SADA: Bridging In-Context Learning and Fine-Tuning via State-Aligned Distillation Adapters

Wenhao Gao [♠♡*], **Tianlong Wang** [♠♢*], **Wei Jia** [♠], **Linhao Zhang** [♠], **Aiwei Liu** [♠], **Miao Fan** [♠], **Xiao Zhou** [♠†]

♠ Basic Model Technology Center, WeChat AI, Tencent Inc., China

♡ Beijing University of Posts and Telecommunications, China

♢ School of Software and Microelectronics, Peking University, China

{wenhaoogao, taronwang, jarviswjia, neuronzhang, coveliu, kathleenfan, chappyzhou}@tencent.com, whgao@bupt.edu.cn, tianlong.wang@stu.pku.edu.cn

Abstract

Prompt-based in-context learning (ICL) and parameter fine-tuning are two dominant paradigms for incorporating external information into large language models (LLMs), but they incur high inference costs or require expensive retraining. To bridge this gap, context-to-parameter mapping converts prompts into temporary adapter weights. However, we identify a critical failure mode in existing methods: *hidden-state collapse*, where the adapter-augmented model's internal states diverge sharply from the full-context oracle in deeper layers. We trace this failure to two coupled gaps: suboptimal **Input-Selection** and inadequate **Supervision-Signal**. To address these issues, we propose SADA (**S**tate-**A**ligned **D**istillation **A**dapters). We establish the *attention-block output* as a principled feature interface to improve input selection and introduce *state-alignment distillation* to enforce consistency between the adapter-augmented model and the full-context oracle. Experiments on long-context language modeling (PG19) and downstream NLU and summarization benchmarks show that SADA consistently outperforms strong baselines like *StreamAdapter* and *GenerativeAdapter*, achieving performance comparable to ICL while significantly reducing memory footprint and latency. We further analyze when parameterized context compression is effective and when explicit context retention remains preferable. Our code is available at https://github.com/Taylor-Gavel/SADA.git.

1 Introduction

In recent years, Large Language Models (LLMs) have achieved strong performance across many tasks (Minaee et al., 2024; Zhao et al., 2023), yet general-purpose models still struggle with domain-specific or complex real-world requirements (Chen et al., 2025a). Consequently, efficient and precise knowledge injection has become a central problem (Lv et al., 2025). Currently, two paradigms dominate this landscape: parameter-update-based fine-tuning (FT) and prompt-based In-Context Learning (ICL) (Dong et al., 2024; Han et al., 2024).

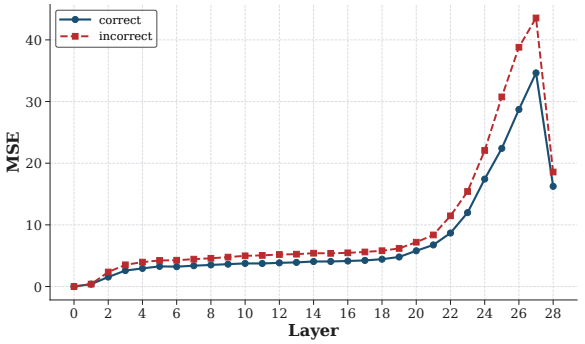


Figure 1: **Hidden-state fidelity predicts correctness.** Layer-wise hidden-state MSE between the injected model and the full-context oracle.

Fine-tuning internalizes new information into model weights, yielding efficient inference once training is done, but it is costly to repeat and hard to control for localized knowledge edits. Updating parameters can also unintentionally overwrite prior capabilities (i.e., catastrophic forgetting) (Li et al., 2024a; van de Ven et al., 2024).

In contrast, ICL keeps evidence explicit in the prompt, making updates reversible and more controllable. However, long contexts are slow and memory-intensive (Wan et al., 2024): compute and the KV cache scale linearly with context length, quickly exhausting GPU memory and limiting batch size.

To bridge the gap between the flexibility of ICL and the efficiency of FT, a new paradigm known as **Context-to-Parameter Mapping** (or Test-Time Adaptation) has emerged. Pioneering works such

* Equal contribution.

† Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 22847–22862

as *GenerativeAdapter* (Chen et al., 2025b) and *StreamAdapter* (Muhtar et al., 2024) convert contextual representations into temporary adapter parameters (e.g., LoRA weights) via a lightweight hypernetwork, aiming to achieve "fine-tuning-like adaptation at inference time" in a single forward pass.

However, we identify a critical failure mode in these approaches regarding *behavioral fidelity*. As illustrated in Figure 1, we compare the layer-wise internal states of the injected model StreamAdapter against a full-context oracle. The results reveal a decisive correlation: when the model answers correctly, its hidden states remain close to the oracle (low MSE); conversely, incorrect answers diverge sharply in deeper layers. This highlights a concrete consequence: supervision on logits alone is insufficient, because once the intermediate reasoning trajectory deviates, the model can fail on complex reasoning or long-range dependencies even if the final outputs are matched.

Viewed through this lens, existing methods fail to achieve such alignment due to two coupled gaps:

- **(i) Input-Selection Gap:** Prior methods such as *StreamAdapter* use the KV cache, while *GenerativeAdapter* uses raw hidden states as features for parameter generation. These choices are not explicitly grounded in how attention-driven context shifts induce effective weight updates, so the mapping module can fail to preserve the oracle trajectory.
- **(ii) Supervision-Signal Gap:** Current objectives primarily optimize for outcome matching (e.g., Cross-Entropy on Ground Truth). As evidenced by Figure 1, such supervision is insufficient to constrain intermediate states that support multi-step reasoning, leading to trajectory divergence in deep layers.

Motivated by this evidence, we propose **SADA** (**S**tate-**A**ligned **D**istillation **A**dapters). To resolve the *input-selection* gap, we analytically justify the **attention output** as a principled feature interface for context-to-parameter conversion, drawing a connection to implicit gradient-based adaptation. To resolve the *supervision-signal* gap, we introduce **State-Alignment Distillation** to align intermediate feature distributions between the injected model and the full-context oracle, improving behavioral fidelity beyond output matching while keeping memory footprint small.

We evaluate SADA on long-context modeling, ICL, and downstream benchmarks. SADA consistently outperforms StreamAdapter and exceeds GenerativeAdapter where reported, while also reducing memory and latency compared to ICL and lowering training cost relative to fine-tuning. We also characterize a clear task boundary: SADA is robust on **concept-oriented tasks** via pattern compression, whereas entity-intensive tasks that demand precise factual recall are better served by retrieval or KV cache compression methods (Zhang et al., 2023; Li et al., 2024b), suggesting that parameterized mapping acts as lossy compression that attenuates detailed entity bindings.

The main contributions of this paper are summarized as follows:

- **SADA Framework:** We propose a parameterized knowledge injection framework featuring a Mapping Module. This module dynamically converts the internal representations of context prompts into model parameters, bridging the gap between ICL flexibility and FT efficiency.
- **Theoretical Foundation & Alignment:** We provide theoretical motivation linking attention outputs to gradient-based adaptation, offering a principled foundation for feature selection in context-to-parameter mapping, and introduce a projection-based state-alignment strategy that better preserves internal reasoning trajectories.
- **Extensive Validation:** Empirical results show that SADA outperforms *StreamAdapter* and *GenerativeAdapter* on long-context modeling and summarization, while improving inference efficiency (memory and latency) relative to standard ICL and reducing training resource consumption compared to fine-tuning.

2 Related Work

Efficient Long-Context Modeling. Long-context ICL can be slow and memory-intensive, so prior work mainly follows two directions: modifying attention and compressing context. Architectural approaches such as sparse attention (Child et al., 2019) and sliding windows (e.g., *StreamingLLM* (Xiao et al., 2024)) reduce compute but may weaken global awareness. Compression methods such as *LLMLingua* (Jiang et al., 2023) and *SnapKV* (Li et al., 2024b) remove tokens or KV pairs to fit a budget, which may

drop information. **In contrast**, SADA avoids growing the KV cache by compressing history into dynamically generated parameters.

