

EPIC: Efficient Position-Independent Context Caching for Serving Large Language Models

Junhao Hu^{1*}, Wenrui Huang², Haoyi Wang¹, Weidong Wang², Tiancheng Hu¹, Qin Zhang³, Hao Feng³, Xusheng Chen³, Yizhou Shan^{3†}, Tao Xie^{1†}

¹Peking University, ²Nanjing University, ³Huawei Cloud

Abstract

Large Language Models (LLMs) are critical for a wide range of applications, but serving them efficiently becomes increasingly challenging as inputs become more complex. **Context caching** improves serving performance by exploiting **inter-request dependency** and **reusing key-value (KV) cache across requests**, thus **improving time-to-first-token (TTFT)**. However, existing **prefix-based context caching** requires **exact token prefix matches**, limiting cache reuse in few-shot learning, multi-document QA, or retrieval-augmented generation, where prefixes may vary.

In this paper, we present **EPIC**, an LLM serving system that introduces **position-independent context caching (PIC)**, enabling **modular KV cache reuse regardless of token chunk position (or prefix)**. EPIC features two key designs: **AttnLink**, which leverages **static attention sparsity** to **minimize recomputation** for accuracy recovery, and **KVSplit**, a customizable **chunking method** that **preserves semantic coherence**. Our experiments demonstrate that EPIC delivers up to **8× improvements in TTFT** and **7× throughput** over existing systems, with negligible or no accuracy loss. By addressing the limitations of traditional caching approaches, EPIC enables more scalable and efficient LLM inference.

1 Introduction

Large Language Models (LLMs)¹ have significantly advanced the process of Artificial General Intelligence (AGI) and are now fundamental to various emerging applications such as question-answering, chatbots, education, and medicine [34]. In addition, LLMs offer a user-friendly interface through “prompts” that consist of concatenated chunks of tokens (text-based words), such as document chunks, few-shot example chunks, or question chunks to instruct LLMs to complete specific tasks. Since LLMs’ increasing capability enables them to tackle a wider range of tasks, their usage pattern has transitioned **from simple chats to multi-turn planning**, reasoning, tool usage, few-shot learning, etc. **This shift resulted in long prompts with repeated tokens**, such as system messages, few-shot learning examples, documents in RAG, etc., whose content is **less changing (static) than user-specific instructions (dynamic)**.

^{*}This work was completed during his internship at Huawei Cloud.

[†]Co-corresponding authors.

¹We only focus on transformer-based LLM models in this paper.

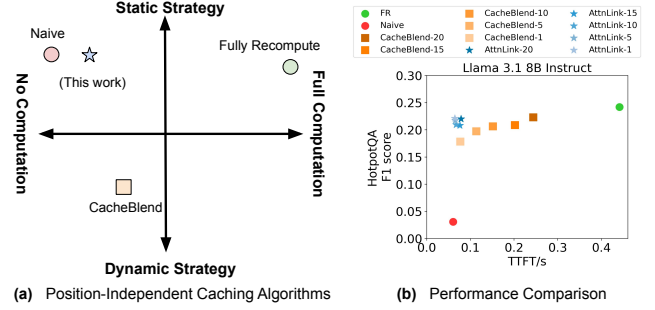


Figure 1. (a) presents position-independent context caching’s design space and four existing algorithms. *Naive* concatenates the pre-generated KV cache of different token chunks and does not recompute any tokens, achieving the lowest time-to-first-token (TTFT) and the lowest accuracy. *Fully Recompute (FR)* recomputes all tokens. *CacheBlend-15* select (at runtime) around 15% tokens to recompute, utilizing dynamic attention sparsity. *AttnLink-20* (this work) select (when offline) 20 tokens on each chunk boundary to recompute, utilizing static attention sparsity. (b) shows that our algorithm outperforms others at both accuracy and *TTFT* while running Llama3.1-8B with *HotpotQA* dataset (see eval for more results).

Context Caching (CC), also known as prompt caching, is an emerging optimization that enables the reuse of intermediate results (i.e., KV Cache) of static tokens in previous requests and accelerates inference by improving time-to-first-token (TTFT). Existing LLM inference systems either **enable CC implicitly**, which is transparent to users [11, 18, 32], or offer **explicit CC APIs** for users to **proactively control the lifecycle of KV cache** [8, 17].

