^\star\}_{b=0}^{L-1}$, and use the same inner quantizer Q (including the same DBF prefit objective $\mathcal{L}_{\mathrm{pre}}$ when $Q = \mathrm{DBF}$) and the same rollback-safeguarded Adam budget for every CBQ subproblem. The constants $\rho, \gamma, \epsilon_{\mathrm{sub}}$ below are certified for the schedules at this fixed K; the theorem therefore applies pointwise to any K for which these certificates hold. A claim uniform over multiple chunk sizes requires the same constants to hold uniformly over those choices of K. Both schedules execute deterministic, finitely many forward passes and parameter updates; hence, for each depth b, the set of depth-b student activations $\hat{X}_b^{(n)}$ arising anywhere in either schedule is finite. Denote

$$\mathcal{T}_b \coloneqq \{X_b^{\star,(n)} : n \in [N]\} \cup \{\hat{X}_b^{(n)} \text{ arising in } \mathcal{S}_{\mathrm{seq}}\} \cup \{\hat{X}_b^{(n)} \text{ arising in } \mathcal{S}_{\mathrm{int}}\} \subset \mathbb{R}^{T \times d_{\mathrm{hidden}}}. \quad (16)$$

Assumption B.1 (Lipschitz teacher blocks)**.** *There exists a constant $\rho > 0$, independent of the depth b and of the schedule, such that*

$$\|f_b^\star(x) - f_b^\star(y)\|_F \le \rho \, \|x - y\|_F \ \text{ for all } x, y \in \mathcal{T}_b,\ b \in \{0, \dots, L-1\}. \quad (17)$$

The non-expansive regime $\rho \le 1$ captures empirical residual-stream behavior of trained transformer blocks; our proofs only require $\tau \coloneqq \gamma\rho < 1$, so locally expansive blocks ($\rho > 1$) are admissible if γ is small (Appendix B.10).

For each depth i, let Θ_i denote the finite set of parameter states assumed by block i at any intermediate state of either schedule (initial fp16 teacher parameters, post-prefit, post-inner-quantizer, post-CBQ-update, and any other pre-pass state that appears as an input to a CBQ subproblem).

Assumption B.2 (Contractive two-block CBQ subproblem)**.** *There exist constants $\gamma \in (0, 1)$ and $\epsilon_{\mathrm{sub}} \ge 0$, independent of the depth and of the schedule, such that:*

(Local representation envelope.) *For every i, every $\theta_i \in \Theta_i$, and every $z \in \mathcal{T}_i$,*

$$\left\| f_i^{\theta_i}(z) - f_i^\star(z) \right\|_F \le \epsilon_{\mathrm{sub}}. \quad (18)$$

(Midpoint contraction.) *For every actual CBQ call made by either schedule on a pair $(i, i{+}1)$, let $(\theta_i^{\mathrm{pre}}, \theta_{i+1}^{\mathrm{pre}})$ be the pre-pass block states and $\{\hat{X}_i^{(n)}\}_{n=1}^{N}$ the student inputs; let $(\theta_i^{\mathrm{post}}, \theta_{i+1}^{\mathrm{post}})$ be the shared post-pass pair returned by the rollback-safeguarded Adam optimizer applied to the plain-MSE two-block CBQ loss $\mathcal{L}_{\mathrm{cbq}}$. Then, for every sample n and every $\hat{X}_i = \hat{X}_i^{(n)}$, writing*

$$r^{\mathrm{in}} \coloneqq \left\| f_i^{\theta_i^{\mathrm{pre}}}(\hat{X}_i) - X_{i+1}^\star \right\|_F, \ \ r^{\mathrm{out}} \coloneqq \left\| f_i^{\theta_i^{\mathrm{post}}}(\hat{X}_i) - X_{i+1}^\star \right\|_F, \quad (19)$$

we have

$$r^{\mathrm{out}} \le \gamma \, r^{\mathrm{in}} + \epsilon_{\mathrm{sub}}. \quad (20)$$

Eq. (20) is the linear-convergence form of a first-order method with a nonzero fixed-point residual; Eq. (18) is the local representation envelope needed to transition from a depth-i input error to a depth-$(i{+}1)$ midpoint error. Three points deserve emphasis. First, Eq. (20) is *per-sample*: the optimizer returns a single parameter pair, but the assumption requires the induced midpoint residual to contract uniformly over every presented sample, which is strictly stronger than contraction on the sample-averaged residual. Second, the rollback safeguard only certifies the non-increase of the monitored two-block loss; the per-sample midpoint contraction in Eq. (20) is an additional empirical certificate (Remark B.3 below). Third, if separate constants are preferred, ϵ_{sub} can be split into a local representation radius ϵ_{rep} (in Eq. (18)) and a CBQ fixed-point radius ϵ_{opt} (in Eq. (20)), and every subsequent inequality holds with ϵ_{sub} replaced by $\max(\epsilon_{\mathrm{rep}}, \epsilon_{\mathrm{opt}})$. The aggregate constant γ is estimated empirically in Appendix B.11.

Remark B.3 (Midpoint contraction is an additional certificate, not a consequence of rollback)**.** *The rollback safeguard is applied to the two-block output reconstruction loss $\mathcal{L}_{\mathrm{cbq}}$, which couples blocks i and $i{+}1$ jointly through $f_{i+1}^{\theta_{i+1}} \circ f_i^{\theta_i}$. The midpoint residual depends only on block i, so rollback does* not *by itself imply the per-sample midpoint contraction in Eq.* (20)*. Assumption B.2 is therefore an additional local certificate of the implemented rollback-safeguarded Adam on the finite collection of CBQ calls actually made by the schedules; the proof of Theorem B.5 uses only this midpoint certificate together with the local representation envelope, not the fact that the two-block loss decreases. The empirical study in Appendix B.11 provides an aggregate diagnostic of this behavior, rather than a uniform certificate.*

Assumption B.4 (Exact re-roll)**.** *Every re-roll performed by either schedule has an associated prefix length d: for the* in-pass re-roll *at the start of every interleaved refinement call (line 1 of Algorithm 2), $d = w_{\min}$; for the* end-of-chunk re-roll *in the outer driver, performed after block b is processed, $d = b + 1$. The re-roll recomputes, for each calibration sample n and every target depth $b' \in \{0, \dots, d\}$,*

$$\hat{X}_{b'}^{(n)} = \big(f_{b'-1}^{\theta_{b'-1}} \circ \cdots \circ f_0^{\theta_0}\big)\big(X_0^{(n)}\big), \;\; X_{b'}^{\star,(n)} = \big(f_{b'-1}^{\star} \circ \cdots \circ f_0^{\star}\big)\big(X_0^{(n)}\big), \tag{21}$$

where $\theta_0, \dots, \theta_{b'-1}$ are the current *parameter states of blocks $0, \dots, b'{-}1$ at the moment of the re-roll (with $b' = 0$ denoting the empty composition). In the same symbolic sense, every single-block forward pass executed elsewhere in either schedule is treated as an exact function application of its block map.*

In our implementation, both re-rolls and every in-loop advance step are deterministic forward passes through their current prefixes using the stored layer-0 inputs. Assumption B.4 abstracts away floating-point non-associativity in the symbolic identities used by the proof. The in-pass re-roll identifies the repeated seam input, while the end-of-chunk re-roll keeps the states and activations included in Θ_i and $\mathcal{T}_i$ consistent with the current prefix.

B.2 Formal theorem, corollary, and remarks

For a schedule $\mathcal{S} \in \{\mathcal{S}_{\mathrm{seq}}, \mathcal{S}_{\mathrm{int}}\}$, write $e_b^{\mathcal{S}} \coloneqq \hat{X}_b^{\mathcal{S}} - X_b^{\star}$ for the depth-b activation mismatch on a calibration sample. The expectation $\mathbb{E}$ in the theorem statements is the empirical average over the fixed calibration set $\mathcal{X}$.

Under Assumptions B.1, B.2, and B.4 and $\tau < 1$, define the *sequential depth-L bound*

$$\mathcal{B}_L^{\mathrm{seq}} \coloneqq \rho\,(1+\gamma)\,\frac{1-\tau^{L-1}}{1-\tau}\,\epsilon_{\mathrm{sub}} + \epsilon_{\mathrm{sub}}, \tag{22}$$

which is the sequential upper bound used for the comparison (derivation in Appendix B.8).

Theorem B.5 (Seam-amortized error contraction, formal)**.** *Under Assumptions B.1, B.2, and B.4 and $\tau = \gamma\rho < 1$, for every $L \geq 2$ and every $1 \leq K \leq L$,*

$$\mathbb{E}\big\|e_L^{\mathrm{int}}\big\|_F \leq \gamma^{S(L,K)}\,\mathcal{B}_L^{\mathrm{seq}} + C(L,K,\rho,\gamma)\,\epsilon_{\mathrm{sub}}, \tag{23}$$

where $S(L,K) = \lceil L/K \rceil - 1$, and one admissible (generally non-tight) choice of the residual constant is

$$\begin{aligned} C(L,K,\rho,\gamma) = {} & \big(1 - \gamma^{S(L,K)}\big) \\ & + \Big[(1+\gamma+\gamma^2) - \gamma^{S(L,K)}(1+\gamma)\Big]\,\rho\,\frac{1-\tau^{L-1}}{1-\tau}. \end{aligned} \tag{24}$$

In particular,

$$C(L,K,\rho,\gamma) \leq 1 + \frac{(1+\gamma+\gamma^2)\,\rho}{1-\tau} \;\; \textit{uniformly in } L, K, \tag{25}$$

and for the standard stability regime $\rho \leq 1$ this simplifies to $C \leq 1 + (1+\gamma+\gamma^2)/(1-\tau)$; for $K = L$ the leading factor is $\gamma^0 = 1$ (one chunk, no seams), and for $K = 2$ it contracts by $\gamma^{\lceil L/2 \rceil - 1}$. The per-seam factor γ is tight at the recurrence level (Proposition B.20).

Remark B.6 (Non-tightness of the displayed residual constant)**.** *The constant in Eq.* (24) *is chosen for a simple closed-form comparison with the sequential bound $\mathcal{B}_L^{\mathrm{seq}}$. It is one admissible upper-bound*

constant produced by the bookkeeping in Proposition B.14—specifically by the pointwise estimate $\beta_k \le \widetilde{R}$ used in Eq. (48)—and is generally not the smallest constant one could write down. In particular, when $K = L$ there is a single chunk and no seam ($S = 0$), so ICBQ and sequential CBQ coincide as schedules at the recurrence level; the displayed residual constant nevertheless reduces only to $C(L, L, \rho, \gamma) = \gamma^2\rho(1 - \tau^{L-1})/(1 - \tau)$ rather than to zero, because the same $\beta_k \le \widetilde{R}$ estimate is applied to non-seam pairs in Eq. (48). A sharper seam-aware comparison that vanishes in the no-seam case is recorded in Proposition B.15 below.

Remark B.7 (Upper-bound semantics)**.** *The formal ICBQ depth-L bound in Theorem B.5 compares ICBQ's expected depth-L error against the explicit sequential upper bound $\mathcal{B}_L^{\mathrm{seq}}$ derived from the same recurrences, rather than against the realized sequential error. The comparison is therefore a bound-level statement: the leading propagated term carries the factor γ^S, while all additive terms that do not carry this factor are collected in $C\,\epsilon_{\mathrm{sub}}$.*

Corollary B.8 (Geometric decay in the number of seams, formal)**.** *When $\mathcal{B}_L^{\mathrm{seq}} > 0$, define the* amortization gain *of ICBQ over sequential CBQ as $G(L, K) := \mathcal{B}_L^{\mathrm{seq}}/\mathcal{B}_L^{\mathrm{int}}$, where $\mathcal{B}_L^{\mathrm{int}} = \gamma^{S(L,K)}\mathcal{B}_L^{\mathrm{seq}} + C(L, K, \rho, \gamma)\,\epsilon_{\mathrm{sub}}$ is the explicit right-hand side of Eq. (23). Under the hypotheses of Theorem B.5, if for some $\delta \ge 0$,*

$$C(L, K, \rho, \gamma)\,\epsilon_{\mathrm{sub}} \le \delta\,\gamma^{S(L,K)}\,\mathcal{B}_L^{\mathrm{seq}}, \tag{26}$$

then

$$G(L, K) \ge \frac{\gamma^{-S(L,K)}}{1 + \delta}. \tag{27}$$

Consequently, for any family of instances in which $C(L, K, \rho, \gamma)\,\epsilon_{\mathrm{sub}} = o\big(\gamma^{S(L,K)}\mathcal{B}_L^{\mathrm{seq}}\big)$,

$$G(L, K) \ge \gamma^{-S(L,K)}\big(1 - o(1)\big) = \gamma^{-(\lceil L/K\rceil - 1)}\big(1 - o(1)\big). \tag{28}$$

Proof. By Theorem B.5, the interleaved upper bound is $\mathcal{B}_L^{\mathrm{int}} = \gamma^{S(L,K)}\mathcal{B}_L^{\mathrm{seq}} + C(L, K, \rho, \gamma)\,\epsilon_{\mathrm{sub}}$ with $C \ge 0$, so $\mathcal{B}_L^{\mathrm{int}} \le (1 + \delta)\gamma^{S(L,K)}\mathcal{B}_L^{\mathrm{seq}}$ under the hypothesis, and $\mathcal{B}_L^{\mathrm{int}} \ge \gamma^{S(L,K)}\mathcal{B}_L^{\mathrm{seq}} > 0$. Dividing yields $G(L, K) \ge 1/\big((1 + \delta)\gamma^{S(L,K)}\big)$, which is Eq. (27). The $o(1)$ version follows by setting $\delta_{L,K} := C\epsilon_{\mathrm{sub}}/(\gamma^{S(L,K)}\mathcal{B}_L^{\mathrm{seq}})$. □

B.3 Proof Overview

We use the notation of the seam-amortized contraction statement. For a fixed calibration sample $x \in \mathcal{X}$, write $e_b := \hat{X}_b - X_b^\star$ for the activation mismatch at depth b, and

$$m_i := \|\hat{X}_{i+1} - X_{i+1}^\star\|_F = \|e_{i+1}\|_F, \;\; i \in \{-1, 0, \ldots, L-2\}, \tag{29}$$

for the midpoint residual of pair $(i, i{+}1)$, which equals the activation mismatch at depth $i{+}1$. We adopt $i = -1$ as a formal boundary index corresponding to the common calibration input, so that $m_{-1} = \|\hat{X}_0 - X_0^\star\|_F = 0$ since both schedules start from $\hat{X}_0 = X_0^\star$; this zero initial condition is used as the base case in Lemma B.9 below. We use the shorthand $\tau := \gamma\rho$ throughout and assume, as in Theorem B.5, that $\tau < 1$. Whenever the same symbol could arise at several moments of the algorithm, it is tied to the specific schedule and time stated in the surrounding sentence; superscripts such as "seq", "int", "pre", "post", and "out" are used only where the distinction is needed for an inequality. All statements below are first pathwise for this fixed calibration sample, and are averaged only in Section B.8.