Parameter-Efficient Fine-Tuning (PEFT). PEFT adapts frozen LLMs via lightweight modules (Han et al., 2024), spanning addition-based Adapters (Houlsby et al., 2019), prompt-based Prefix-Tuning (Li and Liang, 2021), and reparameterization methods like LoRA (Hu et al., 2022). However, these approaches yield static weights that remain fixed after training. **SADA distinguishes itself** by functioning as a hypernetwork that dynamically generates instance-specific parameters from the input stream, combining PEFT's efficiency with the real-time adaptability of In-Context Learning.

Context-to-Parameter Mapping. The most relevant prior works are hypernetwork-based methods like *StreamAdapter* (Muhtar et al., 2024) and *GenerativeAdapter* (Chen et al., 2025b), which map context directly to adapter weights. Unlike these approaches, which typically rely on heuristic feature sources and logit-level supervision, **SADA distinguishes itself** by grounding the mapping in *attention-output* features and enforcing *layer-wise state alignment* via State-Aligned Distillation.

3 Methodology

In this section, we formally introduce **SADA**, a framework designed to bridge the gap between transient context prompting and static parameter fine-tuning. We begin by analyzing the mechanics of ICL through the lens of gradient updates, which provides the theoretical grounding for our architecture. Subsequently, we detail the SADA mechanism, structured around state-space modeling (Dao and Gu, 2024), and conclude with our projection-based distillation training objective. The exposition mirrors the two coarse-grained gaps discussed earlier: we first justify the feature source for context-to-parameter conversion, then describe how to align internal states rather than only matching output logits.

3.1 Motivation: ICL as Layer-wise Implicit Gradient Updates

To motivate a module that maps context into parameters, we interpret in-context information as inducing *layer-wise, first-order* effective weight updates. We deliberately restrict the argument to a single-layer linearization: it provides the right intuition without assuming global linearity across depth.

Layer-wise view. Let a Gated MLP layer be

$$\mathbf{y} = \mathbf{W}_{down}(\phi(\mathbf{W}_{gate}\mathbf{x}) \odot (\mathbf{W}_{up}\mathbf{x})) . \quad (1)$$

In ICL, attention retrieves contextual information and adds it to the residual stream,

$$\mathbf{x}' = \mathbf{x} + \boldsymbol{\delta}_{ctx}. \quad (2)$$

A first-order expansion around $\mathbf{x}$ shows that this input shift is equivalent to *low-rank* perturbations of the layer parameters, i.e., a LoRA-like update with a shared right factor $\boldsymbol{\delta}_{ctx}$:

$$\mathbf{W}' \approx \mathbf{W} + \mathbf{U}(\mathbf{x})\, \boldsymbol{\delta}_{ctx}^{\top}, \quad (3)$$

where $\mathbf{U}(\mathbf{x})$ is an input-dependent factor determined by the local Jacobian (full derivation in Appendix A). This connects ICL to an *implicit gradient-style* modulation: context specifies the update direction while the current state $\mathbf{x}$ determines the local scaling.

Why attention output features? We take $\boldsymbol{\delta}_{ctx}$ as the attention output, $\boldsymbol{\delta}_{ctx} = \mathrm{softmax}(QK^{\top})V$. It is naturally aligned with $\mathbf{x}$ via queries, normalized by softmax, and aggregated across heads, yielding a stable signal that directly reflects how context modifies the residual stream. Using the attention output as input to our mapping module therefore provides a principled and effective feature for generating $\Delta\mathbf{W}$ (Table 3).

Implication. We thus model "memory injection" as dynamically generating layer-specific low-rank updates $\Delta\mathbf{W}$ that modulate both content transformation and routing.

3.2 The SADA Architecture

SADA is an auxiliary memory mechanism that summarizes *evicted tokens* (history) and uses the resulting state to modulate the processing of *local tokens* (current context). At a high level, SADA implements: *compress(history)* $\rightarrow$ *evolve(state)* $\rightarrow$ *inject a low-rank update into selected frozen layers*. Figure 2 illustrates the training signals used to supervise these dynamically generated updates.

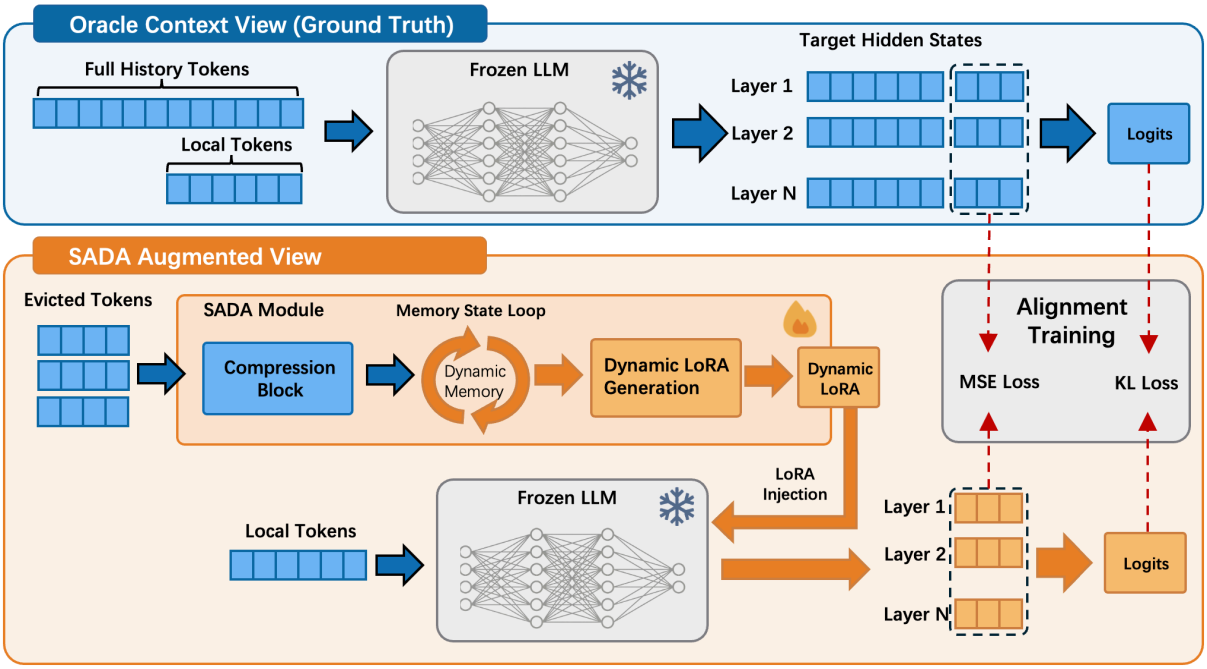


Figure 2: Training strategy for SADA showing distillation signals supervising dynamically generated weight updates.

3.2.1 Memory Encoder: Compression and State Evolution

Let $\mathbf{X}_{hist} \in \mathbb{R}^{L\times d}$ denote evicted tokens, i.e., the historical *attention outputs* (layer-wise residual features) produced by the frozen LLM at the corresponding layer. We split $\mathbf{X}_{hist}$ into chunks $\{\mathbf{X}_k\}_{k=1}^{K}$ of size C. For each chunk, we extract a fixed-size semantic summary using N learnable queries $\mathbf{Q}_{meta}$ via intra-chunk cross attention.

Chunk projection.

$$\mathbf{K}_k = \mathbf{X}_k\mathbf{W}_K,\ \ \mathbf{V}_k = \mathbf{X}_k\mathbf{W}_V, \tag{4}$$

Compression.

$$\mathbf{U}_k = \mathrm{CrossAttn}(\mathbf{Q}_{meta},\ \mathbf{K}_k,\ \mathbf{V}_k) \tag{5}$$

where $\mathbf{W}_K, \mathbf{W}_V$ are standard projections.