Almost all CC systems are prefix-based. Prefix-based CC systems work by reusing the KV cache when the requests share the same initial sequence of tokens (the prefix). This approach is limited because it **only applies when there is an exact match in the prefix tokens across requests**. This limitation means that **if the prefix differs** (e.g., the order of token chunks at the beginning changes), **the entire KV cache cannot be reused**, resulting in full recomputation. This inefficiency can be particularly problematic in cases where **many requests share a substantial amount of overlapping context** but **differ slightly in chunk orders or their initial tokens** (e.g., in RAG or few-shot learning).

Inspired by classical position-independent code that can be executed at any memory address without modification, we propose a new approach for modular context caching called **position-independent caching (PIC)**. PIC allows for the modular reuse of KV cache, enabling any chunk to be reused at arbitrary positions within a prompt, not limited to the prefix. However, the primary challenge with this approach is **accuracy degradation**, which arises from violating the attention mechanism in transformer models.

In this paper, we address one question: **How do we split, generate, store, retrieve, and link** (concatenate and recompute some tokens to recover accuracy) **position-independent KV cache** to **avoid significant accuracy degradation** in the subsequent inference process (Figure 2)?

To the best of our knowledge, CacheBlend [27] is the first attempt to address the PIC problem (Figure 1 (a)) but it has two key limitations. First, the **time and resource complexity of the recomputation** is the same as the original attention mechanism — $O(N^2)$, where N is the number of tokens in the prompt (prompt length). As shown in Figure 1 (b), although CacheBlend-15 dynamically selects approximately 15% of tokens for recomputation, for very long prompts, common in many applications today, this $O(15\%N^2)$ complexity remains slow and prone to out-of-memory (OOM) errors (Figure 8). Second, CacheBlend **only offers a fixed-size, 512-token chunk-splitting method**, which does not generalize well across all different scenarios. Sometimes, it disrupts semantic coherence, as documents are arbitrarily split into 512-token chunks.

In a new attempt to address the PIC problem and overcome the limitations of existing approaches, we propose EPIC² that introduces two key designs. First, EPIC incorporates a novel algorithm, **AttnLink**, that **reduces recomputation complexity** to $O(kN) \sim O(N)$, where $k \ll N$ and increases with chunk numbers instead of N . Achieving this reduced complexity is challenging, as it requires significantly less computation than CacheBlend while maintaining comparable accuracy. We accomplish this by discovering that the **initial tokens of each pre-generated KV cache absorb a disproportionate amount of attention**, preventing subsequent tokens from attending to relevant parts. This phenomenon is known as **Attention Sink** [26]. We propose to **recompute k tokens ($k < 20$) on each chunk boundary**, crippling the attention-absorbing ability of “initial” tokens in each chunk. As shown in Figure 1, **AttnLink** improves TTFT by up to 3x while maintaining comparable accuracy. Second, EPIC **enhances generalizability** across diverse scenarios by incorporating a customizable **chunk-splitting** method, **KVSplit**. Our results indicate that **preserving semantic coherence within each chunk is critical** for maintaining high accuracy with long documents. For shorter documents, fixed-size chunks avoid semantic issues while providing efficiency (Figure 9).

We implement a system EPIC that integrates the **AttnLink** algorithm with a customizable splitting component, **KVSplit**, built on one of the most widely used inference frameworks, vLLM [11]. We evaluate EPIC against the state-of-the-art CacheBlend [27] across six tasks with distinct characteristics and three model architectures with diverse training recipes. The results show that EPIC achieves up to a 3× improvement in TTFT with accuracy losses limited to within 7%, compared to CacheBlend (Figure 1). Furthermore, EPIC provides up to an 8× reduction in TTFT and a 7× increase in throughput when serving multiple requests under varying rates, outperforming CacheBlend.

In summary, this paper makes three major contributions:

- We explicitly propose the PIC problem, consolidate the existing literature within this problem framework, and highlight potential directions for future research.
- We propose **AttnLink** together with customizable splitting method **KVSplit** that reduces up to 3× TTFT in synchronous workloads while keeping accuracy losses limited to within 7%, compared to the state-of-the-art CacheBlend.
- We implement the EPIC system that incorporates openAI-compatible context caching APIs, a KV store, **AttnLink** together with **KVSplit**. EPIC reduces up to 8× TTFT and increases up to 7× throughput when serving multiple requests under varying rates.

2 Background and Motivation

This section first presents a primer on transformers and then discusses the current landscape of context caching. We then highlight the limitations of the existing position-independent caching that motivates our work.