The structure of the proof is the following.

1. Section B.4 (Lemma B.9): a uniform one-step recursion on m_i that holds under both schedules for any pair receiving a single CBQ pass.
2. Section B.5 (Lemma B.10): a formal record of the structural fact that every chunk seam is visited by the inner refinement of two consecutive chunks.
3. Section B.6 (Lemma B.11): isolates the effect of a single boundary revisit, where the second pass adds one extra factor γ relative to a single pass.
4. Section B.7 (Lemma B.13 and Propositions B.14–B.15): chains the per-pair and seam recurrences along the depth axis, producing closed-form upper bounds $\mathcal{U}_i, \mathcal{V}_i$ on the midpoint residual at each pair.

5. Section B.8: extracts the final-prefix identity for $\hat{X}_{L-1}$ via Lemma B.17 (final-visit-time monotonicity) and Lemma B.18 (final-prefix invariance), and then assembles Theorem B.5.
6. Section B.9 (Proposition B.20): the per-seam factor γ^S is tight at the recurrence level.
7. Section B.10 (Proposition B.21): extension to the locally expansive regime $\rho > 1$ under $\tau < 1$.

B.4 Per-pair recursion

Lemma B.9 (Per-pair recursion). *Fix a schedule $\mathcal{S} \in \{\mathcal{S}_{\mathrm{seq}}, \mathcal{S}_{\mathrm{int}}\}$. For every pair $(i, i{+}1)$ that receives a single CBQ pass under $\mathcal{S}$, the midpoint residual obeys*

$$m_i^{\mathrm{out}} \le \gamma\rho\, m_{i-1}^{\mathrm{out}} + (1+\gamma)\,\epsilon_{\mathrm{sub}}, \quad i \ge 0, \tag{30}$$

with $m_{-1}^{\mathrm{out}} \coloneqq 0$ (both schedules start from the common calibration input $\hat{X}_0 = X_0^\star$). The inequality holds pathwise for every calibration sample.

Proof. Fix an arbitrary calibration sample $n \in \{1, \dots, N\}$; we suppress the superscript $^{(n)}$ in what follows since every quantity below is evaluated on this single sample. Let $\hat{X}_i$ and $X_i^\star$ denote the student and teacher activations at depth i in $\mathcal{S}$ immediately before pair $(i, i{+}1)$ is processed. By Assumption B.1, every student activation arising in either schedule at depth i lies in $\mathcal{T}_i$, as does $X_i^\star$; hence $\hat{X}_i, X_i^\star \in \mathcal{T}_i$. By Eq. (29), $m_{i-1}^{\mathrm{out}} = \|\hat{X}_i - X_i^\star\|_F$. Let θ_i^{pre} and $\theta_{i+1}^{\mathrm{pre}}$ denote the parameters of blocks i and $i{+}1$ at that moment (obtained from the prefit step when present, the inner quantizer, and any earlier CBQ updates that $\mathcal{S}$ has performed on those blocks). Both states arise at an intermediate step of $\mathcal{S}$, so $\theta_i^{\mathrm{pre}} \in \Theta_i$ and $\theta_{i+1}^{\mathrm{pre}} \in \Theta_{i+1}$ by the construction of Θ_i, Θ_{i+1} in Assumption B.2. Thus the CBQ call currently being processed is one of the actual calls covered by the midpoint contraction in Assumption B.2: its pre-pass states lie in $\Theta_i \times \Theta_{i+1}$ and its sample input $\hat{X}_i$ lies in $\mathcal{T}_i$.

Bound on the pre-pass residual. The pre-pass midpoint residual of pair $(i, i{+}1)$ is

$$r^{\mathrm{in}} = \big\|f_i^{\theta_i^{\mathrm{pre}}}(\hat{X}_i) - X_{i+1}^\star\big\|_F = \big\|f_i^{\theta_i^{\mathrm{pre}}}(\hat{X}_i) - f_i^\star(X_i^\star)\big\|_F,$$

where we used $X_{i+1}^\star = f_i^\star(X_i^\star)$. By the triangle inequality,

$$\begin{aligned} r^{\mathrm{in}} &\le \underbrace{\big\|f_i^{\theta_i^{\mathrm{pre}}}(\hat{X}_i) - f_i^\star(\hat{X}_i)\big\|_F}_{\le\, \epsilon_{\mathrm{sub}} \text{ by Eq. (18)}} + \underbrace{\big\|f_i^\star(\hat{X}_i) - f_i^\star(X_i^\star)\big\|_F}_{\le\, \rho\,\|\hat{X}_i - X_i^\star\|_F = \rho\, m_{i-1}^{\mathrm{out}} \text{ by Asm. B.1}} \\ &\le \rho\, m_{i-1}^{\mathrm{out}} + \epsilon_{\mathrm{sub}}. \end{aligned} \tag{31}$$

where the first bound applies the local envelope in Eq. (18) (valid because $\theta_i^{\mathrm{pre}} \in \Theta_i$ and $\hat{X}_i \in \mathcal{T}_i$) and the second uses Assumption B.1 (valid because $\hat{X}_i, X_i^\star \in \mathcal{T}_i$).

Applying the midpoint contraction. Using Eq. (20) on the single pass on pair $(i, i{+}1)$,

$$m_i^{\mathrm{out}} = r^{\mathrm{out}} \le \gamma\, r^{\mathrm{in}} + \epsilon_{\mathrm{sub}} \le \gamma\big(\rho\, m_{i-1}^{\mathrm{out}} + \epsilon_{\mathrm{sub}}\big) + \epsilon_{\mathrm{sub}} = \gamma\rho\, m_{i-1}^{\mathrm{out}} + (1+\gamma)\,\epsilon_{\mathrm{sub}}, \tag{32}$$

where the first equality uses $m_i^{\mathrm{out}} = \|\hat{X}_{i+1} - X_{i+1}^\star\|_F = \|f_i^{\theta_i^{\mathrm{post}}}(\hat{X}_i) - X_{i+1}^\star\|_F = r^{\mathrm{out}}$ (since the student advance from $\hat{X}_i$ to $\hat{X}_{i+1}$ uses the post-update block i). This is exactly Eq. (30). The base case $m_{-1}^{\mathrm{out}} = 0$ holds because both schedules share the common calibration input $\hat{X}_0 = X_0^\star$; the convention $m_{-1}^{\mathrm{out}} = \|\hat{X}_0 - X_0^\star\|_F$ follows from Eq. (29) with $i = -1$. □

B.5 Boundary overlap under chunked interleaving

Before isolating the seam contraction quantitatively, we first record, as a separate lemma, the structural "visited twice" property of the ICBQ schedule.

Lemma B.10 (Seam structural property). *Let $L \ge 2$ and $1 \le K \le L$, with $S = S(L,K) = \lceil L/K \rceil - 1$ the number of chunk seams. For every $c \in \{1, \dots, S\}$:*

(a) *the seam index $cK{-}1$ is a valid pair index, i.e. $cK{-}1 \le L{-}2$;*

(b) *the boundary pair* $(cK-1, cK)$ *is the* last *pair processed in chunk* c*'s inner refinement, and the* first *pair processed in chunk* $c+1$*'s inner refinement;*

(c) *under ICBQ, the boundary pair* $(cK-1, cK)$ *is therefore optimized exactly twice.*

Proof. Validity (a). If $S = 0$ there is nothing to prove; assume henceforth $S \geq 1$. Write $L = qK + r$ with $q = \lfloor L/K \rfloor$ and $0 \leq r < K$. If $r = 0$ then $\lceil L/K \rceil = q$, $S = q-1 \geq 1$ (so $q \geq 2$), and $SK = (q-1)K = L-K \leq L-1$ since $K \geq 1$. If $r \geq 1$ then $\lceil L/K \rceil = q+1$, $S = q \geq 1$, and $SK = qK = L-r \leq L-1$ since $r \geq 1$. In both cases $SK \leq L-1$, hence for every $c \in \{1, \ldots, S\}$ the seam index satisfies $cK-1 \leq SK-1 \leq L-2$, which proves (a).

Pair-loop placement (b). For chunk $c \in \{1, \ldots, S\}$, the chunk-pointer satisfies $b_{\text{cs}} = (c-1)K$ at the start of chunk c, and the chunk closes when the outer driver reaches $b = cK-1$ (i.e. $b+1 - b_{\text{cs}} = K$). By the chunk-closure window rule, the window range at this closure is $[w_{\min}, w_{\max}) = [\max(0, (c-1)K-1), \min(cK, L-1))$. Since $cK \leq L-1$ by (a), $w_{\max} = cK$, hence the largest valid pair index in chunk c's pair loop is $w_{\max} - 1 = cK-1$, so the seam pair $(cK-1, cK)$ is the *last* pair processed in chunk c's inner refinement. Conversely, chunk $c+1$ starts with $b_{\text{cs}} = cK$, so its window range begins at $w_{\min} = \max(0, cK-1) = cK-1$ (since $cK \geq K \geq 1$), making $(cK-1, cK)$ the *first* pair processed in chunk $c+1$'s inner refinement.

Multiplicity (c). By (b), the seam pair $(cK-1, cK)$ is processed once at the end of chunk c's pair loop and once at the start of chunk $c+1$'s pair loop, hence optimized exactly twice in total. □

B.6 SINGLE-SEAM LEMMA

The following lemma isolates the contribution of a single boundary revisit in ICBQ, using the one-pass sequential CBQ schedule as the reference case.

Lemma B.11 (Single-seam contraction)**.** *Fix* $c \in \{1, \ldots, S\}$ *and consider the seam pair* $(cK-1, cK)$*. Let* $\hat{X}_{cK-1} \in \mathcal{T}_{cK-1}$ *be the student activation at depth* $cK-1$ *immediately before the first CBQ pass on this pair, and let* $(\theta^{\text{pre}}_{cK-1}, \theta^{\text{pre}}_{cK}) \in \Theta_{cK-1} \times \Theta_{cK}$ *be the pre-first-pass parameters of blocks* $cK-1$ *and* cK*. For the actual ICBQ seam state considered in part* (ii)*, the second component is the fp16 teacher copy,* $\theta^{\text{pre}}_{cK} = \theta^{\star}_{cK}$*; part* (i) *records the contraction certified for this first pass. Let*

$$\widetilde{m} \coloneqq \big\| f^{\theta^{\text{pre}}_{cK-1}}_{cK-1}(\hat{X}_{cK-1}) - X^{\star}_{cK} \big\|_F. \tag{33}$$

Then:

(i) (One certified pass.) *The first CBQ pass on* $(cK-1, cK)$ *from this actual seam state leaves a post-pair midpoint residual*

$$m^{(1),\text{out}}_{cK-1} \leq \gamma\, \widetilde{m} + \epsilon_{\text{sub}}. \tag{34}$$

(ii) (Interleaved.) *Under ICBQ, the seam pair* $(cK-1, cK)$ *is visited* twice *(Lemma B.10): pass* 1 *is the last pair in chunk* c*'s inner refinement, and pass* 2 *is the first pair in chunk* $c+1$*'s inner refinement. Between the two passes the outer driver executes the end-of-chunk re-roll at the close of chunk* c*, the prefit/inner-quantizer/one-layer-advance sequence on blocks* $cK, \ldots, \min\{(c+1)K-1, L-1\}$*, and the in-pass re-roll opening chunk* $c+1$*'s refinement; none of these operations modifies the parameters of blocks* $0, \ldots, cK-1$*. In this setting the post-seam midpoint residual satisfies*

$$m^{\text{int,out}}_{cK-1} \leq \gamma^2\, \widetilde{m} + (1+\gamma)\, \epsilon_{\text{sub}}. \tag{35}$$

Proof. Claim (i). The pre-pass state $(\theta^{\text{pre}}_{cK-1}, \theta^{\text{pre}}_{cK})$ of the first seam pass is admissible by hypothesis, i.e. it lies in $\Theta_{cK-1} \times \Theta_{cK}$; the student activation $\hat{X}_{cK-1}$ presented to the pair lies in $\mathcal{T}_{cK-1}$. Since this is an actual CBQ call of the interleaved schedule, the midpoint contraction in Eq. (20) applies directly and yields Eq. (34).

Claim (ii). Write $\theta^{(1),\text{post}}_{cK-1}$ and $\theta^{(1),\text{post}}_{cK}$ for the states of blocks $cK-1$ and cK after pass 1; Adam updates both block parameters jointly during the look-ahead pass, despite the pre-pass state of block cK being the fp16 teacher copy $\theta^{\star}_{cK}$, which lies in Θ_{cK} by the definition preceding Assumption B.2.