State evolution. Inspired by the *selective state update* idea in Mamba-2 (Dao and Gu, 2024), we maintain continuity across chunks with a gated recurrent update over meta memory states $\mathbf{M}_k \in \mathbb{R}^{N\times d_{state}}$:

$$\begin{aligned}\mathbf{M}_k &= \gamma_k \odot \mathbf{M}_{k-1} + \mathbf{U}_k,\\ \gamma_k &= \sigma(\mathbf{U}_k\mathbf{W}_\gamma + \mathbf{b}_\gamma)^{1/\tau},\end{aligned} \tag{6}$$

where γ_k is a decay gate, τ controls retention, and $\odot$ is element-wise multiplication. The resulting $\mathbf{M}_k$ serves as a compact, evolving representation of the evicted history.

3.2.2 Dynamic Injection via Meta-LoRA

We inject the memory state into a frozen LLM by generating a context-conditioned low-rank update for selected linear layers (e.g., the attention output projection). We use the row-vector convention: for a token representation $\mathbf{x}_t \in \mathbb{R}^{1\times d}$ and a linear weight $\mathbf{W}_0 \in \mathbb{R}^{d\times d_{out}}$, the layer output is $\mathbf{y}_t = \mathbf{x}_t\mathbf{W}_0$. Under this convention, we adopt the LoRA form $\Delta\mathbf{W}_k = \mathbf{A}\mathbf{B}_k$, where $\mathbf{A}$ is static and $\mathbf{B}_k$ is generated from memory.

Specifically, we set the static down-projection $\mathbf{A} \triangleq \mathbf{W}_{MetaA} \in \mathbb{R}^{d\times d_{state}}$ and produce a dynamic up-projection from the current memory state $\mathbf{M}_k \in \mathbb{R}^{N\times d_{state}}$:

$$\mathbf{B}_k = \mathbf{M}_k^\top \mathbf{W}_{MetaB} \in \mathbb{R}^{d_{state}\times d_{out}}, \tag{7}$$

where $\mathbf{W}_{MetaB} \in \mathbb{R}^{N\times d_{out}}$ maps memory slots to the layer output space. The forward pass is:

$$\begin{aligned}\mathbf{y}_t &= \mathbf{x}_t\mathbf{W}_0 + \alpha\big(\mathbf{x}_t\mathbf{W}_{MetaA}\big)\mathbf{B}_k\\ &= \mathbf{x}_t\mathbf{W}_0 + \alpha\,\mathbf{x}_t\underbrace{\big(\mathbf{W}_{MetaA}\mathbf{B}_k\big)}_{\Delta\mathbf{W}_k},\end{aligned} \tag{8}$$

where α is a scaling factor and $\Delta\mathbf{W}_k$ is a rank-d_{state} update conditioned on the compressed history $\mathbf{M}_k$.

Remark. This design separates (i) memory formation in $\mathbf{M}_k$ from (ii) parameter modulation via $\Delta\mathbf{W}_k$, enabling efficient adaptation without storing the full historical context.

3.3 Training Strategy

The overall training pipeline is illustrated in Fig. 2, showing how the mapping module aligns with oracle states from full-context processing. Our method follows a two-stage recipe. In **Stage I (General Distillation)**, we use large-scale pre-training data with a sliding-window scheme and optimize with MSE on hidden states plus KL divergence on logits. In **Stage II (Task Alignment)**, we distill long-context hidden states while optimizing the target output with LM_Loss. Accordingly, SWSD is used in Stage I for long-context distillation, while TAD is used in Stage II for task alignment.

3.3.1 Stage I (General Distillation): Sliding Window State Distillation (SWSD)

For general language generation and long-context processing, we employ SWSD to train the Mapping Module to compress historical context into weight space. For a sequence X of length L, we utilize a window size C' and a stride size Δ. The process follows an iterative cycle:

- **Context Eviction and Encoding**: In each step, we evict the earliest Δ tokens and encode their attention-output hidden states (i.e., the attention-induced residual increment) to predict parameter updates ΔW.
- **Dual-Path Forward Pass**: We run two forward passes for the incoming Δ tokens: a *Teacher Path* with the full-context KV cache and a *Student Path* using ΔW with a truncated context.
- **Gradient Accumulation**: The loss aligns the student's hidden states and logits with the teacher and is accumulated over X before updating SADA parameters.

We optimize the mapping module with hidden-state regression and logit alignment:

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{KL}} \tag{9}$$

where $\mathcal{L}_{\text{MSE}}$ is the mean-squared error averaged across layers,

$$\mathcal{L}_{\text{MSE}} = \frac{1}{|\mathcal{S}|} \sum_{\ell \in \mathcal{S}} \left\| H_\ell^t - H_\ell^s \right\|_2^2, \tag{10}$$

and $\mathcal{L}_{\text{KL}}$ aligns the teacher and student output distributions over the predicted tokens:

$$\mathcal{L}_{\text{KL}} = \frac{1}{\Delta} \sum_{i=1}^{\Delta} \text{KL}(p_i^t \,\|\, p_i^s), \tag{11}$$

where $p_i = \text{softmax}(z_i)$ represents the probability distribution over the vocabulary obtained from the output logits z_i at step i.

3.3.2 Stage II: Task Alignment Distillation (TAD)

To adapt SADA for a broader range of post-training tasks, we employ TAD with a 2-forward-1-backward strategy. For each sample, we form a task-specific context (e.g., exemplars, instructions, or prompts) and a target sequence to be predicted.

1. **First Forward (Oracle States, no-grad)**: We run the frozen base model on the *full* input (context + target) *without gradient* to obtain the teacher cache and oracle hidden states at the target positions.
2. **Second Forward (Parameter Injection)**: The SADA mapping module encodes the teacher cache (e.g., teacher attention outputs) to predict parameter updates ΔW. We then run the student forward using the truncated explicit context with injected ΔW, and process the target sequence.
3. **Task Alignment (single backward)**: We compute a combined loss consisting of (i) hidden-state MSE against the oracle (averaged across selected layers and target positions) and (ii) the language-modeling loss on the target tokens, and backpropagate only through the mapping/SADA module.

In this stage, we combine hidden-state distillation with the language modeling objective:

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{LM}} + \mathcal{L}_{\text{MSE}}. \tag{12}$$

Here, $\mathcal{L}_{\text{LM}}$ denotes the standard language modeling cross-entropy loss over the target sequence:

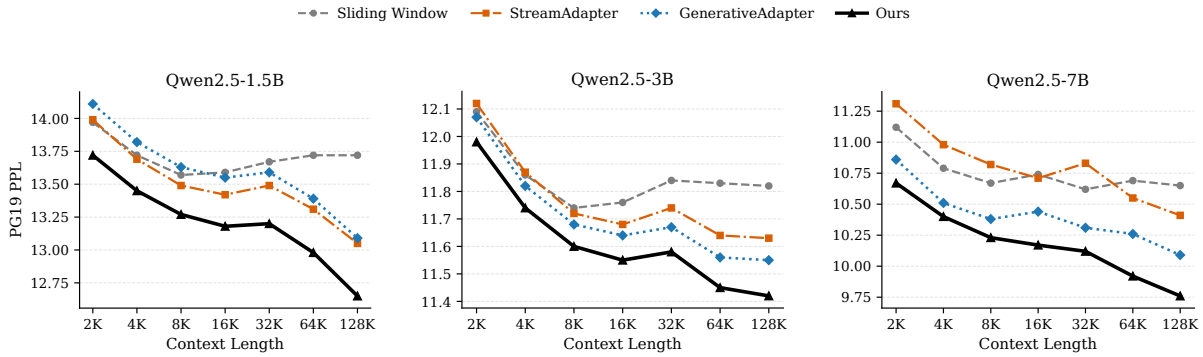
$$\mathcal{L}_{\text{LM}} = -\frac{1}{T} \sum_{t=1}^{T} \log p_\theta(x_t \mid x_{<t}), \tag{13}$$

where T is the target sequence length and p_θ is the student model's output distribution.