2.1 Transformer Primer

The attention-based transformer architecture [24] underpins most LLMs used today. A typical LLM has multiple transformer layers, each with an attention module. The attention can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. Specifically, when attention layers receive a sequence of hidden states as $(x_1, x_2, \dots, x_n) \in \mathbb{R}^{n \times d}$, where n is the length of the sequence and d is the dimension of key and value vectors, the attention layer will first compute the key, query and value vectors of each position:

$$k_i = W_k x_i, q_i = W_q x_i, v_i = W_v x_i.$$

where W_k , W_q , and W_v are learnable weights to compute K, Q, and V. We then calculate the attention score between tokens:

$$a_{ij} = \frac{\exp(q_i^\top k_j / \sqrt{d})}{\sum_{t=1}^i \exp(q_i^\top k_t / \sqrt{d})}$$

²Efficient Position-Independent Caching (EPIC)

Table 1. Context Caching Categorization. Implicit APIs refer to caching mechanisms managed by the inference engine or service provider, where the process is transparent to users. In contrast, explicit APIs give users direct control over the caching process. Prefix-based caching requires an exact match of tokens to reuse cached results, whereas position-independent caching allows the reuse of the KV cache even when the tokens differ.

	Prefix-Based	Position-Independent
Implicit	vLLM, OpenAI	CacheBlend
Explicit	Kimi, Gemini	EPIC (this work)

Finally, the attention model computes the output with value vectors as weight:

$$o_i = \sum_{j=1}^i a_{ij} v_j$$

The output of attention layers can be the input of the next attention layer or just the final output.

Autoregressive Generation & KV Cache LLMs generate tokens autoregressively, meaning they predict one token at a time based on the given input. This process is divided into two phases: the prefill phase and the decode phase. LLMs use KV (Key-Value) cache [20] to speed up generation by storing the attention module’s key and value vectors as described above. In the **prefill** phase, LLMs process the entire input prompt (x_1, x_2, \dots, x_n), computing and caching the key and value vectors for each token. Since this phase involves calculating vectors for the entire sequence, it can be slow for long inputs, and the time taken to generate the first token is measured by the Time-to-First-Token (TTFT) metric. In the **decode** phase, LLMs generate one token per step. The model calculates the probability of the next token x_{n+1} , selects the most likely one, and appends its key and value vectors to the KV cache. This reuse of cached vectors speeds up token generation, as the model only processes the new token instead of recomputing vectors for the entire sequence.

2.2 Context Caching

Since LLMs’ increasing capability enables them to tackle a wider range of tasks, their usage pattern has transitioned from simple chats to multi-turn planning, reasoning, tool usage, few-shot learning, etc. This shift resulted in long prompts with repeated tokens, such as system messages, few-shot learning examples, static documents in RAG, etc.

Context caching (CC), also known as prompt caching, is a technique that exploits inter-request dependency to reuse KV cache across inference requests to avoid repeated computation of the same prompt tokens [9, 14, 32]. Thus, it speeds up the prefill phase when LLMs are used in the above scenarios. In Table 1, we categorize existing CC approaches along two key dimensions. The first dimension addresses the level

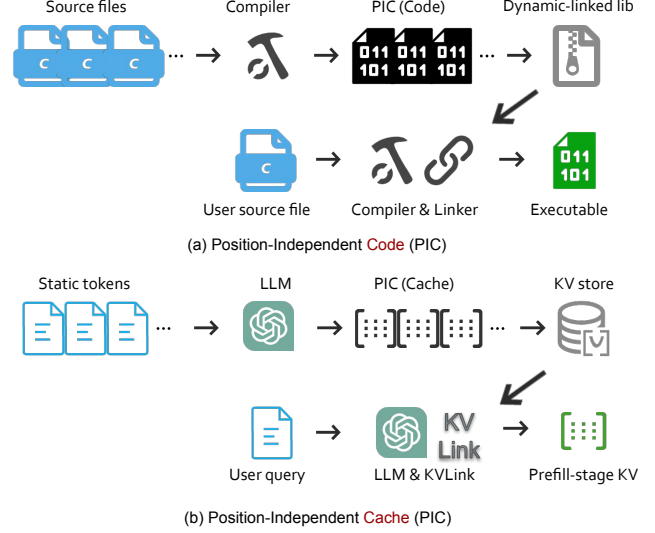


Figure 2. We draw an analogy between position-independent code and position-independent cache to illustrate that KV Cache can be reused across inference requests regardless of its location in the prompt.

of abstraction: caching can either be transparently managed by inference engines, as seen in vLLM [11], SGLang [32], and OpenAI’s prompt cache [18], or it can be explicitly controlled by users, as in Moonshot Kimi [17] and Google Gemini [8]. The second dimension concerns how KV cache are reused. The predominant method across systems is prefix-based caching, where only the initial portion of the sequence is cached and reused. A recent work called CacheBlend [27] introduces a Position-Independent Caching (PIC) method via implicit abstraction. Below, we examine both prefix-based and position-independent caching mechanisms in detail.