Since the pre-pass pair $(\theta^{\mathrm{pre}}_{cK-1}, \theta^{\mathrm{pre}}_{cK}) = (\theta^{\mathrm{pre}}_{cK-1}, \theta^{\star}_{cK})$ is admissible and $\hat{X}_{cK-1} \in \mathcal{T}_{cK-1}$, Assumption B.2 applied to the actual pass 1 provides $(\theta^{(1),\mathrm{post}}_{cK-1}, \theta^{(1),\mathrm{post}}_{cK}) \in \Theta_{cK-1} \times \Theta_{cK}$ together with the pathwise midpoint contraction

$$r^{(1)} := \left\| f^{\theta^{(1),\mathrm{post}}_{cK-1}}_{cK-1}(\hat{X}_{cK-1}) - X^{\star}_{cK} \right\|_F \le \gamma\, \widetilde{m} + \epsilon_{\mathrm{sub}}. \tag{36}$$

Observe that $r^{(1)}$ depends on block cK only through the teacher activation $X^{\star}_{cK} = f^{\star}_{cK-1}(X^{\star}_{cK-1})$ and not through $\theta^{(1),\mathrm{post}}_{cK}$.

We now show that the pre-pass-2 midpoint residual equals $r^{(1)}$. Let $\hat{X}^{(2)}_{cK-1}$ denote the student activation at depth $cK-1$ presented to pass 2, and let $\theta^{(2),\mathrm{pre}}_{cK-1}$ denote the pre-pass-2 state of block $cK-1$. It suffices to establish the two identities

$$\hat{X}^{(2)}_{cK-1} = \hat{X}_{cK-1}, \;\; \theta^{(2),\mathrm{pre}}_{cK-1} = \theta^{(1),\mathrm{post}}_{cK-1}. \tag{37}$$

Prefix invariance. By Lemma B.10(b), pass 1 is the *last* pair processed in chunk c's inner refinement (its index equals $w_{\max} - 1 = cK-1$), and pass 2 is the *first* pair processed in chunk $c+1$'s inner refinement (its index equals $w_{\min} = cK-1$ for the next chunk). We claim that every parameter $\theta_{b''}$ with $b'' \in \{0, \dots, cK-2\}$ takes the *same* value at the start of pass 1 (when chunk c's pair loop begins processing pair $(cK-1, cK)$) and at the start of pass 2; denote this common value $\theta^{(1)}_{b''} \in \Theta_{b''}$. Indeed, pass 1 is a two-block CBQ subproblem on $(cK-1, cK)$, and its rollback-safeguarded Adam update modifies *only* θ_{cK-1} and θ_{cK} (no other parameter vector enters the CBQ loss $\mathcal{L}_{\mathrm{cbq}}$ for this pair, and Algorithm 2's advance step at line 5 is a read-only evaluation that does not update θ). Hence $\theta_{b''}$ with $b'' \le cK-2$ is unchanged by pass 1, so its pre-pass-1 and end-of-pass-1 values agree.

Between the end of pass 1 and the start of pass 2, the outer driver executes exactly

(a) the end-of-chunk activation re-roll at the close of chunk c;

(b) for each $b' \in \{cK, cK+1, \dots, \min\{(c+1)K-1, L-1\}\}$, the prefit step (when present), the inner quantizer Q on $\theta_{b'}$, and a one-layer advance of the student and teacher streams through the updated block b';

(c) the in-pass re-roll that opens chunk $c+1$'s inner refinement.

Operations (a) and (c) recompute activations only, not parameters. Operation (b) modifies $\theta_{b'}$ only for $b' \ge cK$. Hence every parameter $\theta_{b''}$ with $b'' \in \{0, \dots, cK-1\}$ is unchanged between the end of pass 1 and the start of pass 2. Combining this with the invariance under pass 1 for $b'' \le cK-2$,

$$\theta^{\text{start of pass 1}}_{b''} = \theta^{\text{end of pass 1}}_{b''} = \theta^{\text{start of pass 2}}_{b''} \;\text{ for } b'' \in \{0, \dots, cK-2\}, \tag{38}$$

which we take as the definition of $\theta^{(1)}_{b''}$; membership $\theta^{(1)}_{b''} \in \Theta_{b''}$ follows from the construction of $\Theta_{b''}$ in Assumption B.2. For $b'' = cK-1$, operations (a)–(c) leave θ_{cK-1} unchanged, so $\theta^{\text{end of pass 1}}_{cK-1} = \theta^{\text{start of pass 2}}_{cK-1} = \theta^{(1),\mathrm{post}}_{cK-1} = \theta^{(2),\mathrm{pre}}_{cK-1}$ by the very definition of these symbols.

Second identity in Eq. (37). The chain $\theta^{\text{end of pass 1}}_{cK-1} = \theta^{\text{start of pass 2}}_{cK-1}$ just established provides $\theta^{(2),\mathrm{pre}}_{cK-1} = \theta^{(1),\mathrm{post}}_{cK-1}$; in particular $\theta^{(2),\mathrm{pre}}_{cK-1} \in \Theta_{cK-1}$.

First identity in Eq. (37). For $q \in \{1, 2\}$, let $\Phi^{(q)}$ denote the prefix map from depth 0 to depth $cK-1$ using the parameters at the start of pass q:

$$\Phi^{(q)} := \begin{cases} \mathrm{Id}, & cK-1 = 0, \\ f^{\theta^{\text{start of pass } q}_{cK-2}}_{cK-2} \circ \cdots \circ f^{\theta^{\text{start of pass } q}_{0}}_{0}, & cK-1 \ge 1. \end{cases} \tag{39}$$

By Assumption B.4, the in-pass re-roll at the start of pass 2 produces $\hat{X}^{(2)}_{cK-1} = \Phi^{(2)}(X_0)$ on the stored calibration inputs. Inside chunk c's inner-refinement pair loop, the same prefix map is applied incrementally before processing pair $(cK-1, cK)$, so $\hat{X}_{cK-1} = \Phi^{(1)}(X_0)$. By Eq. (38), the two prefix maps use *identical* parameters $\theta^{(1)}_{b''}$ for every $b'' \in \{0, \dots, cK-2\}$; when this index set is

empty, both maps are Id. Hence $\hat{X}^{(2)}_{cK-1} = \hat{X}_{cK-1}$, which establishes the first identity. In particular, $\hat{X}^{(2)}_{cK-1} \in \mathcal{T}_{cK-1}$ since $\hat{X}_{cK-1} \in \mathcal{T}_{cK-1}$ by hypothesis.

Applying the contraction to pass 2. We first record the pre-pass-2 state of block cK. Operation (b) in the interval between pass 1 and pass 2 applies to block cK (the $b' = cK$ step) either the prefit step followed by the inner quantizer Q (when $Q = \mathrm{DBF}$) or Q alone (when $Q = \mathrm{GPTQ}$); in both cases the resulting post-inner-quantizer state lies in Θ_{cK} by the definition preceding Assumption B.2. Subsequent operations in the interval (the in-pass re-roll (c), the remaining prefit/inner-quantizer steps for $b' \in \{cK{+}1, \dots, \min\{(c{+}1)K{-}1, L{-}1\}\}$ if any, and the earlier end-of-chunk re-roll (a)) do not touch θ_{cK}. Hence $\theta^{(2),\mathrm{pre}}_{cK}$ equals this post-inner-quantizer state and lies in Θ_{cK}. Combined with $\theta^{(2),\mathrm{pre}}_{cK-1} \in \Theta_{cK-1}$ and $\hat{X}^{(2)}_{cK-1} \in \mathcal{T}_{cK-1}$, this actual pass 2 is covered by the midpoint-contraction certificate Eq. (20). The teacher activation $X^\star_{cK}$ is unchanged between the two passes because the teacher blocks are fixed by Assumption B.1 and are reproduced verbatim by the teacher-side re-rolls of Assumption B.4. Combining the two identities in Eq. (37),

$$r^{(2),\mathrm{in}} \coloneqq \big\| f^{\theta^{(2),\mathrm{pre}}_{cK-1}}_{cK-1}(\hat{X}^{(2)}_{cK-1}) - X^\star_{cK} \big\|_F = \big\| f^{\theta^{(1),\mathrm{post}}_{cK-1}}_{cK-1}(\hat{X}_{cK-1}) - X^\star_{cK} \big\|_F = r^{(1)}. \tag{40}$$

Crucially, $r^{(2),\mathrm{in}}$ depends on block cK only through the teacher target $X^\star_{cK}$ and not through $\theta^{(2),\mathrm{pre}}_{cK}$; the transition $\theta^{(1),\mathrm{post}}_{cK} \mapsto \theta^{(2),\mathrm{pre}}_{cK}$ induced by the intervening prefit/inner-quantizer steps therefore does not enter the midpoint residual. This transition does enter the subsequent Adam trajectory of pass 2, since the aggregate CBQ loss $\mathcal{L}_{\mathrm{cbq}}$ depends on both block parameters; Assumption B.2 accommodates this by certifying the post-pass midpoint contraction for the actual admissible pre-pass pair $(\theta^{(2),\mathrm{pre}}_{cK-1}, \theta^{(2),\mathrm{pre}}_{cK}) \in \Theta_{cK-1} \times \Theta_{cK}$ produced by operation (b). Applying Eq. (20) to pass 2 with input residual $r^{(2),\mathrm{in}} = r^{(1)}$ and combining with Eq. (36) yields

$$m^{\mathrm{int,out}}_{cK-1} \le \gamma\, r^{(2),\mathrm{in}} + \epsilon_{\mathrm{sub}} = \gamma\, r^{(1)} + \epsilon_{\mathrm{sub}} \le \gamma^2\, \widetilde{m} + (1+\gamma)\, \epsilon_{\mathrm{sub}}, \tag{41}$$

which is Eq. (35). □

Remark B.12. *The ratio of ICBQ's post-seam residual to the sequential one is γ in the leading term, at the cost of a single additional $\gamma\,\epsilon_{\mathrm{sub}}$ in the residual term. The leading-term gain is therefore visible whenever this additive term is small relative to the incoming seam residual.*

B.7 Chain along the depth axis

We now chain Lemmas B.9 and B.11 along the full depth $0, \dots, L{-}1$. Let $u_i \coloneqq m^{\mathrm{seq,out}}_i$ and $v_i \coloneqq m^{\mathrm{int,out}}_i$ denote the post-pair midpoint residuals of the two schedules at pair $i = 0, \dots, L{-}2$. Write

$$\tau \coloneqq \gamma\rho, \quad R \coloneqq (1+\gamma)\,\epsilon_{\mathrm{sub}}, \quad \widetilde{R} \coloneqq (1+\gamma+\gamma^2)\,\epsilon_{\mathrm{sub}}, \tag{42}$$

and let $s(i) \coloneqq |\{c \in \{1, \dots, S\} : cK{-}1 \le i\}|$ be the number of chunk seams whose index is at most i. Both schedules share the initial condition $u_{-1} = v_{-1} = 0$ (the calibration-sample-matched inputs) and satisfy the recurrences in the following lemma; the remaining work of this subsection is to unroll them.

Lemma B.13 (Per-pair recurrences)**.** *For every $i \in \{0, \dots, L{-}2\}$,*

(S) $u_i \le \tau\, u_{i-1} + R$ *uniformly over i* (sequential CBQ)*;*

(I) $v_i \le \tau\, v_{i-1} + R$ *if i is a non-seam pair, and $v_i \le \gamma\tau\, v_{i-1} + \widetilde{R}$ if $i = cK{-}1$ is a seam pair* (ICBQ)*.*

Proof. Statement (S). Sequential CBQ processes every pair exactly once, so Lemma B.9 applied to pair $(i, i{+}1)$ yields $u_i \le \gamma\rho\, u_{i-1} + (1+\gamma)\,\epsilon_{\mathrm{sub}} = \tau\, u_{i-1} + R$.

Non-seam case of (I). For a non-seam pair i that receives a single CBQ pass, Lemma B.9 gives $v_i \le \tau\, v_{i-1} + R$, identical to (S).

Seam case of (I). Let $i = cK{-}1$ be a seam pair. In the interleaved schedule, let $\hat{X}_{cK-1}$ denote the student activation at depth $cK{-}1$ presented to pass 1, and let $(\theta^{\mathrm{pre}}_{cK-1}, \theta^{\mathrm{pre}}_{cK})$ denote the pre-pass-1 state of blocks $cK{-}1$ and cK. Both components of this pair arise at an intermediate step

of $\mathcal{S}_{\text{int}}$: $\theta^{\text{pre}}_{cK-1}$ is either the state left by the previous pair update in chunk c's inner refinement (when $cK-1 \geq 1$ and chunk c's pair loop has processed at least one pair before the seam) or the post-inner-quantizer state of block $cK-1$ (when the seam is the first pair processed in chunk c, e.g. for $c = 1, K = 1$); both belong to Θ_{cK-1} by the definition preceding Assumption B.2. The other component $\theta^{\text{pre}}_{cK} = \theta^{\star}_{cK}$ is the fp16 teacher copy (block cK has not yet been processed by the prefit or inner-quantizer step at the start of pass 1), which belongs to Θ_{cK} by the same definition. Hence $(\theta^{\text{pre}}_{cK-1}, \theta^{\text{pre}}_{cK}) \in \Theta_{cK-1} \times \Theta_{cK}$, verifying the hypothesis of Lemma B.11. By the construction of $\mathcal{T}_{cK-1}$ in Assumption B.1, $\hat{X}_{cK-1}, X^{\star}_{cK-1} \in \mathcal{T}_{cK-1}$.

We claim that $\|\hat{X}_{cK-1} - X^{\star}_{cK-1}\|_F = v_{i-1}$ (with the convention $v_{-1} = 0$) in both of the following cases:

- If $cK-1 \geq 1$ (which includes $c \geq 2$, and $c = 1$ with $K \geq 2$), chunk c's inner-refinement loop processes pair $(cK-2, cK-1)$ before pair $(cK-1, cK)$—the latter is the last pair in chunk c's window—so $\hat{X}_{cK-1}$ equals the post-pair-$(cK-2)$ student activation, and $\|\hat{X}_{cK-1} - X^{\star}_{cK-1}\|_F = m^{\text{int,out}}_{cK-2} = v_{i-1}$ by Eq. (29).
- If $cK-1 = 0$ (which happens only for $c = 1, K = 1$), the seam pair $(0, 1)$ is the only pair in chunk 1's window, and the in-pass re-roll (Assumption B.4) gives $\hat{X}_0 = X^{\star}_0$; hence $\|\hat{X}_0 - X^{\star}_0\|_F = 0 = v_{-1}$ by the boundary convention.