4 Experiment

4.1 Setup

We evaluate SADA across two primary dimensions: long-context language modeling and post-training alignment on downstream NLU and summarization tasks.



(a) Qwen2.5-1.5B-Instruct (b) Qwen2.5-3B-Instruct (c) Qwen2.5-7B-Instruct

Figure 3: **Long-context modeling results on PG19 across context lengths.** Perplexity (PPL; lower is better) from 2K to 128K for three Qwen2.5 backbones. SADA consistently achieves the lowest PPL across all tested lengths and model scales, with especially clear gains in extended-context regimes.

Table 1: Main results on NLU (Seen/Unseen; accuracy) and long-form summarization (GovReport and QMSum; ROUGE) with Qwen2.5-Instruct backbones. ICL uses the full prompt and serves as an upper bound.

	Qwen2.5-1.5B-Instruct				Qwen2.5-3B-Instruct				Qwen2.5-7B-Instruct			
	NLU		Summarization		NLU		Summarization		NLU		Summarization	
Method	**Seen**	**Unseen**	**GovReport**	**QMSum**	**Seen**	**Unseen**	**GovReport**	**QMSum**	**Seen**	**Unseen**	**GovReport**	**QMSum**
ICL	80.22	57.48	25.98	21.62	79.31	63.68	31.97	24.31	83.49	69.96	31.51	23.55
Fine-tuning	56.92	54.62	7.87	17.77	77.34	61.75	8.21	14.86	80.98	65.06	18.84	19.33
SnapKV	77.07	55.73	10.72	14.36	77.92	62.04	12.93	14.78	83.80	65.64	16.84	15.63
StreamAdapter	85.01	55.62	18.03	11.08	87.25	60.43	22.32	19.17	88.53	64.10	23.66	16.34
GenerativeAdapter	85.08	56.11	18.17	12.12	86.60	61.69	21.14	17.82	87.83	65.09	21.50	17.49
Ours	**86.50**	**56.68**	**19.07**	**18.64**	**88.18**	**63.33**	**24.57**	**19.68**	**89.90**	**68.79**	**26.52**	**20.54**

Baselines. We compare against five representative approaches: **ICL** (full prompt + full KV cache at inference, used as a full-context reference), **Fine-tuning** (SFT on task data), **SnapKV** (KV-cache pruning/compression), and two context-to-parameter hypernetwork baselines, **StreamAdapter** and **GenerativeAdapter** (see related work for details).

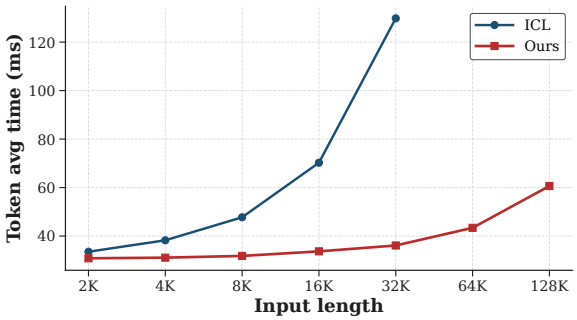
Models. We utilize **Qwen2.5-1.5B-Instruct**, **Qwen2.5-3B-Instruct**, and **Qwen2.5-7B-Instruct** as the foundation models for post-training alignment experiments.

Datasets. For long-context capabilities, we evaluate language modeling perplexity using the **PG19** dataset(Rae et al., 2019). For post-training task alignment, we evaluate on a mixture of **NLU** benchmarks and long-form **summarization** datasets (GovReport (Huang et al., 2021) and QMSum (Zhong et al., 2021)), and use a Seen/Unseen split to measure in- vs. out-of-distribution generalization.

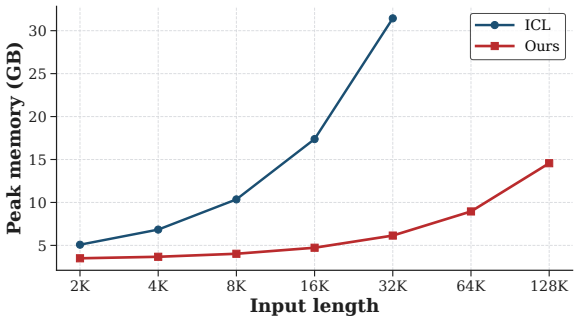
4.2 Long-Context Modeling (PG19).

Training Details. We set the training sequence length to 8,192 and employed a sliding-window approach for progressive distillation, matching Stage I distillation settings. We evaluated the perplexity (PPL) of our method against *StreamAdapter*, *GenerativeAdapter*, and baseline sliding-window approaches using the PG19 dataset.

Main Results. We evaluate PG19 test perplexity across context lengths ranging from 2K to 128K, using stride-averaged PPL to compare our method against Sliding Window, *StreamAdapter*, and *GenerativeAdapter*. As shown in Fig. 3, SADA achieves the lowest PPL across all tested lengths, consistently outperforming sliding-window and state-of-the-art adapter baselines at all three model scales, particularly in extended contexts. Notably, these improvements persist well beyond the training-context regime, indicating that state-aligned distillation supports reliable extrapolation to much longer contexts without retaining the full KV cache.



(a) Average per-token latency (ms) on PG19 across input lengths.



(b) Peak GPU memory (GB) under the same sweep, highlighting KV-cache savings.

Figure 4: Efficiency evaluation on PG19 contrasting SADA and ICL based on the Qwen2.5-1.5B-Instruct model: (a) average per-token latency and (b) peak GPU memory across input lengths.

Efficiency Test. To quantify inference-time efficiency, we measure (i) average per-token latency and (ii) peak GPU memory on **Qwen2.5-1.5B-Instruct** while sweeping input length from 2K to 128K on PG19. As shown in Fig. 4, SADA reduces both latency and memory relative to prompt-based ICL by replacing long KV-cache retention with parameterized compression of the historical context. Although SADA introduces a one-time offline training cost, the resulting inference gains make this trade-off favorable in repeated long-context deployment scenarios.

4.3 Post-Training Alignment on NLU and Summarization.

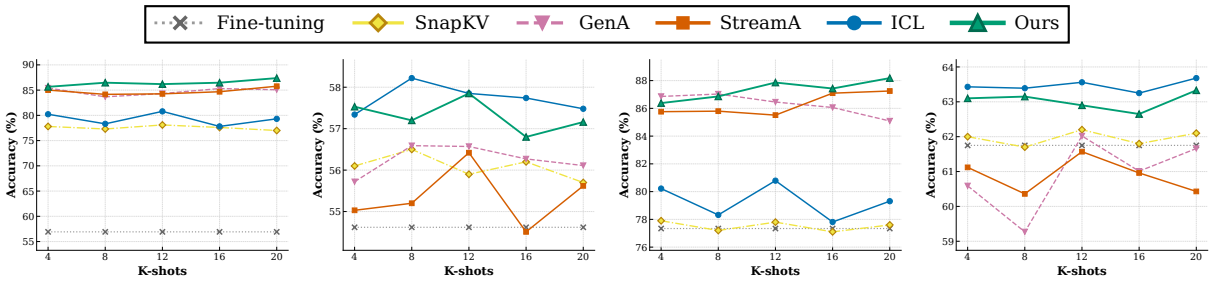
Training Details. We perform post-training alignment using a mixture of NLU datasets: BoolQ (Clark et al., 2019), CoPA (Melissa et al., 2011), SST-2 (Richard et al., 2013), CB (De Marneffe et al., 2019), and RTE (Bentivogli et al., 2009). To analyze sensitivity, we scale the few-shot context from 4 to 20 instances; the final two serve as the in-context prefix, while others are compressed into model parameters. Training is conducted over 3 epochs with a batch size of 64 and a learning rate of 6×10^{-5} on eight 40GB GPUs. We evaluate performance on **Seen** (in-distribution) and **Unseen** (OOD) benchmarks, the latter including ARC (Clark et al., 2018), HellaSwag (Zellers et al., 2019), OpenBookQA (Mihaylov et al., 2018), PIQA (Bisk et al., 2020), and WinoGrande (Sun and Emami, 2024). Table 1 reports 20-shot results by default.