2.2.1 Prefix-Based Caching. Prefix-based caching (formerly known as prefix caching) is the current implementation of CC. It works by reusing the KV cache when the requests share the same initial sequence of tokens (the prefix). Almost all existing CC offerings and designs are prefix-based [5, 11, 17, 18, 32].

This approach, however, is limited because it only applies when there is an exact match in the prefix tokens across requests. This limitation means that if the prefix differs slightly (e.g., one or two words are different at the beginning), the entire cache cannot be reused, forcing the model to recompute all the key-value pairs for that request. This inefficiency can be particularly problematic in cases where many requests share a substantial amount of overlapping context but differ slightly in their initial tokens (e.g., in RAG). As a result, this constraint leads to redundant calculations, even when most of the token sequence could be reused.

2.2.2 Position-Independent Caching. Inspired by the classical position-independent code that can be executed at any memory address without modification [25], we propose a new approach for context caching called position-independent caching (PIC). This approach addresses the limitations of prefix-based caching by enabling KV cache reuse across requests, even when token sequences differ, without being constrained by the shared prefix or the point of divergence in the sequence.

The position-independent caching approach consists of four key components, closely aligned with steps in positional-independent code compiling (Figure 2):

1. **KVSplit:** An optional step that splits a long prompt into independent chunks (e.g., large documents used in RAG). This is akin to splitting C code into files for better maintainability.
2. **KVGen:** Feeds these chunks into the LLM to generate individual KV cache, much like precompiling C files into position-independent code.
3. **KVStore:** Saves the generated KV cache in a storage system for reuse, similar to storing compiled code in static or dynamic libraries.
4. **KVLink:** Retrieves KV cache, concatenates them, and selectively recomputes some tokens to form the KV cache of the prefill stage. This is analogous to linking PICs with source files to create a working executable.

2.3 Limitations of Existing Position-Independent Context Caching

The position-independent approach was first explored in a recent work called CacheBlend [27], which reuses precomputed KV cache regardless of token position by selectively recomputing a small subset of tokens. However, CacheBlend has two notable limitations. First, the time and resource complexity of the KVLink process in CacheBlend remains as high as the original attention mechanism, with a complexity of $O(N^2)$, where N is the number of tokens in the prompt (i.e., prompt length). Although CacheBlend recomputes roughly 15% of tokens to reconstruct the KV cache and maintain accuracy, for very long prompts—common in many modern applications—this $O(15\%N^2)$ complexity is still too slow and can lead to out-of-memory (OOM) errors (as shown in Figure 8). CacheBlend also consumes significant resources to reconstruct portions of the concatenated KV cache, aiming to make them numerically close to fully recomputed cache values. However, we argue that this reconstruction is unnecessary for preserving accuracy. In model quantization, for example, the KV cache of fp16 models and their int8/int4 counterparts can differ, yet maintain similar accuracy levels. Inspired by the sparsity characteristics of attention [26], our work introduces an algorithm that only recomputes k tokens on each chunk boundary — potentially as few as 1% [31] —

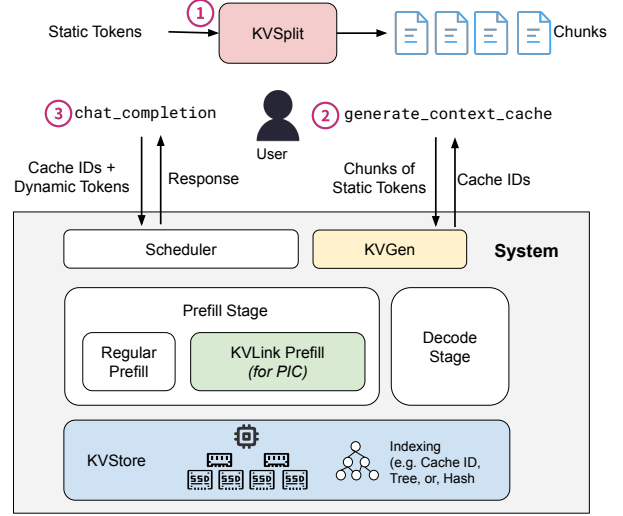


Figure 3. The Architecture of EPIC Serving System.

which reduces the overall complexity to $O(kN)$, significantly improving efficiency without sacrificing accuracy.