In either case $\|\hat{X}_{cK-1} - X^{\star}_{cK-1}\|_F = v_{i-1}$.

Let $\widetilde{m} = \|f^{\theta^{\text{pre}}_{cK-1}}_{cK-1}(\hat{X}_{cK-1}) - X^{\star}_{cK}\|_F$ denote the pre-seam midpoint residual presented to pass 1 (in the notation of Lemma B.11). By the same triangle-inequality derivation as in Eq. (31) (local envelope in Eq. (18) applied with $\theta_{cK-1} = \theta^{\text{pre}}_{cK-1}, z = \hat{X}_{cK-1}$, and Assumption B.1 applied with $x = \hat{X}_{cK-1}, y = X^{\star}_{cK-1}$),

$$\widetilde{m} \leq \rho \|\hat{X}_{cK-1} - X^{\star}_{cK-1}\|_F + \epsilon_{\text{sub}} = \rho v_{i-1} + \epsilon_{\text{sub}}. \tag{43}$$

Applying Lemma B.11(ii) (the two-pass seam contraction $v_i \leq \gamma^2 \widetilde{m} + (1+\gamma)\epsilon_{\text{sub}}$) with the hypotheses just verified, and combining with Eq. (43) and $\tau = \gamma\rho$,

$$\begin{aligned} v_i &\leq \gamma^2 \widetilde{m} + (1+\gamma)\epsilon_{\text{sub}} \\ &\leq \gamma^2 \big(\rho v_{i-1} + \epsilon_{\text{sub}}\big) + (1+\gamma)\epsilon_{\text{sub}} \\ &= \gamma\rho \cdot \gamma v_{i-1} + (\gamma^2 + 1 + \gamma)\epsilon_{\text{sub}} \\ &= \gamma\tau v_{i-1} + \widetilde{R}, \end{aligned}$$

where the last equality uses $\widetilde{R} = (1 + \gamma + \gamma^2)\epsilon_{\text{sub}}$. □

Unrolling Lemma B.13(S) from $u_{-1} = 0$ yields a closed-form sequential upper bound $\mathcal{U}_i$; unrolling Lemma B.13(I) provides a closed-form interleaved upper bound $\mathcal{V}_i$ in which a contribution originating from pair k is multiplied by an additional factor γ for each subsequent seam.

Proposition B.14 (Pair-level unrolling)**.** *Under Assumptions B.1, B.2, and B.4 and $\tau < 1$, for every $i \in \{0, \ldots, L-2\}$ the midpoint residuals satisfy $u_i \leq \mathcal{U}_i$ and $v_i \leq \mathcal{V}_i$ with the deterministic upper bounds*

$$\mathcal{U}_i \coloneqq R \sum_{k=0}^{i} \tau^{i-k} = R \frac{1 - \tau^{i+1}}{1 - \tau}, \tag{44}$$

$$\mathcal{V}_i \coloneqq \widetilde{R} \sum_{k=0}^{i} \tau^{i-k} \gamma^{s(i)-s(k)} \leq \widetilde{R} \frac{1 - \tau^{i+1}}{1 - \tau}. \tag{45}$$

Moreover the two bounds admit the direct comparison

$$\mathcal{V}_i \leq \gamma^{s(i)} \mathcal{U}_i + \big(\widetilde{R} - \gamma^{s(i)} R\big) \frac{1 - \tau^{i+1}}{1 - \tau}, \tag{46}$$

where the factor $\gamma^{s(i)}$ quantifies the cumulative seam-amortized gain in the propagated term and $(\widetilde{R} - \gamma^{s(i)} R) \geq 0$ upper-bounds the additive terms not carrying this full factor.

Proof. Proof of Eq. (44). From $u_{-1} = 0$, Lemma B.13(S) gives, by a one-step induction, the base $u_0 \le \tau \cdot 0 + R = R = \mathcal{U}_0$ and the step $u_i \le \tau u_{i-1} + R \le \tau \mathcal{U}_{i-1} + R = R\big(\tau \sum_{k=0}^{i-1} \tau^{i-1-k} + 1\big) = R \sum_{k=0}^{i} \tau^{i-k} = \mathcal{U}_i$.

Proof of Eq. (45). Define, for $\ell \in \{0, \dots, L-2\}$,

$$\alpha_\ell \coloneqq \begin{cases} \tau & \text{if } \ell \text{ is non-seam,} \\ \gamma\tau & \text{if } \ell = cK-1 \text{ is seam,} \end{cases} \quad \beta_\ell \coloneqq \begin{cases} R & \text{if } \ell \text{ is non-seam,} \\ \widetilde{R} & \text{if } \ell = cK-1 \text{ is seam.} \end{cases} \tag{47}$$

Both α_ℓ, β_ℓ are nonnegative, and $0 \le \beta_\ell \le \widetilde{R}$ (since $\widetilde{R} = R + \gamma^2 \epsilon_{\text{sub}} \ge R$). With the convention that an empty product is 1, induction from $v_{-1} = 0$ using Lemma B.13(I) gives

$$v_i \le \sum_{k=0}^{i} \beta_k \prod_{\ell=k+1}^{i} \alpha_\ell \le \widetilde{R} \sum_{k=0}^{i} \prod_{\ell=k+1}^{i} \alpha_\ell, \tag{48}$$

because the base case is $v_0 \le \alpha_0 v_{-1} + \beta_0 = \beta_0$, and the induction step follows from

$$v_i \le \alpha_i v_{i-1} + \beta_i \le \sum_{k=0}^{i-1} \beta_k \prod_{\ell=k+1}^{i} \alpha_\ell + \beta_i. \tag{49}$$

The second inequality in Eq. (48) uses $\beta_k \le \widetilde{R}$ and the nonnegativity of the product. The product $\prod_{\ell=k+1}^{i} \alpha_\ell$ picks up a factor τ at every $\ell \in \{k+1, \dots, i\}$ and an *extra* factor γ at every seam in that range. The number of seams in $\{k+1, \dots, i\}$ is $s(i) - s(k)$ by definition of $s(\cdot)$, so

$$\prod_{\ell=k+1}^{i} \alpha_\ell = \tau^{i-k} \gamma^{s(i)-s(k)}. \tag{50}$$

The product identity therefore gives the interleaved closed form

$$v_i \le \mathcal{V}_i = \widetilde{R} \sum_{k=0}^{i} \tau^{i-k} \gamma^{s(i)-s(k)}.$$

Its geometric upper bound follows from $\gamma^{s(i)-s(k)} \le 1$, valid because $s(i) \ge s(k)$ and $\gamma \in (0, 1)$.

Direct comparison. The quantities $\mathcal{U}_i, \mathcal{V}_i$ are deterministic once $\tau, \gamma, R, \widetilde{R}$ are fixed. Comparing their closed forms and pulling the geometric factor $\gamma^{s(i)}$ out of $\mathcal{U}_i$,

$$\mathcal{V}_i - \gamma^{s(i)} \mathcal{U}_i = \sum_{k=0}^{i} \tau^{i-k} \Big[\widetilde{R} \gamma^{s(i)-s(k)} - \gamma^{s(i)} R \Big]. \tag{51}$$

For each term, $\gamma^{s(i)-s(k)} \le 1$ and $\widetilde{R} \ge 0$ imply $\widetilde{R} \gamma^{s(i)-s(k)} \le \widetilde{R}$, so

$$\mathcal{V}_i - \gamma^{s(i)} \mathcal{U}_i \le \sum_{k=0}^{i} \tau^{i-k} \big(\widetilde{R} - \gamma^{s(i)} R \big) = \big(\widetilde{R} - \gamma^{s(i)} R \big) \frac{1 - \tau^{i+1}}{1 - \tau}. \tag{52}$$

Rearranging yields the stated comparison. The coefficient $\widetilde{R} - \gamma^{s(i)} R$ is nonnegative: since $\gamma \in (0, 1)$ and $s(i) \ge 0$, $\gamma^{s(i)} \le 1$, and thus $\gamma^{s(i)} R \le R \le \widetilde{R}$. Hence the comparison is a nontrivial upper bound of $\mathcal{V}_i$, and the two components have the claimed interpretation: $\gamma^{s(i)} \mathcal{U}_i$ is the propagated contribution amortized by $s(i)$ seams, and $(\widetilde{R} - \gamma^{s(i)} R)(1 - \tau^{i+1})/(1 - \tau)$ collects the additive terms that do not carry the full $\gamma^{s(i)}$ factor. □

The bound in Eq. (46) is convenient for matching the sequential envelope $\mathcal{U}_i$, but it overestimates the residual contribution of non-seam pairs by replacing each β_k by the seam-only upper bound $\widetilde{R}$. The next proposition records a seam-aware refinement that splits the contribution of seam and non-seam pairs explicitly; it is used only to justify Remark B.6 and is not invoked elsewhere in the proof.

Proposition B.15 (Seam-aware sharper comparison). *Let $\mathcal{J}_i := \{k \in \{0,\dots,i\} : k = cK-1$ for some $c \in \{1,\dots,S\}\}$ be the set of seam indices up to pair i. Under the hypotheses of Proposition B.14,*

$$v_i \le \sum_{k=0}^{i} \big(R + \gamma^2 \epsilon_{\mathrm{sub}} \mathbf{1}\{k \in \mathcal{J}_i\}\big) \tau^{i-k} \gamma^{s(i)-s(k)}. \tag{53}$$

Consequently,

$$v_i \le \gamma^{s(i)} \mathcal{U}_i + D_i \epsilon_{\mathrm{sub}}, \tag{54}$$

where

$$D_i := (1+\gamma) \sum_{k=0}^{i} \tau^{i-k} \big(\gamma^{s(i)-s(k)} - \gamma^{s(i)}\big) + \gamma^2 \sum_{k \in \mathcal{J}_i} \tau^{i-k} \gamma^{s(i)-s(k)}. \tag{55}$$

In particular, when $S(L,K) = 0$ (single chunk, no seams), $\mathcal{J}_i = \emptyset$ and $s(i) = s(k) = 0$ for every $k \le i$, so $D_i = 0$ and $v_i \le \mathcal{U}_i$ exactly.

Proof. The interleaved forcing coefficient is $\beta_k = R$ at non-seam pairs and $\beta_k = \widetilde{R} = R + \gamma^2 \epsilon_{\mathrm{sub}}$ at seam pairs, that is, $\beta_k = R + \gamma^2 \epsilon_{\mathrm{sub}} \mathbf{1}\{k \in \mathcal{J}_i\}$. Keeping this seam-dependent forcing in the exact variation-of-constants expansion

$$v_i \le \sum_{k=0}^{i} \beta_k \prod_{\ell=k+1}^{i} \alpha_\ell$$

(without invoking $\beta_k \le \widetilde{R}$) and using the same product identity gives the seam-aware closed form. To extract the comparison with the sequential envelope, write

$$v_i - \gamma^{s(i)} \mathcal{U}_i = R \sum_{k=0}^{i} \tau^{i-k} \big(\gamma^{s(i)-s(k)} - \gamma^{s(i)}\big) + \gamma^2 \epsilon_{\mathrm{sub}} \sum_{k \in \mathcal{J}_i} \tau^{i-k} \gamma^{s(i)-s(k)},$$

using $R = (1+\gamma)\epsilon_{\mathrm{sub}}$ and rearranging. Both summands are nonnegative because $s(i) \ge s(k)$ implies $\gamma^{s(i)-s(k)} \ge \gamma^{s(i)}$ (as $\gamma \in (0,1)$). When $S = 0$, $\mathcal{J}_i = \emptyset$ trivially zeros the second sum, and the first sum vanishes term by term because $s(i) - s(k) = 0$. Hence $D_i = 0$ and the bound becomes $v_i \le \gamma^{s(i)} \mathcal{U}_i = \mathcal{U}_i$. □

Corollary B.16 (Last-pair upper bounds). *At $i = L-2$ (so $s(L-2) = S$), $u_{L-2} \le \mathcal{U}_{L-2}$ and $v_{L-2} \le \mathcal{V}_{L-2}$, with*

$$\mathcal{U}_{L-2} = R \frac{1-\tau^{L-1}}{1-\tau}, \tag{56}$$

$$\mathcal{V}_{L-2} \le \gamma^S \mathcal{U}_{L-2} + \big(\widetilde{R} - \gamma^S R\big) \frac{1-\tau^{L-1}}{1-\tau}. \tag{57}$$

Proof. Apply Proposition B.14 to the last valid pair $i = L-2$. It remains only to check that $s(L-2) = S$. Recall that the seam count is $S = \lceil L/K \rceil - 1$ and that the seams are indexed by $c \in \{1,\dots,S\}$ with seam index $cK-1$. If $S = 0$, there are no seams and $s(L-2) = 0 = S$. Assume henceforth $S \ge 1$. Write $L = qK + r$ with $q = \lfloor L/K \rfloor$ and $0 \le r < K$. Two cases:

- If $r = 0$, then $\lceil L/K \rceil = q$, so $S = q-1 \ge 1$, i.e. $q \ge 2$, and $SK = (q-1)K = L-K \le L-1$ because $K \ge 1$.
- If $r \ge 1$, then $\lceil L/K \rceil = q+1$, so $S = q \ge 1$ and $SK = qK = L-r \le L-1$ because $r \ge 1$.