Main Results. Table 1 reports accuracy/ROUGE on Seen/Unseen NLU and long-form summarization under an ICL-style protocol, where a fixed set of demonstrations is provided and only the most recent few are kept as an explicit prefix. On Seen NLU, SADA performs best for both backbones, consistently improving over the strongest mapping baseline by a clear margin, while remaining close to full-prompt ICL on Unseen benchmarks. On summarization, SADA yields substantial gains over adapter baselines, but still falls short of full-prompt ICL. This pattern indicates that parameterized compression transfers in-distribution task behavior effectively, while explicit prompts retain an advantage under distribution shift and evidence-heavy generation, which we further probe in the QA/summarization analysis and the alignment ablations.

Discussion. Across model sizes and shot budgets (Fig. 5), SADA fits in-distribution patterns strongly (Seen) while narrowing—but not fully closing—the gap to full-prompt ICL under distribution shift (Unseen). This trade-off is consistent with replacing explicit context retention by parameterized compression.

4.4 Ablation Studies

Conceptual vs. Detailed Memory. Building on the PG19 checkpoint, we further fine-tuned on downstream Sum and QA tasks and observed a divergence: As shown in Tab. 2, QA drops markedly while summarization degrades more gently. This suggests SADA preserves high-level semantics but struggles with fine-grained retrieval compared to prompt-only ICL. GenerativeAdapter yields stronger QA than SADA but remains far below ICL, whereas SADA achieves the best summarization among adapter-based baselines, highlighting a trade-off between entity recall and se-



(a) Qwen2.5-1.5B seen. (b) Qwen2.5-1.5B unseen. (c) Qwen2.5-3B seen. (d) Qwen2.5-3B unseen.

Figure 5: ICL evaluation results for SADA across model sizes and settings.

mantic compression. A plausible explanation is that low-rank parameter updates retain global topical structure and discourse flow, but weaken entity bindings and exact spans needed for QA, especially on multi-hop settings requiring precise entity chains and temporal relations. This suggests pairing SADA with retrieval/KV retention for entity-centric queries, while remaining effective for concept-oriented tasks that benefit from semantic compression.

Table 2: Qwen2.5-1.5B-Instruct on QA (HotpotQA (Yang et al., 2018), 2WikiMQA (Ho et al., 2020)) and summarization (GovReport (Huang et al., 2021), QMSum (Zhong et al., 2021)).

	QA		Sum	
Qwen2.5-1.5B	**HotpotQA**	**2WikiMQA**	**GovReport**	**QMSum**
ICL	42.41	35.58	25.98	21.62
SnapKV	**38.51**	**29.47**	10.72	14.36
StreamAdapter	20.28	19.13	18.03	11.08
GenerativeAdapter	28.19	25.19	18.17	12.12
Ours	23.86	26.15	**19.07**	**18.64**

Impact of Loss Components. Our task-alignment stage (Stage II) couples language modeling supervision with hidden-state distillation: (i) hidden-state alignment ($\mathcal{L}_{\text{MSE}}$) that constrains the internal reasoning trajectory, and (ii) LM_Loss ($\mathcal{L}_{\text{LM}}$) that ensures correct target generation. To focus on the alignment stage, we keep Stage I fixed and run an ICL-based ablation protocol on **Qwen2.5-1.5B-Instruct** (TAD setting): we fix the prompt construction and compression policy (retain the last two demonstrations as the explicit prefix and compress the remaining exemplars into the generated parameters), and evaluate under a **20-shot** setting on the Seen/Unseen benchmark split described earlier. We vary only the alignment-stage loss while keeping the architecture and training hyperparameters unchanged: (*1*) **Full** objective $\mathcal{L}_{\text{LM}} + \mu\mathcal{L}_{\text{MSE}}$; (*2*) **w/o $\mathcal{L}_{\text{MSE}}$**, which removes state trajectory supervision and tests whether task supervision alone is sufficient; (*3*) **w/o $\mathcal{L}_{\text{LM}}$**, which removes task supervision and tests whether distillation alone can recover task accuracy.

Both loss terms matter, but the LM loss is more critical: removing $\mathcal{L}_{\text{LM}}$ causes a large drop, while removing $\mathcal{L}_{\text{MSE}}$ yields a smaller decline. LM supervision anchors task performance, and state matching stabilizes behavior transfer under compression.

Table 3: Proposed ICL ablations on **Qwen2.5-1.5B-Instruct** under a 20-shot setting (Seen/Unseen split as in Fig. 5).

Variant	**Seen Acc.**	**Unseen Acc.**	**Avg.**
Impact of Loss Components			
Full ($\mathcal{L}_{\text{LM}} + \mathcal{L}_{\text{MSE}}$)	**87.38**	**57.16**	**72.27**
w/o $\mathcal{L}_{\text{MSE}}$	85.84	56.32	71.08
w/o $\mathcal{L}_{\text{LM}}$	79.90	55.46	67.68
Input Feature Sources			
Attention output	**87.38**	**57.16**	**72.27**
Attention input	84.20	53.45	68.83
KV features (K/V aggregated)	86.34	56.44	71.19
Post-MLP activations	85.28	56.47	70.88
Token embeddings	85.13	55.77	70.45

Input Feature Sources. We additionally ablate the *feature source* fed into the Mapping Module, since our method hinges on which representation is most "convertible" into effective parameter updates. Under the same TAD setup as above (1.5B, 20-shot, identical loss weights), we replace the mapping input with: (*1*) **Token embeddings** (pre-Transformer), (*2*) **Attention input** (the normalized residual stream entering the attention block, i.e., the features used to form $Q/K/V$), (*3*) **Attention output** (post-attention residual), (*4*) **Post-MLP activations**, and (*5*) **KV features** (aggregated keys/values from the teacher cache).

This validates our claim in Section 3.1 that the

attention *output* is the optimal interface for parameter modulation. Empirically, we expect embeddings to underperform due to their syntactic nature and post-MLP features to suffer from nonlinear distortion. KV features, while competitive, likely lag due to head-wise fragmentation and formatting sensitivity.

5 Conclusion

We introduced SADA, a framework that maps context into low-rank parameter updates via attention-output features and state-aligned distillation. Across PG19 and ICL benchmarks, SADA improves long-context modeling, delivers strong in-distribution accuracy, outperforming StreamAdapter and GenerativeAdapter, and reduces latency and memory compared to prompt ICL. Ablations clarify the roles of LM supervision and state matching, while analysis shows remaining gaps on entity-heavy and out-of-distribution settings.

6 Limitations

Lossy Compression. Transforming context into parameters acts as a lossy compression mechanism. While effective for high-level semantic tasks like summarization, it degrades performance on entity-intensive tasks (e.g., QA) that require precise factual recall, where explicit KV caches remain superior.

OOD Generalization Gap. Our experiments show a slight performance gap on "Unseen" benchmarks compared to standard ICL. Explicit context prompting proves more robust for novel task distributions not encountered during the distillation phase, whereas parameterized mapping may slightly overfit to training distributions.

Training Overhead. Unlike standard ICL, SADA requires a one-time offline training procedure through two-stage distillation before deployment. This additional cost increases the barrier to adoption compared with training-free prompting, and prevents instant adaptation to arbitrary new contexts. A detailed analysis of the practical overhead and the training/inference complexity is provided in Appendix D.