Second, CacheBlend’s approach to maintaining accuracy is limited by its rigid fixed-length chunking strategy, which does not generalize well across different scenarios. By splitting documents into fixed-length chunks, CacheBlend disrupts the semantic coherence within each chunk. In this work, we propose a customizable splitting method that adjusts dynamically based on the content, preserving each chunk’s semantic independence. Our results show that the customizable splitting component is critical for achieving consistently high accuracy across various applications.

3 System Overview

We present EPIC’s overall architecture in Figure 3. EPIC is an inference serving system with a scheduler that processes requests containing a prompt or both a prompt and cache IDs, a prefill execution stage including a regular prefill and a newly proposed KVLink component for handling position-independent caching, a decode execution stage, a newly proposed KVGen component that manages context caching requests, and a KVStore that stores historical KV cache that are reused across multiple inference requests.

EPIC is structured around the following user workflow. First, ① users optionally employ the KVSplit component to divide static tokens (e.g., documents, system prompts) into semantically independent chunks. Second, ② users send a request with these static token chunks, and the request is handled by KVGen. KVGen generates KV cache for the static tokens using an ordinary prefill component, stores the KV cache in KVStore, and returns the corresponding cache IDs. Third, ③ users send a request with dynamic tokens (e.g., user-specific information or instructions) and cache IDs, and the request is handled by the scheduler. The scheduler executes

the prefill stage using the *KVLink* component to generate the first token, followed by the decoding stage to generate the remaining tokens. When called, the *KVLink* retrieves the corresponding KV cache associated with the cache IDs from *KVStore*, concatenates them, and selectively recomputes tokens to form the KV cache for the prefill stage.

In the rest of the paper, we will deep dive into the *KVLink* component on how to realize position-independent caching. Crucially, we will introduce a new algorithm called *AttnLink* specifically designed for PIC.

4 Design of *KVLink*

This section discusses the system design and algorithms for PIC. Assuming that users have already generated and stored the context cache of static tokens in the *KVStore*, the *KVLink* component operates by taking the KV cache of these static tokens, concatenating them in any order (e.g., based on the sequence specified by users in their message list), appending dynamic tokens, and selectively recomputing a subset of tokens to generate the full KV cache during the prefill stage. *KVLink* introduces a new prefill stage design, including algorithms tailored for selective recomputation, as shown below the dashed line in Figure 4. Below, we will first describe the existing algorithms and then present our proposed algorithm, *AttnLink*.

4.1 Existing Algorithms

Figure 4 presents three algorithms that can be used to implement position-independent caching; each strikes a unique trade-off between accuracy and computational overhead.

1. *Naive*: This algorithm does not recompute any static tokens, except those new dynamic tokens in the user query (as the KV cache for these tokens has not been pre-generated). While this approach minimizes recomputation overhead, the attention score map shows that this algorithm prevents the user query from attending to relevant parts of the context. Most attention scores are concentrated in the initial tokens of each chunk [26], resulting in the lowest accuracy.
2. *Fully Recompute (FR)*: This algorithm recomputes all tokens as if no KV cache were reused. This method can achieve the highest accuracy as it allows for full attention at the cost of the highest computation overhead (i.e., disable context caching).
3. *CacheBlend* [27]: This algorithm selectively recomputes approximately 15% of the tokens, targeting numerical equivalence to *FR*, as observed in the attention map. It reduces computational overhead to 15% of the cost incurred by *FR* while maintaining accuracy within a 1% to 5% margin, as shown in Figure 6.

4.2 Design of *AttnLink* algorithm

To further reduce the overhead of *KVLink* while preserving accuracy, we introduce *AttnLink* (as shown in the last line in Figure 4). *AttnLink* is a novel recomputation algorithm that recomputes k tokens at each chunk boundary with $k/2$ tokens selected from the preceding chunk and $k/2$ from the following chunk. Here, $k \ll N$ and increases slowly with N , resulting in a time and resource complexity of recomputation of $O(kN) \sim O(N)$. In the following subsections, we address two key questions about *AttnLink*: 1) How do we recompute the selected set of tokens? 2) Why do we select k tokens on each chunk boundary?