In both cases $SK \le L-1$. Hence for every $c \in \{1,\dots,S\}$, the seam index satisfies $cK-1 \le SK-1 \le L-2$, and is therefore counted by $s(L-2)$. Conversely, by the definition of S, there are no seams with $c > S$, so no seam index exceeds $SK-1 \le L-2$. We conclude $s(L-2) = S$. □

B.8 From last-pair residual to depth-L mismatch

We propagate the last-pair midpoint residual through the final block $L-1$ in two formal steps: a monotonicity-of-final-visit-times lemma (Lemma B.17) and a final-prefix-invariance lemma (Lemma B.18). For every $b \in \{0, \dots, L-2\}$, let T_b denote the algorithmic time of the *last* CBQ pass on pair $(b, b+1)$: in Sequential CBQ this is the unique pass on that pair; in ICBQ it is the unique pass if $(b, b+1)$ is a non-seam pair, and the second pass if $(b, b+1)$ is a seam pair (i.e. $b = cK-1$ for some $c \in \{1, \dots, S\}$). We extend by $T_{L-1} := T_{L-2}$, so that block $L-1$'s final state is pinned at T_{L-2}. For each $b \in \{0, \dots, L-1\}$, write θ_b^{post} for the parameter state of block b at time T_b.

Lemma B.17 (Monotonicity of final-visit times and block finalization)**.** *Under the schedules described by Algorithms 1–2,*

(a) *the times satisfy $T_0 < T_1 < \cdots < T_{L-2} = T_{L-1}$;*

(b) *for every $b \in \{0, \dots, L-1\}$, no operation that modifies θ_b occurs strictly after T_b; consequently, block b's state at every time $t \geq T_b$ equals θ_b^{post}, and $\theta_b^{\mathrm{post}} \in \Theta_b$ by Assumption B.2.*

Proof. (a) The outer driver iterates $b = 0, 1, \dots, L-1$ in increasing order, and the inner pair loop processes pairs in increasing order of the first-block index within each chunk (Algorithms 1–2). For non-seam pairs, the unique CBQ pass occurs inside the chunk-closing inner refinement at some outer iteration b', and its algorithmic time is strictly increasing in the pair index b (later pairs are processed later inside the same pair loop, and later chunks close later than earlier chunks). For a seam pair $(cK-1, cK)$, the second pass occurs at chunk $c+1$'s closure, which is strictly later than chunk c's closure (where the first pass occurs); the second pass on the seam is itself the first pair processed in chunk $c+1$'s pair loop, so its algorithmic time precedes every other CBQ pass in chunk $c+1$. Combining these observations, $T_0 < T_1 < \cdots < T_{L-2}$. The convention $T_{L-1} = T_{L-2}$ is by definition.

(b) The operations that may modify θ_b are

(A1) the outer-driver prefit/inner-quantizer step on block b, which occurs exactly once at outer iteration b; call its algorithmic time Q_b;

(A2) CBQ updates arising from pairs that include b, namely $(b-1, b)$ (defined only for $b \geq 1$) and $(b, b+1)$ (defined only for $b \leq L-2$); re-rolls recompute activations only and modify no parameter.

For type (A2): the last CBQ pass on $(b, b+1)$ is T_b by definition (for $b \leq L-2$); the last CBQ pass on $(b-1, b)$ is T_{b-1} (for $b \geq 1$). By (a), $T_{b-1} < T_b$ for $1 \leq b \leq L-2$, and $T_{L-2} = T_{L-1}$ for $b = L-1$. Hence no type-(A2) operation occurs strictly after T_b.

For type (A1): Q_b occurs at the end of outer iteration b, which strictly precedes every chunk-closing inner refinement triggered at outer iteration $b' \geq b$. If $b \leq L-2$, the first CBQ pass on pair $(b, b+1)$ occurs inside the chunk closing at some $b' \geq b$, hence strictly after Q_b; in particular $Q_b < T_b$. If $b = L-1$, then $T_{L-1} = T_{L-2}$ lies within the chunk closing at outer iteration $L-1$ (the final chunk), which is strictly after Q_{L-1}, the quantizer step of outer iteration $L-1$; hence $Q_{L-1} < T_{L-1}$.

Combining the two cases, no operation modifying θ_b occurs strictly after T_b, so block b's state at every $t \geq T_b$ equals θ_b^{post}. Admissibility $\theta_b^{\mathrm{post}} \in \Theta_b$ follows from the definition preceding Assumption B.2. □

Lemma B.18 (Final-prefix identity for $\hat{X}_{L-1}$)**.** *Under Assumptions B.1, B.2, and B.4 and the schedules described in Algorithms 1–2, for every $b \in \{0, \dots, L-2\}$ the student activation at depth $b+1$ at time T_b (immediately after the advance step following that CBQ pass) satisfies*

$$\hat{X}_{b+1}^{[T_b]} = \big(f_b^{\theta_b^{\mathrm{post}}} \circ \cdots \circ f_0^{\theta_0^{\mathrm{post}}}\big)(X_0). \tag{58}$$

In particular, $\hat{X}_{L-1} = (f_{L-2}^{\theta_{L-2}^{\mathrm{post}}} \circ \cdots \circ f_0^{\theta_0^{\mathrm{post}}})(X_0)$, and this value is preserved by any subsequent activation-only re-roll (no parameter is modified after T_{L-1}).

Proof. We argue by induction on $b \in \{0, \dots, L-2\}$.

Base case $b = 0$. By Assumption B.4 applied with $d = 0$ (empty composition), $\hat{X}_0 = X_0$, and the advance at T_0 gives $\hat{X}_1^{[T_0]} = f_0^{\theta_0^{\mathrm{post}}}(\hat{X}_0) = f_0^{\theta_0^{\mathrm{post}}}(X_0)$, which is Eq. (58) at $b = 0$.

Inductive step. Assume Eq. (58) holds at T_{b-1} (with $b \geq 1$). In the interval $(T_{b-1}, T_b]$ the algorithm executes some combination of outer-driver block-level steps on indices $b' \geq b$ (the blocks whose outer-driver quantization has not yet taken place by time T_{b-1}), end-of-chunk and in-pass re-rolls, and CBQ passes on pairs $(i, i{+}1)$ with $i \geq b$ (since the last CBQ pass on any pair $(i, i{+}1)$ with $i < b$ is $T_i \leq T_{b-1}$, the right endpoint of the preceding interval). By Lemma B.17, $\theta_{b'}$ for every $b' \leq b{-}1$ remains pinned at its final state $\theta_{b'}^{\mathrm{post}}$ throughout this interval.

We now track $\hat{X}_b$ during $(T_{b-1}, T_b]$. The only operations that can modify $\hat{X}_b$ are:

(i) re-rolls (which recompute the student prefix up to some depth $\geq b$);

(ii) the advance step at line 5 of Algorithm 2 following a CBQ pass on pair $(i, i{+}1)$ with $i{+}1 = b$, i.e. pair $(b{-}1, b)$.

By Lemma B.17, no CBQ pass on pair $(b{-}1, b)$ occurs strictly after T_{b-1}, so advance steps of type (ii) do not occur in the open interval $(T_{b-1}, T_b]$. Hence $\hat{X}_b$ is modified only by re-rolls in this interval. By Assumption B.4, any re-roll in the interval produces $\hat{X}_b = (f_{b-1}^{\theta_{b-1}^{\mathrm{current}}} \circ \cdots \circ f_0^{\theta_0^{\mathrm{current}}})(X_0)$; since $\theta_{b'}^{\mathrm{current}} = \theta_{b'}^{\mathrm{post}}$ for every $b' \leq b{-}1$ throughout the interval, this equals the *final-prefix* value $P_b \coloneqq (f_{b-1}^{\theta_{b-1}^{\mathrm{post}}} \circ \cdots \circ f_0^{\theta_0^{\mathrm{post}}})(X_0)$. The inductive hypothesis at depth $b{-}1$ states $\hat{X}_b^{[T_{b-1}]} = P_b$, so at the left endpoint of the interval $\hat{X}_b$ is already at the final-prefix value; between re-rolls, $\hat{X}_b$ is stored and untouched by any other operation, hence frozen at P_b; every re-roll resets it to P_b. Consequently, $\hat{X}_b^{[t]} = P_b$ for every $t \in (T_{b-1}, T_b]$.

At time T_b, the CBQ pass on pair $(b, b{+}1)$ sees this $\hat{X}_b$ as its input; immediately after the pass, the advance gives $\hat{X}_{b+1}^{[T_b]} = f_b^{\theta_b^{\mathrm{post}}}(\hat{X}_b) = (f_b^{\theta_b^{\mathrm{post}}} \circ \cdots \circ f_0^{\theta_0^{\mathrm{post}}})(X_0)$, which is Eq. (58) at b.

Specialising $b = L{-}2$ yields $\hat{X}_{L-1} = (f_{L-2}^{\theta_{L-2}^{\mathrm{post}}} \circ \cdots \circ f_0^{\theta_0^{\mathrm{post}}})(X_0)$. By Lemma B.17(b), no parameter is modified after $T_{L-1} = T_{L-2}$; any later activation-only re-roll therefore recomputes $\hat{X}_{L-1}$ to the same final-prefix value, leaving it unchanged. □

Proof of Theorem B.5. Pathwise depth-L bound. By Lemma B.18,

$$\hat{X}_{L-1} = \big(f_{L-2}^{\theta_{L-2}^{\mathrm{post}}} \circ \cdots \circ f_0^{\theta_0^{\mathrm{post}}}\big)(X_0) \tag{59}$$

holds at the moment immediately after the CBQ pass at time T_{L-2}, and it is preserved by any subsequent activation-only re-roll. In particular, after the final chunk closes, the end-of-chunk re-roll (or equivalently an explicit final forward pass through the frozen final block) computes $\hat{X}_L = f_{L-1}^{\theta_{L-1}^{\mathrm{post}}}(\hat{X}_{L-1})$. This is an exact application of the final block map by Assumption B.4; no later operation modifies θ_{L-1} by Lemma B.17(b) with $b = L{-}1$.

Hence $\hat{X}_{L-1} \in \mathcal{T}_{L-1}$ by Assumption B.1, $X_{L-1}^\star \in \mathcal{T}_{L-1}$ by construction, and $\|\hat{X}_{L-1} - X_{L-1}^\star\|_F = m_{L-2}^{\mathrm{out}}$ by definition in Eq. (29). By the triangle inequality followed by the local envelope in Eq. (18) (applied with $\theta_{L-1} = \theta_{L-1}^{\mathrm{post}} \in \Theta_{L-1}$ and $z = \hat{X}_{L-1} \in \mathcal{T}_{L-1}$) and Assumption B.1,

$$\begin{aligned} \|e_L\|_F &= \|f_{L-1}^{\theta_{L-1}^{\mathrm{post}}}(\hat{X}_{L-1}) - f_{L-1}^\star(X_{L-1}^\star)\|_F \\ &\leq \|f_{L-1}^{\theta_{L-1}^{\mathrm{post}}}(\hat{X}_{L-1}) - f_{L-1}^\star(\hat{X}_{L-1})\|_F + \|f_{L-1}^\star(\hat{X}_{L-1}) - f_{L-1}^\star(X_{L-1}^\star)\|_F \\ &\leq \epsilon_{\mathrm{sub}} + \rho\|\hat{X}_{L-1} - X_{L-1}^\star\|_F = \rho m_{L-2}^{\mathrm{out}} + \epsilon_{\mathrm{sub}}. \end{aligned} \tag{60}$$

This inequality is pathwise for each calibration sample.

Sequential bound. For the sequential schedule, apply Corollary B.16: the pathwise bound $u_{L-2}^{\mathrm{out},(n)} \leq \mathcal{U}_{L-2}$ holds for every calibration sample $n \in \{1, \ldots, N\}$; the right-hand side $\mathcal{U}_{L-2}$ is sample-independent, so this is a uniform pathwise bound. The last-layer Lipschitz inequality then gives,

pathwise on each sample, $\|e_L^{\mathrm{seq},(n)}\|_F \le \rho\mathcal{U}_{L-2}+\epsilon_{\mathrm{sub}}$ for every n. Taking the calibration expectation $\mathbb{E}[\,\cdot\,] = \frac{1}{N}\sum_{n=1}^{N}(\cdot)^{(n)}$ preserves the inequality because the right-hand side is sample-independent, and the sum $\rho\mathcal{U}_{L-2}+\epsilon_{\mathrm{sub}}$ equals $\mathcal{B}_L^{\mathrm{seq}}$ from Eq. (22):

$$\mathbb{E}\|e_L^{\mathrm{seq}}\|_F \le \rho\mathcal{U}_{L-2}+\epsilon_{\mathrm{sub}} = \mathcal{B}_L^{\mathrm{seq}}. \tag{61}$$

Interleaved bound. For ICBQ, apply Eq. (60) pathwise to each sample n, then use the pathwise bound $v_{L-2}^{\mathrm{out},(n)} \le \mathcal{V}_{L-2}$ from Corollary B.16 (with $\mathcal{V}_{L-2}$ sample-independent), and finally apply the deterministic bound-level comparison in Eq. (57):

$$\|e_L^{\mathrm{int},(n)}\|_F \le \rho\, v_{L-2}^{\mathrm{out},(n)} + \epsilon_{\mathrm{sub}} \le \rho\,\mathcal{V}_{L-2} + \epsilon_{\mathrm{sub}} \le \rho\left[\gamma^S\mathcal{U}_{L-2} + \big(\widetilde{R}-\gamma^S R\big)\frac{1-\tau^{L-1}}{1-\tau}\right] + \epsilon_{\mathrm{sub}}.$$

Using the identity $\epsilon_{\mathrm{sub}} = \gamma^S\epsilon_{\mathrm{sub}} + (1-\gamma^S)\epsilon_{\mathrm{sub}}$, reshuffling terms by purely algebraic rearrangement, and taking calibration expectations (all right-hand sides are sample-independent, so the inequality is preserved),

$$\begin{aligned}
\mathbb{E}\|e_L^{\mathrm{int}}\|_F &\le \rho\,\mathcal{V}_{L-2} + \epsilon_{\mathrm{sub}} \\
&\le \rho\Big[\gamma^S\mathcal{U}_{L-2} + \big(\widetilde{R}-\gamma^S R\big)\frac{1-\tau^{L-1}}{1-\tau}\Big] + \epsilon_{\mathrm{sub}} \\
&= \gamma^S\big[\rho\mathcal{U}_{L-2}+\epsilon_{\mathrm{sub}}\big] + \big(\widetilde{R}-\gamma^S R\big)\frac{\rho\,(1-\tau^{L-1})}{1-\tau} + (1-\gamma^S)\,\epsilon_{\mathrm{sub}} \\
&= \gamma^S\,\mathcal{B}_L^{\mathrm{seq}} + C(L,K,\rho,\gamma)\,\epsilon_{\mathrm{sub}},
\end{aligned} \tag{62}$$

where the last equality uses the definition in Eq. (22) of $\mathcal{B}_L^{\mathrm{seq}}$ together with the identification

$$C(L,K,\rho,\gamma)\,\epsilon_{\mathrm{sub}} = \big(\widetilde{R}-\gamma^S R\big)\frac{\rho\,(1-\tau^{L-1})}{1-\tau} + (1-\gamma^S)\,\epsilon_{\mathrm{sub}}. \tag{63}$$

Expanding the residual coefficient $\widetilde{R}-\gamma^S R = \big[(1+\gamma+\gamma^2)-\gamma^S(1+\gamma)\big]\epsilon_{\mathrm{sub}}$ yields the admissible residual constant whenever $\epsilon_{\mathrm{sub}} > 0$. If $\epsilon_{\mathrm{sub}} = 0$, then $R = \widetilde{R} = 0$ and $\mathcal{B}_L^{\mathrm{seq}} = 0$, so the preceding pathwise recurrences give $\mathbb{E}\|e_L^{\mathrm{int}}\|_F = 0$; choosing the same displayed value of $C(L,K,\rho,\gamma)$ leaves $C(L,K,\rho,\gamma)\,\epsilon_{\mathrm{sub}} = 0$. Hence Eq. (24) is valid in all cases.