7 Ethical considerations

This paper presents work whose goal is to advance the field of Machine Learning. Since our work concerns context-to-parameter mapping and involves lossy compression mechanisms described in Section 6, which may impact factual recall in entity-intensive tasks, we expect our work to be used under the guidance of human values.

8 Acknowledgements

We thank the anonymous reviewers for their insightful suggestions. All computational resources for this work were supported by the Basic Model Technology Center, WeChat AI, Tencent Inc., China.

References

Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. 2009. The fifth PASCAL recognizing textual entailment challenge.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Haolong Chen, Hanzhi Chen, Zijian Zhao, Kaifeng Han, Guangxu Zhu, Yichen Zhao, Ying Du, Wei Xu, and Qingjiang Shi. 2025a. An overview of domain-specific foundation model: key technologies, applications and challenges. *Science China Information Sciences*, 69(1).

Tong Chen, Hao Fang, Patrick Xia, Xiaodong Liu, Benjamin Van Durme, Luke Zettlemoyer, Jianfeng Gao, and Hao Cheng. 2025b. Generative adapter: Contextualizing language models in parameters with a single forward pass. In *The Thirteenth International Conference on Learning Representations*.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. 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)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Tri Dao and Albert Gu. 2024. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *CoRR*, abs/2405.21060.

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The CommitmentBank: Investigating projection in naturally occurring discourse.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *Transactions on Machine Learning Research*.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Neil Houlsby, Andrei Giouvanos, Zornitsa Kozareva, Möhrmann Wei, Michael Hall, Romal Thoppilan, and Google Sung. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online. Association for Computational Linguistics.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023. LLMLingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13358–13376, Singapore. Association for Computational Linguistics.

Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. 2024a. Revisiting catastrophic forgetting in large language model tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4297–4308, Miami, Florida, USA. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. 2024b. SnapKV: LLM knows what you are looking for before generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Kangtao Lv, Haibin Chen, Yujin Yuan, Langming Liu, Shilei Liu, Yongwei Wang, Wenbo Su, and Bo Zheng. 2025. How to inject knowledge efficiently? knowledge infusion scaling law for pre-training large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26204–26219, Suzhou, China. Association for Computational Linguistics.

Roemmele Melissa, Bejan Cosmin Adrian, and Gordon Andrew S. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. 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*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Dilxat Muhtar, Yelong Shen, Yaming Yang, Xiaodong Liu, Yadong Lu, Jianfeng Liu, Yuefeng Zhan, Hao Sun, Weiwei Deng, Feng Sun, and 1 others. 2024. Streamadapter: Efficient test time adaptation from contextual streams. *arXiv preprint arXiv:2411.09289*.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, and Timothy P. Lillicrap. 2019. Compressive transformers for long-range sequence modelling. *Preprint*, arXiv:1911.05507.

Socher Richard, Perelygin Alex, Jean Wu, Chuang Jason, Manning Christopher D., Ng Andrew, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Jing Han Sun and Ali Emami. 2024. EvoGrad: A dynamic take on the Winograd schema challenge with human adversaries. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6701–6716, Torino, Italia. ELRA and ICCL.

Gido M van de Ven, Nicholas Soures, and Dhireesha Kudithipudi. 2024. Continual learning and catastrophic forgetting. *arXiv preprint arXiv:2403.05175*.

Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, Mosharaf Chowdhury, and Mi Zhang. 2024. Efficient large language models: A survey. *Transactions on Machine Learning Research*. Survey Certification.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Re, Clark Barrett, Zhangyang Wang, and Beidi Chen. 2023. H2o: Heavy-hitter oracle for efficient generative inference of large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

A Layer-wise Derivation for Input Shift

This appendix provides the first-order, single-layer derivation sketched in Section 3.1. We linearize a gated MLP layer and express the impact of an attention-derived input shift as low-rank weight perturbations.

Setup. For a gated MLP layer,

$$\mathbf{y} = \mathbf{W}_{down}\left(\phi(\mathbf{W}_{gate}\mathbf{x}) \odot (\mathbf{W}_{up}\mathbf{x})\right), \quad (14)$$

where ϕ is the activation (e.g., SiLU) and $\odot$ is the Hadamard product. Under ICL, the effective input is $\mathbf{x}' = \mathbf{x} + \boldsymbol{\delta}_{ctx}$ with $\boldsymbol{\delta}_{ctx}$ produced by attention.

First-order expansion. Expanding around $\mathbf{x}$ (ignoring $\mathcal{O}(\|\boldsymbol{\delta}_{ctx}\|^2)$):

$$\begin{aligned}\mathbf{h}(\mathbf{x}+\boldsymbol{\delta}_{ctx}) \approx{}& \phi(\mathbf{W}_{gate}\mathbf{x}) \odot (\mathbf{W}_{up}\mathbf{x}) \\ &+ \left[\phi'(\mathbf{z}_g) \odot (\mathbf{W}_{gate}\boldsymbol{\delta}_{ctx})\right] \odot (\mathbf{W}_{up}\mathbf{x}) \\ &+ \phi(\mathbf{z}_g) \odot (\mathbf{W}_{up}\boldsymbol{\delta}_{ctx}),\end{aligned} \quad (15)$$

where $\mathbf{z}_g = \mathbf{W}_{gate}\mathbf{x}$. Propagating to the output,

$$\Delta\mathbf{y} \approx \mathbf{W}_{down}\left(\Delta\mathbf{h}_{gate} + \Delta\mathbf{h}_{val}\right). \quad (16)$$

Each term is re-expressible as a rank-1 perturbation:

$$\Delta\mathbf{h}_{val} = \left[\mathrm{diag}(\phi(\mathbf{z}_g))\mathbf{W}_{up}\right]\boldsymbol{\delta}_{ctx}, \quad (17)$$

$$\Delta\mathbf{h}_{gate} = \left[\mathrm{diag}(\phi'(\mathbf{z}_g) \odot \mathbf{W}_{up}\mathbf{x})\mathbf{W}_{gate}\right]\boldsymbol{\delta}_{ctx}. \quad (18)$$

Thus

$$\begin{aligned}\mathbf{W}'_{up} &= \mathbf{W}_{up} + \mathbf{U}_{up}(\mathbf{x})\,\boldsymbol{\delta}_{ctx}^{\top}, \\ \mathbf{W}'_{gate} &= \mathbf{W}_{gate} + \mathbf{U}_{gate}(\mathbf{x})\,\boldsymbol{\delta}_{ctx}^{\top},\end{aligned} \quad (19)$$

with $\mathbf{U}_{up}, \mathbf{U}_{gate}$ determined by $\mathbf{W}_{down}$, ϕ, and ϕ'. This is a layer-wise, first-order approximation; multi-layer nonlinear accumulation is not assumed linear.

B Hyperparameter Settings

B.1 SADA Module Hyperparameters

Chunk size. We set the chunk size $C = 64$, matching the chunking operator used in Section 3.3 and the definition in Section 3.2.1 where evicted tokens are partitioned into fixed-size segments.

Learnable queries. We use $N = 16$ learnable queries ($\mathbf{Q}_{meta}$ in Section 3.3) for intra-chunk cross attention, yielding a fixed-size representation per chunk.

Low-rank head dimension. We set the per-head low-rank dimension $d_r = 16$ (corresponding to the low-rank projection used to form the meta key/value features). The resulting memory width is $d_{state} = H \cdot d_r$, where H is the number of attention heads of the backbone model.

B.2 Training Hyperparameters

Optimization. Unless stated otherwise, we train with AdamW (betas $(0.9, 0.95)$, $\epsilon = 10^{-6}$, weight decay 0) and a cosine learning-rate schedule with linear warmup (warmup steps $= 100$; total steps determined by the number of update steps). We clip gradients with a global norm of 1.0 and train in BF16 with FlashAttention-2 enabled.