4.2.1 How do we recompute the selected set of tokens?

Assuming that we have selected k' (k tokens from each chunk plus the user query) tokens on from a total of N (prompt length) tokens, we recompute them as follows. First, we obtain the embedding matrix E (with shape (k', d)) of the k' tokens, where d is the hidden size. Second, at layer i , we compute the new K , Q , and V matrices (each with shape (k', d)) for these k' tokens:

$$Q = EW_Q, K = EW_K, V = EW_V \quad (1)$$

where W_Q , W_K , and W_V are model parameters with shape $(d, d)^3$. Third, we expand the K and V matrices by incorporating the pre-generated KV cache of the $N - k'$ unselected tokens at correct positions, forming K_{exp} and V_{exp} (both with shape (N, d)). Fourth, we compute the attention matrix A (with shape (k', N)) by multiplying Q (with shape (k', d)) with K_{exp}^T (with shape (d, N)), allowing the k' tokens to attend to all N tokens:

$$A = \text{softmax}(QK_{exp}^T \cdot \text{MASK}) \quad (2)$$

where MASK assures that the k' tokens only attend to tokens before them. Finally, we multiply A (with shape (k', N)) with V_{exp} (with shape (N, d)) to obtain the output (or input to the next layer) with shape $((k', d))$:

$$O = AVW_O \quad (3)$$

where W_O is a matrix with shape (d, d) .

4.2.2 Why do we select k tokens on each chunk boundary?

We select k tokens on each chunk boundary based on the following three reasoning steps. First, the attention maps from the *Naive* approach indicate that most attention scores for the four decoded tokens are concentrated on the first few tokens of each chunk (illustrated by the four glowing yellow vertical lines). Since each chunk is generated independently, each of their initial tokens has a tendency to absorb attention from subsequent tokens — a phenomenon referred to as the attention sink [26]. As a result, later decoding tokens

³For notation simplicity, d represents all possible hidden dimension sizes, which may be further divided into the number of heads and head dimensions.

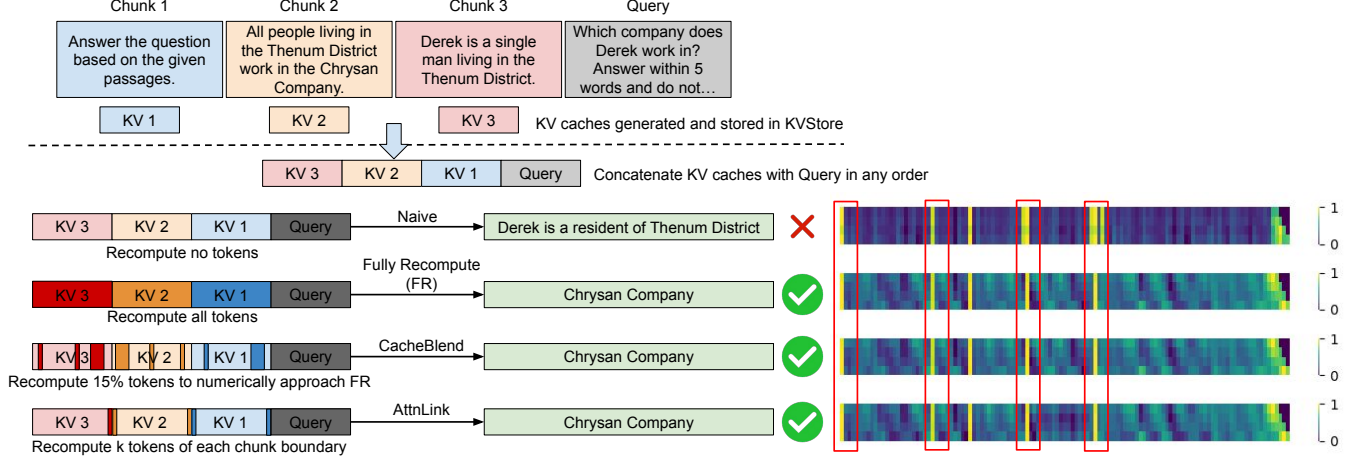


Figure 4. KVLink’s PIC Algorithms. Above the dashed line, the KV cache is assumed to be pre-stored. Below the dashed line, *KVLink* retrieves the KV cache, concatenates it, appends the user query, and selectively recomputes certain tokens. The comparison of four recombination algorithms is illustrated using light and dark shades: light represents pre-generated KV cache, while dark indicates the portion that requires recomputation. *Naive* performs no recomputation, relying entirely on the pre-stored cache. *Fully Recompute (FR)* recomputes the entire KV cache from scratch. *CacheBlend* selectively recomputes approximately 15% of the tokens. Finally, *AttnLink* limits recomputation to only a few tokens at chunk boundaries. The bottom right visualizes attention maps for four decoded tokens, with *Naive* outputting an unnecessarily long sequence for the first four decoded tokens. To highlight the differences between *Naive*’s attention map and those of the other algorithms, we normalize the QK^T results to the [0, 1] range using min-max scaling instead of Softmax.

struggle to allocate sufficient attention to the relevant context needed to address the user’s query (illustrated by the darkness except for the glowing yellow vertical lines).