Depth-uniform residual constant. Using $\gamma^S \ge 0$,

$$\widetilde{R}-\gamma^S R = \big[(1+\gamma+\gamma^2)-\gamma^S(1+\gamma)\big]\epsilon_{\mathrm{sub}} \le (1+\gamma+\gamma^2)\,\epsilon_{\mathrm{sub}}, \tag{64}$$

together with $(1-\gamma^S) \le 1$ and $(1-\tau^{L-1})/(1-\tau) \le 1/(1-\tau)$ (from $\tau < 1$), we obtain the depth-uniform bound

$$C(L,K,\rho,\gamma) \le (1-\gamma^S) + (1+\gamma+\gamma^2)\frac{\rho(1-\tau^{L-1})}{1-\tau} \le 1 + \frac{(1+\gamma+\gamma^2)\rho}{1-\tau}, \tag{65}$$

independent of L and K. Under the standard stability regime $\rho \le 1$ this simplifies to $C \le 1 + (1+\gamma+\gamma^2)/(1-\tau)$, as stated in Eq. (25); the near-divergent case $\rho > 1$ is handled in Appendix B.10 under the same sufficient condition $\rho < 1/\gamma$ (equivalently $\tau < 1$). Eq. (62) is exactly the statement of Theorem B.5. □

Remark B.19 (On the interpretation of the upper bound)**.** *The derivation above produces the ICBQ bound $\mathcal{B}_L^{\mathrm{int}} := \gamma^S\mathcal{B}_L^{\mathrm{seq}} + C\epsilon_{\mathrm{sub}}$ and the sequential bound $\mathcal{B}_L^{\mathrm{seq}}$ in terms of the same quantities $\rho,\gamma,\epsilon_{\mathrm{sub}},\tau$. The factor γ^S is attached only to the propagated component of the bound; the additive residual terms are collected in $C\epsilon_{\mathrm{sub}}$ and can dominate when the sequential bound is already small. The actual realized sequential residual $\mathbb{E}\|e_L^{\mathrm{seq}}\|_F$ may be strictly smaller than $\mathcal{B}_L^{\mathrm{seq}}$; thus, Theorem B.5 is a* bound-to-bound *comparison rather than a realized-residual comparison, which is how Corollary B.8's amortization-gain prediction should be read.*

B.9 RECURRENCE-LEVEL SHARPNESS

Proposition B.20 (Scalar-recurrence saturation and seam-exponent sharpness)**.** *Fix any $L \geq 2$, any chunk length $K \geq 1$ (which determines the seam set $\mathcal{J}_K = \{cK-1 : c = 1, \dots, S\}$), and any constants $R, \widetilde{R} \geq 0$, $\tau \in [0,1)$, $\gamma \in (0,1]$. Define the* saturating *scalar recurrences obtained by replacing the inequalities in Lemma B.13 by equalities:*

$$u_i^* = \tau\, u_{i-1}^* + R, \;\; v_i^* = \begin{cases} \tau\, v_{i-1}^* + R, & i \notin \mathcal{J}_K, \\ \gamma\tau\, v_{i-1}^* + \widetilde{R}, & i \in \mathcal{J}_K, \end{cases} \quad u_{-1}^* = v_{-1}^* = 0. \tag{66}$$

Then:

(a) (Envelopes.) *Any pathwise sequence (u_i, v_i) with $u_i, v_i \geq 0$ that satisfies the inequalities in Lemma B.13 with the boundary condition $u_{-1} = v_{-1} = 0$ obeys $u_i \leq u_i^*$ and $v_i \leq v_i^*$ for every $i \geq -1$.*

(b) (Closed form.) *With*

$$\alpha_\ell^* = \begin{cases} \tau, & \ell \notin \mathcal{J}_K, \\ \gamma\tau, & \ell \in \mathcal{J}_K, \end{cases} \quad \beta_\ell^* = \begin{cases} R, & \ell \notin \mathcal{J}_K, \\ \widetilde{R}, & \ell \in \mathcal{J}_K, \end{cases}$$

the saturated sequences satisfy

$$u_i^* = R \sum_{k=0}^{i} \tau^{i-k}, \;\; v_i^* = \sum_{k=0}^{i} \beta_k^* \prod_{\ell=k+1}^{i} \alpha_\ell^* = \sum_{k=0}^{i} \beta_k^* \tau^{i-k} \gamma^{s(i)-s(k)}. \tag{67}$$

(c) (Sharpness of the seam exponent.) *For every $0 \leq k \leq i \leq L-2$, the linear response at index i to an additive impulse injected at index k is exactly*

$$\prod_{\ell=k+1}^{i} \alpha_\ell^* = \tau^{i-k} \gamma^{s(i)-s(k)}. \tag{68}$$

Consequently, for $\gamma \in (0,1)$, no recurrence-level argument that uses only Lemma B.13's multipliers can uniformly replace the exponent $s(i) - s(k)$ by any larger exponent; doing so would underbound this exact impulse response. For $\gamma = 1$ the statement is vacuous because seams add no contraction.

Proof. (a) We argue $u_i \leq u_i^*$ and $v_i \leq v_i^*$ by simultaneous induction on $i \geq -1$. The base case $i = -1$ is immediate from $u_{-1} = v_{-1} = 0 = u_{-1}^* = v_{-1}^*$. For the inductive step, fix $i \geq 0$ and assume $u_{i-1} \leq u_{i-1}^*$ and $v_{i-1} \leq v_{i-1}^*$. The maps $x \mapsto \tau x + R$ and $x \mapsto \gamma\tau x + \widetilde{R}$ are monotone nondecreasing on $[0, \infty)$. Applying the appropriate map gives $u_i \leq \tau u_{i-1} + R \leq \tau u_{i-1}^* + R = u_i^*$ and, analogously, $v_i \leq v_i^*$, splitting cases according to whether $i \in \mathcal{J}_K$.

(b) Unrolling the saturating scalar recurrences gives the displayed formula for u_i^*. The formula for v_i^* is the standard variation-of-constants expansion for a non-autonomous linear recursion. The product contains one factor τ for each index in $\{k+1, \dots, i\}$ and one additional factor γ for each seam in that same index range. The number of such seams is $s(i) - s(k)$, giving Eq. (67).

(c) Fix $0 \leq k \leq i \leq L-2$ and add a nonnegative impulse a to the forcing term at index k, leaving all other forcing terms and all multipliers unchanged. Let Δ_j be the difference between the perturbed and unperturbed saturated sequences. Then $\Delta_j = 0$ for $j < k$, $\Delta_k = a$, and $\Delta_j = \alpha_j^* \Delta_{j-1}$ for every $j > k$. Hence

$$\Delta_i = a \prod_{\ell=k+1}^{i} \alpha_\ell^* = a\tau^{i-k} \gamma^{s(i)-s(k)}.$$

This equality is an exact linear response of the recurrence. If, for some $\eta > 0$ and $\gamma \in (0,1)$, one attempted to replace the factor $\gamma^{s(i)-s(k)}$ by the smaller factor $\gamma^{s(i)-s(k)+\eta}$ uniformly over nonnegative impulses, the bound would fail for this perturbation because $a\tau^{i-k}\gamma^{s(i)-s(k)} > a\tau^{i-k}\gamma^{s(i)-s(k)+\eta}$ whenever $a > 0$ and $\tau > 0$; the case $\tau = 0$ is degenerate: the response is zero for $i > k$, while for $i = k$ the empty product is one and the same contradiction applies. Thus the seam exponent in Eq. (68) is exact at the recurrence level. □

The theorem's $C\,\epsilon_{\mathrm{sub}}$ residual is the price of keeping these additive contributions explicit at the depth-L bound.

B.10 THE $\rho > 1$ (NEAR-DIVERGENT) REGIME

The per-pair recursion in Eq. (30) requires only $\tau = \gamma\rho < 1$ for $C(L, K, \rho, \gamma)$ to be uniformly bounded in depth. When the fp16 blocks are expansive on parts of the empirical calibration trajectories ($\rho > 1$), ICBQ remains stable provided $\gamma\rho < 1$, while the sequential recursion (which has $u_i \le \gamma\rho u_{i-1} + (1+\gamma)\epsilon_{\mathrm{sub}}$ too, cf. Lemma B.9) requires the same condition. The *advantage* of ICBQ in this regime is the same bound-level advantage as in Theorem B.5: the propagated part carries γ^S, while the additive residual remains uniformly bounded as long as $\tau < 1$. Stated explicitly:

Proposition B.21 (Stability gain)**.** *Under Assumptions B.1, B.2, and B.4 with the locally expansive allowance $1 < \rho < 1/\gamma$, both the sequential and interleaved upper bounds are finite uniformly in depth. Moreover, if $C(L, K, \rho, \gamma)\epsilon_{\mathrm{sub}} \le \delta\gamma^{S(L,K)}\mathcal{B}_L^{\mathrm{seq}}$ for some $\delta \ge 0$ and $\mathcal{B}_L^{\mathrm{seq}} > 0$, then the ratio of the sequential upper bound to the ICBQ upper bound is at least $\gamma^{-S(L,K)}/(1+\delta)$.*

Proof. Since $\gamma \in (0, 1)$, the condition $1 < \rho < 1/\gamma$ is equivalent to $\tau = \gamma\rho < 1$ with locally expansive teacher blocks. Therefore $(1-\tau^{L-1})/(1-\tau) \le 1/(1-\tau)$ uniformly in L, and

$$\mathcal{B}_L^{\mathrm{seq}} = \rho(1+\gamma)\frac{1-\tau^{L-1}}{1-\tau}\epsilon_{\mathrm{sub}} + \epsilon_{\mathrm{sub}} \le \left(\frac{\rho(1+\gamma)}{1-\tau} + 1\right)\epsilon_{\mathrm{sub}}, \tag{69}$$

which is independent of L. Theorem B.5 together with Eq. (25) gives

$$\mathcal{B}_L^{\mathrm{int}} := \gamma^{S(L,K)}\mathcal{B}_L^{\mathrm{seq}} + C(L, K, \rho, \gamma)\epsilon_{\mathrm{sub}} \le \mathcal{B}_L^{\mathrm{seq}} + \left(1 + \frac{(1+\gamma+\gamma^2)\rho}{1-\tau}\right)\epsilon_{\mathrm{sub}}, \tag{70}$$

where we used $\gamma^{S(L,K)} \le 1$ and the depth-uniform bound in Eq. (25) on C. Hence the interleaved upper bound is also finite uniformly in L. For the ratio claim, suppose the displayed residual condition holds. Then from Eq. (23) and the definition of $\mathcal{B}_L^{\mathrm{int}}$ in Corollary B.8,

$$\mathcal{B}_L^{\mathrm{int}} = \gamma^{S(L,K)}\mathcal{B}_L^{\mathrm{seq}} + C(L, K, \rho, \gamma)\epsilon_{\mathrm{sub}} \le (1+\delta)\gamma^{S(L,K)}\mathcal{B}_L^{\mathrm{seq}}. \tag{71}$$

Since $C(L, K, \rho, \gamma)\epsilon_{\mathrm{sub}} \ge 0$, $\mathcal{B}_L^{\mathrm{seq}} > 0$, and $\gamma^{S(L,K)} > 0$, we also have $\mathcal{B}_L^{\mathrm{int}} \ge \gamma^{S(L,K)}\mathcal{B}_L^{\mathrm{seq}} > 0$, so both sides of the displayed inequality are strictly positive. Dividing by the positive quantity $\mathcal{B}_L^{\mathrm{int}} \cdot (1+\delta)\gamma^{S(L,K)}$ and rearranging,

$$\frac{\mathcal{B}_L^{\mathrm{seq}}}{\mathcal{B}_L^{\mathrm{int}}} \ge \frac{1}{(1+\delta)\gamma^{S(L,K)}} = \frac{\gamma^{-S(L,K)}}{1+\delta}, \tag{72}$$

which is the claimed ratio bound. □

The main experiments exhibit a related qualitative pattern: Sequential CBQ produces PPL $> 10^4$ for Qwen3-8B in one reported setting, whereas ICBQ yields PPL below 50 for both Qwen3-8B and Llama-3-8B. These observations are consistent with the schedule mechanism but do not estimate the theorem's uniform constants or validate the bound quantitatively.

B.11 EMPIRICAL MEASUREMENT OF γ

We diagnose the per-pair contraction empirically on Llama-2-7B by recording, after each of the $32-1 = 31$ two-block refinement calls with ternary DBF, the square-root post-over-pre MSE ratio. The corresponding histogram has mean $\bar{\gamma} \approx 0.67$ and maximum $\bar{\gamma}_{\max} \le 0.92$.