Stage I (PG19 / SWSD). We train for 1 epoch with per-device batch size 1 and gradient accumulation steps 8 on 8 GPUs (global batch size 64). The learning rate is 6×10^{-5}.

Stage II (ICL/NLU / TAD). Starting from the PG19 checkpoint, we train for 3 epochs using the same per-device batch size 1, gradient accumulation steps 8, and 8 GPUs (global batch size 64). The learning rate is 6×10^{-5}.

C Dialogue Templates

This appendix reports the chat-style dialogue templates used in our ICL benchmarks. We group templates by Seen vs. Unseen tasks. Placeholders are denoted with angle brackets (e.g., `<premise>`).

C.1 Seen Tasks

User: `<sentence>`
Question: Is this sentence positive or negative?
Assistant: `positive | negative`

User: `<premise> therefore | because`
Assistant: `<choice>`

User: `<passage>`
Question: `<question?>`
Assistant: `yes | no`

User: `<premise>`
Question: `<hypothesis>`. True, False, or Neither?
Assistant: `True | False | Neither`

User: `<premise>`
Question: `<hypothesis>` True or False?
Assistant: `True | False`

C.2 Unseen Tasks

User: `<question>`
Please answer the question.
Assistant: `<option label>`

User: `<question>`
Please answer the question.
Assistant: `<option label>`

User: `<context>`
Assistant: `<option label>`

User: Answer the question: `<goal>`.
Assistant: `<option label>`

User: `<sentence with blank as _>`
Options: `A. <choice A>`
`B. <choice B>`
`...`
Assistant: `<label>. <choice text>`

C.3 QA Tasks

User: `<passages>`
Answer the question based on the given passages.
Only give me the answer.
Question: `<question>`
Assistant: `<answer>`

User: `<passages>`
Answer the question based on the given passages.
Only give me the answer.
Question: `<question>`
Assistant: `<answer>`

C.4 Summarization Tasks

User: `<report>`
Write a one-page summary of the report.
Assistant: `<summary>`

User: `<transcript>`
Answer the query in one or more sentences.
Query: `<query>`
Assistant: `<answer>`

Table 4: Practical training overhead and amortized inference benefit on **Qwen2.5-1.5B-Instruct** at **32K** context length.

Method	Upfront Training Time	Inference Latency / Sample	Speedup
ICL	0 (None)	16.64s	1.0×
SADA (Ours)	∼11 hours (Stage I + II)	3.84s	4.3× faster

D Training Overhead, Amortized Cost, and Complexity Analysis

This appendix provides additional analysis on the training overhead of SADA, which complements the discussion in Section 6. We report both a practical cost-benefit view based on measured wall-clock latency and a theoretical complexity comparison with standard prompt-based ICL.

D.1 Practical Overhead and Amortized Cost Analysis

A key trade-off of SADA is that it is not training-free: unlike standard ICL, it requires a one-time offline two-stage distillation procedure before deployment. In return, this offline cost enables substantially more efficient long-context inference by replacing explicit long-context KV-cache retention with parameterized compression.

Table 4 summarizes the practical overhead of SADA on **Qwen2.5-1.5B-Instruct**. In our setting, the full two-stage training procedure (Stage I + Stage II) takes approximately **11 hours** on **8×40G** GPUs. At inference time, however, SADA reduces the latency of a **32K-context** sample from **16.64s** under standard ICL to **3.84s**, corresponding to a **4.3× speedup**.

This trade-off is favorable in repeated long-context deployment settings. The latency saving per sample is

$$\Delta t = 16.64 - 3.84 = 12.8 \text{ seconds}.$$

Therefore, the one-time training investment is amortized after roughly

$$\frac{11 \times 3600}{12.8} \approx 3.1 \times 10^3$$

samples, i.e., around **3,000 samples**. This suggests that for production scenarios involving a large number of long-context queries, the offline training overhead can be offset by the reduced inference cost.

Accordingly, SADA is better suited to **fixed-task, high-frequency inference** settings than to **zero-shot scenarios requiring instant adaptation to arbitrary new contexts**. This trade-off is already reflected in our discussion of limitations in the main text.

D.2 Training and Inference Complexity Analysis

We now provide a high-level complexity analysis of SADA. The central idea is that SADA compresses evicted history into a fixed-size memory state and dynamically generated low-rank parameters, avoiding the linear growth of a global KV cache with respect to the full history length.

Notation. Let L denote the total context length, d the hidden dimension, C the chunk size, N the number of learnable memory queries, d_{state} the memory state dimension, and W_{local} the retained local context window during inference.

D.2.1 Inference Complexity

Memory complexity. Under standard prompt-based ICL, the model retains KV pairs for the full visible history, leading to memory usage that grows with context length:

$$\mathcal{M}_{\text{ICL}} = O(L \cdot d).$$

In contrast, SADA stores a fixed-size meta-memory state $\mathbf{M}_k \in \mathbb{R}^{N \times d_{state}}$ together with a bounded local window. Therefore, with respect to the evicted history length, the memory footprint of SADA is independent of L:

$$\mathcal{M}_{\text{SADA}} = O(N \cdot d_{state}) + O(W_{local} \cdot d).$$

Since N, d_{state}, and W_{local} are fixed by design, SADA avoids the unbounded memory growth of full-history KV caching.

Time complexity. We decompose inference over a sequence of length L into three components.

(i) History compression. The evicted history is partitioned into L/C chunks. For each chunk, SADA computes

$$\mathbf{U}_k = \text{CrossAttn}(\mathbf{Q}_{meta}, \mathbf{K}_k, \mathbf{V}_k).$$

The cost per chunk is

$$O(N \cdot C \cdot d),$$

so over all chunks the total compression cost is

$$\frac{L}{C} \cdot O(N \cdot C \cdot d) = O(L \cdot N \cdot d).$$

Because N is fixed, this is linear in L.

(ii) State evolution. The memory state is updated recurrently as

$$\mathbf{M}_k = \gamma_k \odot \mathbf{M}_{k-1} + \mathbf{U}_k.$$

The cost per chunk is

$$O(N \cdot d_{state}),$$

which leads to total cost

$$\frac{L}{C} \cdot O(N \cdot d_{state}) = O(L).$$

(iii) Parameter generation and local-window decoding. The dynamic low-rank parameters are generated from the memory state once per chunk:

$$\mathbf{B}_k = \mathbf{M}_k^\top \mathbf{W}_{MetaB},$$

which costs

$$O(d_{state} \cdot N \cdot d_{out}).$$

This term is constant with respect to L. The frozen backbone then processes only the retained local window W_{local} together with the injected update, rather than the full history.

Taken together, SADA processes the full sequence with total cost approximately linear in context length:

$$\mathcal{T}_{\text{SADA, total}} = O(L),$$

up to constant factors depending on the backbone and local window size. By contrast, standard full-context Transformer inference over a sequence of length L incurs quadratic total attention cost:

$$\mathcal{T}_{\text{ICL, total}} = O(L^2).$$

At a high level, SADA therefore trades repeated full-history attention for one-pass history compression plus fixed-window decoding.

Per-step interpretation. During decoding, standard ICL becomes increasingly expensive as the visible context grows, since the model must repeatedly attend to and move a larger KV cache. In contrast, SADA maintains a bounded local window and a fixed-size compressed state, so the effective per-step cost remains stable with respect to the evicted history length.

D.2.2 Training Complexity

SADA uses a teacher-student distillation scheme. This introduces a one-time offline cost, but keeps the backbone frozen and therefore greatly reduces the trainable state compared with full fine-tuning.

Stage I: Sliding Window State Distillation (SWSD). Stage I contains two computation paths:

1. **Teacher path:** the frozen full-context model generates oracle hidden states and logits. Its dominant cost follows full-context Transformer attention, i.e., $O(L^2)$.