Second, as shown by the attention map of *AttnLink*, by recomputing the first few tokens of each chunk along with their surrounding tokens, we can diminish their status as “initial” tokens to cripple their ability to absorb attention from subsequent decoding tokens. Consequently, the decoding tokens can then allocate sufficient attention to the relevant parts of the context (illustrated by the large area of lightness in *AttnLink*’s attention map).

Third, our observations indicate that among the 15% of tokens recomputed by *CacheBlend*, a significant proportion often includes the tokens at each chunk boundary.

Compared to *CacheBlend*, *AttnLink* recomputes a small, nearly constant number of tokens while preserving accuracy (Figure 6) for two main reasons. First, the ability to recompute only a limited number of tokens while maintaining effectiveness may be attributed to the utilization of attention sparsity [26, 31], which shows that a small subset of tokens (less than 1%) is essential for inference. Second, a deeper understanding and strategic use of attention mechanisms can outweigh the need for numerical equivalence, which often requires extensive recomputation that is not critical for achieving high accuracy. For instance, while the KV cache of an FP16 model and its INT8 counterparts may differ, they still achieve comparable accuracy.

5 Implementation

We implement EPIC based on vLLM 0.4.1 [11], with 2K lines of code in Python. We incorporate the four PIC recombination algorithms presented in Section 4. We port *CacheBlend* from their public repository⁴.

We implement *KVSplit* as a standalone component independent of EPIC. We provide two basic implementations: one splits tokens based on semantic boundaries (e.g., sentence, paragraph, or passage), while the other uses fixed-size chunks (e.g., 512 tokens). Users can choose the desired granularity of chunks and **optionally** employ *KVSplit*; they can also integrate external tools, such as Retrieval-Augmented Generation (RAG) for chunking. Improved chunking techniques correlate with enhanced accuracy (Figure 9) and greater reusability.

We implement the *KVStore* based on vLLM’s original memory management and prefix caching subsystem. We make the following changes. First, we add a cache-ID-based indexing mechanism using the sequence group ID as the cache ID. Second, since the original vLLM manages historical KV cache residing in HBM only, we extend it to include DRAM and local filesystem, akin to Mooncake [21]. Third, we modify the scheduler to retain block tables and memory for context caching requests. We also implement helper APIs that allow users to manage the lifecycle of KV cache, such as

⁴<https://github.com/YaoJiayi/CacheBlend>. Accessed on Sep 2024.

`expire_cache(cache_id)`. As building a highly efficient *KVStore* is not the core focus of this paper, we build a minimal working system. EPIC can use external *KVStore* systems like LMCache [15], which we leave for future work.

We implement *KVGen* as a standalone module that handles CC APIs. We implement *KVLink* as a parallel module of the original prefill. First, we adapt the model architecture to support masked attention across tokens scattered in different positions. Second, we modify the attention backends to handle data placement, movement, and the computational steps required by the preceding algorithm. These changes ensure efficient recomputation and accurate linking between static and dynamic tokens.

EPIC offers APIs similar to those in Kimi [17] and Gemini [7]. To create position-independent caching, users can call: `generate_context_cache(static_token_chunks[]) -> cache_ids`. This API accepts a list of static token chunks, pre-generates the corresponding KV cache, and returns their cache IDs. For PIC-accelerated LLM inference, users can call `chat_completion(message_list) -> Response`, where the message list can contain both new dynamic tokens and previously returned cache IDs, arranged in any order. This flexibility marks a key distinction from state-of-the-art prefix-based context caching approaches, such as those used in Kimi [17] and Gemini [8].

6 Evaluation

We highlight key research questions and our findings below:

- **RQ1.** What is the accuracy-latency trade-off of the *AttnLink* algorithm? *Takeaway:* *AttnLink* improves TTFT by 3x while maintaining accuracy loss within 0% to 7%, compared to CacheBlend.
- **RQ2.** What is the latency and throughput of the EPIC system under different workloads? *Takeaway:* EPIC improves TTFT by 8x and increases throughput by 7x compared to CacheBlend (Figure 7).
- **RQ3.** How does EPIC perform with very long contexts? *Takeaway:* TTFT grows nearly linearly with context length using *AttnLink*, compared to the quadratic growth observed with CacheBlend.
- **RQ4.** How does *KVSplit* affect the accuracy of EPIC? *Takeaway:* The splitting method choice significantly impacts accuracy and performance. *AttnLink* with semantic-based splitting remains on the Pareto frontier compared to CacheBlend with 512-token splitting.