Uniform vs. aggregate contraction. The quantity $\bar{\gamma}$ reported here is an aggregate diagnostic: it is the average square-root post-over-pre MSE ratio over the CBQ windows actually executed on the calibration set, hence it summarizes how the implemented optimizer behaves *in aggregate*. It is *not*, by itself, the uniform per-sample contraction constant required by Assumption B.2. A theorem-level certificate would instead require a uniform estimate such as

$$\widehat{\gamma}_{\mathrm{unif}}(\epsilon_{\mathrm{opt}}) = \max_{\substack{j,n:\\ r_{j,\mathrm{in}}^{(n)} > 0}} \frac{\left(r_{j,\mathrm{out}}^{(n)} - \epsilon_{\mathrm{opt}}\right)_+}{r_{j,\mathrm{in}}^{(n)}}, \tag{73}$$

where j ranges over the actual CBQ calls of the schedule and n over the calibration samples in each call's batch, and $r^{(n)}_{j,\mathrm{in}}, r^{(n)}_{j,\mathrm{out}}$ are the pre- and post-pass midpoint residuals of call j on sample n. With this convention, $\widehat{\gamma}_{\mathrm{unif}}(\epsilon_{\mathrm{opt}})$ is the smallest γ that makes Assumption B.2 hold uniformly on the actually visited finite sets, given a chosen fixed-point radius ϵ_{opt}.

Reading of Theorem B.5. Using $\bar{\gamma}$ or $\bar{\gamma}_{\max}$ in Corollary B.8 therefore produces only a *schedule-level diagnostic* of the boundary-overlap mechanism, not a formal certificate. With $L = 32$ and $K = 4$ (so $S = 7$), the aggregate-level diagnostics give an amortization-gain proxy of $\bar{\gamma}^{-7} \approx 16\times$ at the mean and $\bar{\gamma}_{\max}^{-7} \approx 1.8\times$ at the aggregate maximum. The empirical per-model range 1.08–$14.5\times$ reported in the main experiments sits between these proxies, consistent with the formal γ in Theorem B.5 being some uniform value between $\bar{\gamma}$ and $\bar{\gamma}_{\max}$. Establishing $\widehat{\gamma}_{\mathrm{unif}}(\epsilon_{\mathrm{opt}})$ at scale would require logging per-sample midpoint residuals at every CBQ call and is left to future work; the present empirical evidence supports the qualitative validity of the schedule mechanism rather than a quantitative theorem-level certificate.

B.12 Numerical simulation

A standalone `numpy` script reproducing the rate of Theorem B.5 on a 64-block scalar recurrence toy is included in the supplementary material; running it prints a table of $E_c^{\mathrm{int}}/E_c^{\mathrm{seq}}$ vs. c that matches γ^c to four significant digits over $c \in \{1, \ldots, 15\}$.

C Experimental and Implementation Details

C.1 Experimental setup details

Base models and references. The seven base models used in the main experiments are Llama-3.2-3B (Meta Llama Team, 2024), Llama-2-7B (Touvron et al., 2023), Mistral-7B (Jiang et al., 2023), Qwen3-8B (Yang et al., 2025), Llama-3-8B (Grattafiori et al., 2024), Llama-2-13B (Touvron et al., 2023), and Qwen3-14B (Yang et al., 2025).

Quantization and optimization settings. The ternary DBF runs use progressive prefit (50 steps, float32, AdamW). For models below 7B, we use `nsamples` $= 128$ and `dbf.iters` $= 200$; for models at 7B and above, we use `nsamples` $= 256$ and `dbf.iters` $= 400$. GPTQ runs use W3 and W2g128 with prefit disabled, GPTQ block size 128, and Hessian damping 0.01. The integer-bit comparison uses the same weight-only regime as the ternary DBF table, with the inner quantizer replaced by GPTQ. No adapter/LoRA parameters are introduced.

Evaluation protocol and benchmarks. Perplexity is measured with the `lm-evaluation-harness` protocol (Gao et al., 2021). Zero-shot evaluation uses BoolQ, PIQA, HellaSwag, WinoGrande, ARC-e, ARC-c, and OpenBookQA (Clark et al., 2019; Bisk et al., 2020; Zellers et al., 2019; Sakaguchi et al., 2021; Clark et al., 2018; Mihaylov et al., 2018). Main-text tables focus on C4-calibrated reporting, while supplementary tables include additional calibration views where relevant.

C.2 Calibration bias

WikiText-2 calibration on base models with non-standard attention, such as Qwen3 with grouped-query attention (Yang et al., 2025) or Gemma-3 with local sliding-window attention (Gemma Team et al., 2025), can cover a narrow distribution of positional indices. Under ternary DBF, this restricted calibration distribution can produce a mismatch that the local CBQ loss does not fully capture. In our setting, C4 calibration at the same sample count gave more stable results on the deeper models, making it the default for the main ternary DBF results.

C.3 Implementation and hyperparameters

Hardware. All runs on $1\times$ NVIDIA H100-80GB. Memory peak with ternary DBF: 32 GB for 7B, 55 GB for 13B models.

Peak memory: ICBQ vs. CBQ. The same practical memory regime is observed under Sequential CBQ and ICBQ: in our implementation, the extra state introduced by ICBQ over a single-sweep CBQ is the passive re-roll buffer `original_inps`, whose size is set by calibration workload and is independent of depth-related CBQ scheduling variables (e.g., L and windowing order for fixed workload). Instantiating the stored-input buffer size at our operating points gives:

Model	d_{hidden}	N	T	dtype	$\Delta\text{mem}_{\text{ICBQ}}$
TinyLlama-1.1B	2048	128	2048	fp16	≈ 1.0 GB
Llama-2-7B / Llama-3-8B / Qwen3-8B / Mistral-7B	4096	256	2048	fp16	≈ 4.3 GB
Llama-2-13B / Qwen3-14B	5120	256	2048	fp16	≈ 5.2 GB

The remaining models use smaller hidden widths ($d_{\text{hidden}} \in \{1024, 1152, 2304\}$), which gives proportionally smaller ICBQ overheads (approximately 0.5, 0.6, and 1.2 GB at $N = 128$, $T = 2048$, fp16). In these cases, the overhead is below 15% of the ternary DBF peak quoted above. Because ICBQ never materializes two active two-block windows simultaneously, the practical peak does not enter a new regime when switching from Sequential CBQ to ICBQ. Measured run logs show similar peak GPU memory across paired settings, consistent with the L, K-independent peak-memory bound up to implementation-level allocator effects.

Wall-clock time: ICBQ vs. CBQ (C4 calibration). To complement the peak-memory comparison, Table 5 reports end-to-end wall-clock runtime for the C4-calibrated ternary DBF runs, comparing

Sequential CBQ to the full ICBQ schedule with seam revisit on eleven models. The mean wall-clock factor is $1.21\times$, which gives an upper-end estimate of the runtime cost of the full schedule. Disabling seam revisit would remove the extra pair updates and shrink this gap, but interleaving and re-roll overhead would remain.

Table 5: **Wall-clock comparison on C4-calibrated ternary DBF runs (11 models).** Sequential CBQ versus ICBQ (ours), using Slurm elapsed time. Ratio > 1 indicates ICBQ takes longer.

Model	Seq. CBQ elapsed	ICBQ elapsed	ICBQ/Seq.	Delta
TinyLlama-1.1B	01:20:53	01:27:51	$1.09\times$	$+7.0$ min
Qwen3-0.6B	01:13:18	01:27:42	$1.20\times$	$+14.4$ min
Gemma-2-2B	02:11:44	02:43:17	$1.24\times$	$+31.6$ min
Gemma-3-1B	01:05:13	01:27:08	$1.34\times$	$+21.9$ min
Llama-3.2-3B	02:06:30	02:32:36	$1.21\times$	$+26.1$ min
Mistral-7B	08:24:14	09:39:28	$1.15\times$	$+75.2$ min
Llama-2-7B	09:01:33	10:41:15	$1.18\times$	$+99.7$ min
Llama-3-8B	08:22:38	09:55:29	$1.18\times$	$+92.8$ min
Qwen3-8B	09:40:55	11:32:26	$1.19\times$	$+111.5$ min
Qwen3-14B	15:00:02	19:38:56	$1.31\times$	$+278.9$ min
Llama-2-13B	15:29:09	18:10:54	$1.17\times$	$+161.8$ min
Mean (11 models)	—	—	$1.21\times$	—

Optimization (ternary DBF runs). Prefit: AdamW, $\eta_{\text{pre}} = 10^{-4}$, 50 steps, float32. DBF: internal iterations $\{200, 400\}$ for $<$ 7B / $\geq$ 7B. CBQ refinement: Adam with the repository's optimization defaults (`cbq_epochs`$= 20$, `cbq_lr`$= 5 \times 10^{-5}$, rollback enabled). Chunk size $K = 4$, with one interleaved refinement sweep per chunk.

Optimization (GPTQ runs). Inner quantizer: weight-only GPTQ at either $W3$ (three bits per channel) or $W2g128$ (two bits, group size 128). GPTQ defaults used throughout: block size 128, Hessian damping `percdamp` $= 0.01$, no activation reordering (`actorder` $=$ false; we observed that enabling actorder changes PPL by < 0.1 and does not interact with the schedule, hence we report with the default value). Prefit is disabled ($T_{\text{pre}} = 0$) for every GPTQ comparison row. The CBQ refinement stage uses the same settings as the ternary DBF runs.

Calibration. 128 samples of length 2048 from WikiText-2 (or 256 samples for $\geq$ 7B), random starting positions as produced by the shared `get_loaders()` utility. Repository seed default 0 (matching `conf/config.yaml`). Identical calibration splits are shared between DBF and GPTQ runs on the same model.

Reproducibility. Every row of the main ternary DBF results is reproduced with the same calibration pipeline and optimization settings reported above, using chunk size 4, plain CBQ refinement, and 50 prefit steps for ternary DBF runs. The GPTQ rows are produced under the same pipeline with the inner quantizer switched to GPTQ and the bit/group-size pair (`w_bit`, `w_group`) $\in \{(3, -1), (2, 128)\}$.

D SUPPLEMENTARY RESULTS

Formatting note. In all tables of this section, **bold** marks the strongest value in each comparison group.

Main-set ternary DBF results under WikiText-2 calibration. Table 6 reports the seven-model DBF counterpart to the main C4-calibrated table, using WikiText-2 calibration only and PPL metrics only (no zero-shot). The qualitative trend is the same: ICBQ is the strongest across all cases for runs calibrated on WikiText-2 as well.

Supplementary DBF PPL results. Table 6 reports supplementary ternary DBF PPL results on four small models in the same format as the main ternary DBF table: WikiText-2 and C4 evaluation under both WikiText-2 and C4 calibration, comparing No-Ref., Sequential CBQ, and ICBQ. As shown in Table 6, the qualitative trend is unchanged: ICBQ is the strongest refinement across these cases.

Ternary-DBF zero-shot results. Table 6 consolidates ternary DBF C4-calibrated zero-shot scores on 11 models with the same three schedules as the main ternary DBF comparison: No-Ref. (ternary DBF only), Sequential CBQ, and ICBQ. As summarized in Table 6, ICBQ most often attains the best score across models and tasks, with a few task-specific ties or Seq. wins.

GPTQ zero-shot results. Table 7 mirrors the ternary DBF zero-shot presentation for GPTQ at $W3$ and $W2g128$ on the main seven-model set. As shown in Table 7, the pattern is consistent with the PPL results: ICBQ is usually best or tied, with a small number of task-specific Seq. wins.

Seam revisit ablation additional results (GPTQ). Table 8 mirrors Table 3 for interleaved GPTQ–CBQ on the same seven models (C4 calibration, $T_{\text{pre}} = 0$, full-model quantization), reporting W3 and W2g128. "No revisit" disables the second optimization of chunk-boundary pairs; all other settings match the seam rows. Seam revisiting improves C4 PPL in most cells on both bit-widths, with gaps typically below $\sim$0.3 PPL, though a few models favor no revisit and zero-shot does not always agree. Together with the DBF ablation, this suggests seam revisiting is a small, backend-dependent refinement rather than the main source of ICBQ gains.

Effect of prefit. Table 9 isolates the prefit step on the same seven-model set (PPL on C4) by setting $T_{\text{pre}} = 0$ under both schedules. The "No-prefit" rows keep all other hyperparameters identical to the DBF Prefit-enabled rows (DBF iters, CBQ epochs, learning rate, rollback, chunk size, calibration samples), so this is the minimal perturbation that isolates the contribution of prefit within the progressive driver. Even without prefit, ICBQ remains better than Sequential CBQ in the reported runs, indicating that the schedule effect is not limited to the prefit-enabled setting.

Table 6: Supplementary ternary DBF results. Part a. Main-set ternary DBF results under WikiText-2 calibration (PPL only). Part b. PPL results on 4 small models in the same format as the main ternary DBF results (WikiText-2 and C4, two calibration settings, and three schedules). Part c. Zero-shot results on 11 models under C4 calibration. Zero-shot values are normalized accuracy when available, otherwise accuracy (higher is better).

Part a. Main-set ternary DBF results under WikiText-2 calibration (PPL only).

	No-Ref. (ternary DBF only)		Seq. CBQ		**ICBQ (ours)**	
Metric	PPL (Wiki-2)	PPL (C4)	PPL (Wiki-2)	PPL (C4)	PPL (Wiki-2)	PPL (C4)
Standard models						
Mistral-7B	44.27	329.91	16.24	53.91	**10.09**	**25.59**
Llama-2-7B	11.23	20.05	8.38	14.59	**7.78**	**13.30**
Llama-2-13B	7.63	12.69	6.55	11.15	**6.32**	**10.59**
Deep and non-standard-attention models						
Llama-3.2-3B	40.24	66.50	25.30	49.90	**22.98**	**46.04**
Llama-3-8B	46.12	87.19	27.16	61.21	**24.01**	**48.92**
Qwen3-8B	4661.50	diverge	281.45	8275.66	**18.95**	**37.68**
Qwen3-14B	512.82	2522.21	35.24	119.40	**15.52**	**34.68**

Part b. Supplementary DBF PPL results on 4 small models.