2. **Student path:** SADA compresses history and runs the frozen backbone with truncated context and injected low-rank updates. Its sequence cost is approximately linear in L, i.e., $O(L)$.

The overall compute in Stage I is therefore dominated by the teacher forward pass.

Stage II: Task Alignment Distillation (TAD). Stage II uses the "2-forward-1-backward" strategy described in the main text:

1. an oracle forward without gradients on the full context,

2. a SADA student forward with compressed context,

3. a backward pass through the mapping module only.

Again, the dominant cost comes from the teacher forward on the full context, while the student-side compression and local-window decoding remain substantially lighter.

Training memory considerations. A key practical advantage of SADA is that the backbone remains frozen during both stages. As a result, optimizer states and gradient updates are required only for the mapping module parameters, rather

Table 5: High-level comparison of inference and training characteristics.

Metric	Standard ICL	SADA (Ours)	Main Effect
Inference memory	$O(L \cdot d)$	$O(Nd_{state}) + O(W_{local}d)$	No unbounded history growth
Inference total cost	$O(L^2)$	$O(L)$	More efficient long-context processing
Effective decode cost	grows with context length	stable wrt evicted history	Faster long-context decoding
Upfront training cost	none	one-time offline cost	amortized in repeated deployment
Trainable state	N/A	$O(\lvert\theta_{\text{SADA}}\rvert)$	Much smaller than full FT

than for the full LLM. Let $|\theta_{\text{SADA}}|$ denote the number of trainable mapping-module parameters and $|\theta_{\text{LLM}}|$ denote the backbone parameter count, with $|\theta_{\text{SADA}}| \ll |\theta_{\text{LLM}}|$. Then the trainable-state overhead of SADA scales with

$$O(|\theta_{\text{SADA}}|),$$

rather than

$$O(|\theta_{\text{LLM}}|),$$

as in full fine-tuning. In practice, this substantially lowers peak training memory relative to end-to-end long-context fine-tuning, although teacher activations and student hidden states still contribute non-negligible runtime memory.

D.2.3 Summary

Table 5 summarizes the comparison between standard ICL and SADA.

Overall, SADA shifts part of the cost of long-context reasoning from online inference to offline compression and distillation. This shift is disadvantageous for single-use or fully open-ended zero-shot settings, but can be favorable when the same long-context task pattern is executed repeatedly and inference efficiency is critical.

E Additional Analysis of Conceptual vs. Entity-Intensive Tasks

This appendix provides additional analysis of the task boundary discussed in Section 2. In the main text, we observed that SADA is more effective on concept-oriented tasks such as summarization, while remaining weaker on entity-intensive tasks such as multi-hop QA. To better characterize this gap, we analyze (i) hidden-state fidelity, (ii) the effect of low-rank capacity, and (iii) qualitative error patterns.

E.1 Hidden-State Fidelity: Summarization vs. QA

To better understand why SADA struggles more on entity-intensive tasks, we compare the layer-wise hidden-state discrepancy between SADA and the full-context oracle on a summarization task (**GovReport**) and a multi-hop QA task (**HotpotQA**) using **Qwen2.5-1.5B-Instruct**. We report the mean squared error (MSE) averaged over three layer regions.

As shown in Table 7, the two task types remain relatively close in early layers, but a clearer divergence emerges in middle and deep layers. The absolute MSE gap increases from **0.11** in early layers to **0.25** in middle layers and further to **1.46** in deep layers. This pattern suggests that SADA can preserve broad semantic structure reasonably well, but becomes less faithful when the task requires precise entity tracking and retrieval over multiple reasoning steps.

This observation is consistent with the hidden-state-collapse phenomenon discussed in the main text: once the intermediate trajectory begins to deviate in layers associated with retrieval and composition, the mismatch accumulates and becomes harder to recover in later layers.

E.2 Capacity Analysis: The Impact of LoRA Rank

We further hypothesize that part of the entity-retrieval gap comes from the low-rank bottleneck in the generated parameter update $\Delta\mathbf{W} = \mathbf{AB}$. If the rank is too small, the compressed parameterization may preserve coarse semantic patterns while discarding fine-grained entity details. To test this hypothesis, we vary the LoRA rank r from 8 to 64 and evaluate HotpotQA accuracy on **Qwen2.5-1.5B-Instruct**.

Table 8 shows a monotonic improvement as the rank increases. In particular, scaling from the default setting $r = 16$ to $r = 64$ yields a **3.08-point** gain in HotpotQA accuracy. This supports the interpretation that the low-rank constraint contributes to the loss of fine-grained entity information: increasing rank provides additional capacity and partially closes the gap to full-context ICL.

At the same time, larger ranks also increase training cost and memory overhead. We therefore

Table 6: Representative bad cases from the main QA benchmarks, HotpotQA and 2WikiMQA, with minimal supporting evidence. SADA often remains semantically on target, but fails to recover the exact answer instance.

Dataset	Question	Context Evidence	Gold	SADA
2WikiMQA	What is the date of death of the director of film *Nallavan Vazhvan*?	"*Nallavan Vazhvan* ... was produced and directed by **P. Neelakantan**." "Palaniyaandi Neelakantan (2 October 1916 – **3 September 1992**) was a Tamil film director."	3 September 1992	16 January 2013
2WikiMQA	Who is Sir William Gore, 3rd Baronet's paternal grandfather?	"Sir William Gore, 3rd Baronet ... was the oldest son of **Sir Ralph Gore, 2nd Baronet**." "Sir Ralph Gore, 2nd Baronet ... was the eldest son of **Sir Paul Gore, 1st Baronet**."	Sir Paul Gore, 1st Baronet	William Gore
HotpotQA	1956 Night Series Cup games were played at what inner suburb of Melbourne that is 3km south of Melbourne's central business district?	"The 1956 VFL Night Premiership Cup ... games being played at the Lake Oval, **Albert Park**." "**Albert Park** is an inner suburb of Melbourne ... 4 km south of Melbourne's Central Business District."	Albert Park	the Melbourne Cricket Ground

Table 7: Layer-wise hidden-state MSE between SADA and the full-context oracle on summarization and QA.

Layer Region	GovReport	HotpotQA	QA / Sum Ratio
Early (L0–10)	3.13	3.24	1.04×
Middle (L11–20)	4.27	4.52	1.06×
Deep (L21–28)	10.68	12.14	1.14×

Table 8: Impact of LoRA rank on HotpotQA accuracy.

LoRA Rank r	Accuracy (%)	Relative Gain
8	19.24	-4.62
16 (default)	23.86	–
32	25.29	+1.43
64	26.94	+3.08

view the default configuration as a compromise between efficiency and fidelity: lower-rank updates retain the lightweight nature of SADA, but at the cost of more aggressive information compression on entity-heavy tasks.

F Qualitative Bad Case Analysis

To complement the aggregate results in the main text, we provide a small qualitative analysis of representative failures from our main QA benchmarks, **HotpotQA** and **2WikiMQA**. A recurring pattern is that SADA often predicts the correct *answer type* while missing the exact *instance*. This behavior is particularly visible on questions that require precise recovery of dates, years, or entity mentions from multi-hop supporting context.

The three cases cover different answer types: a date, a person in a kinship chain, and a location. In the first case, SADA outputs another plausible date rather than the correct death date. In the second, it stays within the same family/name cluster but selects the wrong individual. In the third, it retrieves a related sports venue instead of the precise suburb asked by the question. These are not off-topic hallucinations: the model remains aligned with the question's semantic field, but fails to preserve the precise factual binding required for exact multi-hop retrieval.

These cases support the interpretation developed in Section E. Parameterized context compression appears sufficient for preserving coarse semantic structure, expected answer type, and topical relevance, but it is less reliable when the task depends on exact factual bindings. By contrast, full-context ICL retains explicit access to the original supporting tokens, making it better suited to questions whose correctness depends on exact dates, years, and entity identities rather than approximate semantic matching.