	No-Ref. (ternary DBF only)				Seq. CBQ				**ICBQ (ours)**			
Calibration data	Wiki-2		C4		Wiki-2		C4		Wiki-2		C4	
PPL dataset	Wiki-2	C4	Wiki-2	C4	Wiki-2	C4	Wiki-2	C4	Wiki-2	C4	Wiki-2	C4
Qwen3-0.6B	89.84	215.91	165.73	143.15	62.58	137.71	101.4	106.7	**54.64**	**109.88**	**88.0**	**92.2**
Gemma-3-1B	198.94	407.36	462.94	278.29	86.08	175.41	171.9	130.7	**74.02**	**144.19**	**140.3**	**111.9**
TinyLlama-1.1B	27.97	47.90	58.03	43.81	18.84	33.58	30.2	30.4	**17.04**	**30.51**	**25.8**	**26.7**
Gemma-2-2B	85.36	151.92	142.18	92.71	85.31	164.43	106.4	89.2	**79.62**	**149.23**	**99.1**	**86.1**

Part c. Ternary-DBF zero-shot results on 11 models (C4 calibration).

Model	Schedule	BoolQ	PIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA
Qwen3-0.6B	No-Ref.	0.4180	**0.5702**	0.2925	0.5075	0.3245	0.2159	0.2500
	Seq. CBQ	0.5728	0.5642	0.2995	**0.5178**	0.3194	0.2048	0.2520
	ICBQ (ours)	**0.5942**	0.5637	**0.3071**	**0.5178**	**0.3304**	**0.2167**	**0.2660**
Gemma-3-1B	No-Ref.	**0.5061**	0.5533	0.2775	**0.5114**	0.3220	**0.2201**	0.2660
	Seq. CBQ	0.4826	0.5740	0.2813	0.5028	0.3283	0.2065	0.2680
	ICBQ (ours)	0.4979	**0.5811**	**0.2928**	0.5067	**0.3375**	0.2065	**0.3100**
TinyLlama-1.1B	No-Ref.	0.5862	0.6110	0.3368	0.5272	0.3497	0.2227	0.2860
	Seq. CBQ	0.5832	0.6295	0.3601	0.5249	0.3801	0.2406	**0.3060**
	ICBQ (ours)	**0.5905**	**0.6376**	**0.3725**	**0.5257**	**0.3838**	**0.2551**	0.2800
Gemma-2-2B	No-Ref.	0.5832	**0.6502**	0.3568	**0.5501**	0.4419	0.2474	0.2740
	Seq. CBQ	0.5890	0.6420	0.3529	0.5367	0.4432	0.2440	0.2740
	ICBQ (ours)	**0.6266**	0.6485	**0.3779**	0.5391	**0.4739**	**0.2517**	**0.3040**
Llama-3.2-3B	No-Ref.	0.6517	0.6202	0.3937	0.5478	0.3855	0.2543	0.2840
	Seq. CBQ	0.6673	0.6534	0.4253	0.5462	0.4314	**0.2765**	**0.2940**
	ICBQ (ours)	**0.6774**	**0.6801**	**0.4392**	**0.5604**	**0.4815**	0.2730	0.2880
Mistral-7B	No-Ref.	0.3856	0.5887	0.2913	0.5264	0.3022	0.2227	0.2620
	Seq. CBQ	0.5131	0.6697	0.4524	0.5446	0.4508	0.2858	0.2940
	ICBQ (ours)	**0.6875**	**0.7236**	**0.5855**	**0.5825**	**0.5568**	**0.3328**	**0.3140**
Llama-2-7B	No-Ref.	0.5807	0.6817	0.5259	0.5264	0.4928	0.2679	0.3240
	Seq. CBQ	0.6407	0.7138	0.5708	0.5785	0.5564	0.3131	**0.3580**
	ICBQ (ours)	**0.6914**	**0.7203**	**0.5875**	**0.6014**	**0.5880**	**0.3345**	0.3500
Llama-3-8B	No-Ref.	0.6734	0.6572	0.4305	0.5580	0.4314	0.2500	0.3000
	Seq. CBQ	**0.7064**	0.6931	0.4810	0.5485	0.4966	0.2944	0.2980
	ICBQ (ours)	0.6963	**0.7002**	**0.5038**	**0.5770**	**0.5299**	**0.3251**	**0.3180**
Qwen3-8B	No-Ref.	0.3862	0.5180	0.2606	0.4901	0.2567	0.2688	0.2660
	Seq. CBQ	0.5606	0.5876	0.3184	0.4901	0.3291	0.2526	0.2840
	ICBQ (ours)	**0.7407**	**0.6910**	**0.5057**	**0.5904**	**0.6111**	**0.3763**	**0.3440**
Llama-2-13B	No-Ref.	0.7505	0.7307	0.6261	0.6322	0.6662	0.3746	0.3920
	Seq. CBQ	**0.8083**	**0.7938**	0.6012	**0.7206**	0.7950	0.4829	0.3500
	ICBQ (ours)	0.8080	0.7927	**0.6016**	0.7151	**0.7955**	**0.4855**	**0.3520**
Qwen3-14B	No-Ref.	0.3786	0.5408	0.2632	0.4964	0.2976	0.2278	0.2660
	Seq. CBQ	**0.8933**	0.7987	**0.6103**	**0.7301**	0.8413	0.5862	**0.3500**
	ICBQ (ours)	0.8930	**0.8020**	0.6087	**0.7301**	**0.8426**	**0.5896**	0.3440

Table 7: **GPTQ zero-shot results on the main seven-model set (C4 calibration).** Weight-only GPTQ, no progressive prefit ($T_{\text{pre}} = 0$), and the same schedules as the main GPTQ comparison: GPTQ only, Sequential CBQ, and ICBQ with seam revisit. Values are normalized accuracy when available, otherwise accuracy (higher is better).

Model	Bit-width	Schedule	BoolQ	PIQA	Hella-Swag	Wino-Grande	ARC-e	ARC-c	OBQA
Mistral-7B	$W3$	GPTQ only	0.7425	0.7840	**0.7409**	0.6740	0.6915	0.4070	0.3780
		Seq. CBQ	0.7817	0.7905	0.7314	0.6709	0.7353	0.4292	**0.4340**
		ICBQ (ours)	**0.7933**	**0.7911**	0.7342	**0.6788**	**0.7382**	**0.4386**	**0.4340**
	$W2g128$	GPTQ only	0.5859	0.6942	0.5337	0.5627	0.4668	0.3029	0.2800
		Seq. CBQ	**0.6972**	0.7367	0.6039	0.6085	0.5833	**0.3456**	0.3240
		ICBQ (ours)	0.6813	**0.7437**	**0.6127**	**0.6298**	**0.5955**	0.3413	**0.3400**
Llama-2-7B	$W3$	GPTQ only	0.7281	0.7503	0.6775	0.6306	0.6082	0.3720	0.3840
		Seq. CBQ	**0.7450**	**0.7671**	0.6742	**0.6527**	**0.6629**	0.3788	**0.4060**
		ICBQ (ours)	0.7428	0.7666	**0.6828**	0.6480	0.6578	**0.3805**	0.3980
	$W2g128$	GPTQ only	0.5116	0.6551	0.5149	0.5643	0.4327	0.2858	0.3060
		Seq. CBQ	**0.6550**	0.6948	0.5539	0.5904	**0.5328**	**0.3012**	**0.3560**
		ICBQ (ours)	0.6529	**0.7106**	**0.5624**	**0.6133**	0.5316	0.3003	0.3540
Llama-2-13B	$W3$	GPTQ only	0.7502	0.7797	**0.7425**	**0.7024**	0.7088	**0.4428**	0.4260
		Seq. CBQ	**0.7639**	0.7813	0.7340	0.6906	0.7146	0.4249	0.4440
		ICBQ (ours)	**0.7639**	**0.7818**	0.7356	0.6914	**0.7214**	0.4283	**0.4460**
	$W2g128$	GPTQ only	0.5740	0.7214	0.6159	0.6062	0.5636	0.3285	0.3600
		Seq. CBQ	**0.7343**	0.7318	0.6375	0.6346	**0.6221**	**0.3652**	0.3820
		ICBQ (ours)	0.7245	**0.7465**	**0.6453**	**0.6543**	0.6204	0.3567	**0.3860**
Llama-3.2-3B	$W3$	GPTQ only	0.6465	0.6344	0.5535	0.5754	0.3960	0.2807	0.3140
		Seq. CBQ	**0.7398**	0.7035	0.5737	**0.5919**	0.5471	0.3319	0.3340
		ICBQ (ours)	0.7266	**0.7040**	**0.5862**	0.5809	**0.5673**	**0.3387**	**0.3440**
	$W2g128$	GPTQ only	0.4581	0.5256	0.3040	0.5130	0.2748	**0.2509**	0.2560
		Seq. CBQ	**0.6275**	0.6240	0.3873	0.5162	**0.3939**	0.2474	**0.2960**
		ICBQ (ours)	0.6242	**0.6306**	**0.4088**	**0.5406**	0.3927	0.2474	0.2920
Llama-3-8B	$W3$	GPTQ only	0.5832	0.6099	**0.6068**	0.5935	0.3817	0.2765	0.3320
		Seq. CBQ	0.7520	**0.7519**	0.5935	**0.6614**	0.6553	0.4147	0.3680
		ICBQ (ours)	**0.7920**	0.7465	0.6014	0.6504	**0.6679**	**0.4215**	**0.3880**
	$W2g128$	GPTQ only	0.4694	0.5365	0.3415	0.5209	0.2950	0.2295	0.2500
		Seq. CBQ	**0.6746**	0.6719	0.4608	0.5754	0.4907	0.2875	0.2840
		ICBQ (ours)	0.6544	**0.6893**	**0.5016**	**0.5848**	**0.5147**	**0.3029**	**0.3260**
Qwen3-8B	$W3$	GPTQ only	0.6581	0.7160	0.6204	0.5533	0.5248	0.3311	0.3260
		Seq. CBQ	**0.7820**	**0.7459**	0.6332	0.6204	0.6827	**0.4633**	**0.3680**
		ICBQ (ours)	0.7703	0.7443	**0.6350**	**0.6322**	**0.7311**	**0.4633**	**0.3680**
	$W2g128$	GPTQ only	0.4752	0.5713	0.3891	0.5178	0.3291	0.2312	0.2580
		Seq. CBQ	**0.7446**	**0.7084**	0.5200	**0.5998**	**0.6014**	0.3541	**0.3480**
		ICBQ (ours)	0.7425	0.7024	**0.5313**	0.5919	0.5753	**0.3567**	0.3360
Qwen3-14B	$W3$	GPTQ only	0.8138	0.7677	0.7156	0.6630	0.7205	0.4744	0.4220
		Seq. CBQ	0.8584	**0.7731**	0.7163	0.6732	**0.7828**	0.5119	0.4360
		ICBQ (ours)	**0.8587**	0.7704	**0.7221**	**0.6993**	0.7706	**0.5247**	**0.4540**
	$W2g128$	GPTQ only	0.6645	0.6844	0.5474	0.5454	0.4659	0.2910	0.3000
		Seq. CBQ	**0.8055**	0.7307	0.6070	**0.6504**	0.7033	**0.4403**	0.3620
		ICBQ (ours)	0.7884	**0.7448**	**0.6150**	0.6456	**0.7134**	0.4317	**0.3920**

Table 8: **Effect of seam revisiting on GPTQ interleaved CBQ** (seven-model set, C4 calibration, full-model GPTQ with $T_{\text{pre}} = 0$). *Seam revisit* is the standard ICBQ schedule (chunk-boundary pairs optimized twice); in the *no revisit* setting each adjacent pair is visited exactly once.

	W3 (per-channel)				W2g128 (group size 128)			
	C4 PPL (↓)		Avg. ZS (↑)		C4 PPL (↓)		Avg. ZS (↑)	
Base model	Seam	No revisit	Seam	No revisit	Seam	No revisit	Seam	No revisit
Mistral-7B	**10.32**	10.33	**0.6583**	0.6567	**14.20**	14.27	0.5635	**0.5700**
Llama-2-7B	9.75	9.75	0.6109	**0.6130**	**13.54**	13.56	**0.5322**	0.5243
Llama-2-13B	8.18	8.18	0.6526	**0.6530**	**10.82**	10.83	**0.5905**	0.5892
Llama-3.2-3B	**23.14**	23.22	**0.5497**	0.5495	**41.56**	41.81	**0.4480**	0.4472
Llama-3-8B	**23.52**	23.56	**0.6097**	0.6071	**38.68**	39.35	**0.5105**	0.5102
Qwen3-8B	**21.09**	21.61	**0.6206**	0.6172	28.29	**28.19**	0.5480	**0.5552**
Qwen3-14B	17.51	**17.49**	0.6857	**0.6875**	**22.21**	22.35	**0.6187**	0.6145

Table 9: **Interleaving helps even without prefit.** PPL on C4 (↓) with ternary DBF, with and without the prefit step ($T_{\text{pre}} = 50$ vs. 0), on the same seven-model set.

Base model	Prefit + Seq. CBQ	Prefit + Interleaved CBQ (ours)	No-prefit + Seq. CBQ	No-prefit + Interleaved CBQ
Mistral-7B	31.1	**16.8**	19.71	**18.9**
Llama-2-7B	13.6	**12.5**	15.50	**15.4**
Llama-2-13B	10.48	**10.09**	11.85	**11.80**
Llama-3.2-3B	40.63	**37.01**	51.66	**50.00**
Llama-3-8B	48.98	**42.32**	68.53	**66.35**
Qwen3-8B	5752.9	**29.90**	47.28	**41.59**
Qwen3-14B	44.82	**24.67**	35.30	**31.42**