6.1 Experiment Setup

In this section, we discuss the datasets, workloads constructed from these datasets, models, metrics, and the environment in which we carry out experiments.

6.1.1 Dataset. We select five datasets from LongBench [1] and create a synthetic *Needle in a haystack* [6] dataset. First, **LongBench** consists of 21 different tasks across 6 major

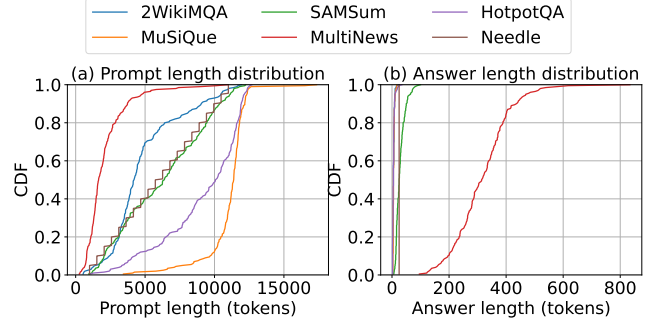


Figure 5. Prompt and answer length distribution.

categories, including ①multi-doc QA, ②multi-doc summarization, ③few-shot-examples-guided task, ④single-doc QA, ⑤multi-doc retrieval tasks, and ⑥code generation. The five datasets we select include those used in CacheBlend [27] — 2WikiMQA (①), MuSiQue (①), SAMSUM (③), and MultiNews (②) — plus HotpotQA (①), covering the first three categories. Second, *Needle in a Haystack* is popular for assessing LLMs’ understanding and retrieval ability from long contexts. It involves inserting one fact into different positions of unrelated contexts of varying lengths and asking the LLM to answer a question about this fact.

The rationale for selecting and excluding datasets is as follows. First, we include *HotpotQA* because it identifies the two passages containing the answer, aiding in detailed analysis. Second, we exclude ④single-document QA datasets, as ①multi-document QA adequately represents them. Third, we omit ⑥code generation datasets due to concerns over evaluation metrics (which focus on edit similarity rather than testing results). Fourth, for ⑤multi-document retrieval tasks, we substitute *Needle in a Haystack*, which is a more suitable benchmark for assessing retrieval capabilities.

All selected datasets contain 200 test cases, with total lengths ranging from 5k to 20k tokens, covering static tokens (e.g., documents, system prompts, few-shot examples), dynamic tokens (e.g., user-specific instructions), and answers. First, prompt length ranges from 0 to 20,000 tokens, with the majority falling between 1,000 and 12,000 tokens. Approximately 95% to 99% of the prompt consists of static tokens or contexts, while the dynamic tokens are less than 50 tokens. Second, answer length ranges from 0 to 900 tokens, with most answers being short (within 100 tokens), except for *MultiNews*, which requires summarization of documents.

6.1.2 Metrics. We use the following metrics to evaluate performance and model accuracy.

- **Time-To-First-Token (TTFT)** is the time from when users send a request (a prompt) to LLMs to when users receive the first token. This metric aims to measure the time taken in the prefill stage, which could be

reduced by solving the PIC problem. A smaller *TTFT* indicates a faster algorithm.

- *F1 score* [4] is used to evaluate *2WikiMQA*, *MuSiQue*, *HotpotQA* in LongBench, and *needle in a haystack*. This metric aims to evaluate the similarity between LLMs’ output and the ground-truth answer based on their common words. A higher *F1 score* indicates an algorithm with higher accuracy.
- *Rough-L score* [13] is used to evaluate *SAMSum* and *MultiNews* in LongBench. This metric aims to evaluate the similarity between LLMs’ output and the ground-truth answer by calculating the length of their longest common subsequence. A higher *Rough-L score* indicates an algorithm with higher accuracy.

6.1.3 Models. We evaluate our algorithm and system using three popular models: Mistral 7B Instruct [3], Llama 3.1 8B Instruct [16], and Yi Coder 9B Chat [28]. They are three of the most powerful open-source LLMs with diverse structures and training recipes. Instead of using the quantized version of larger models, we stick to smaller models that fit in our limited GPU resources.

6.1.4 Baselines. We compare *AttnLink* with the other three recomputation algorithms discussed in Section 4: *FR*, *Naive* and *CacheBlend* [27]. Additionally, we evaluate different variants of *CacheBlend*, denoted as *CacheBlend-r*, where *r* represents the ratio of tokens recomputed. Similarly, we test different variants of *AttnLink*, denoted as *AttnLink-k*, where *k* refers to the number of tokens recomputed at each chunk boundary.

6.1.5 Environment. We run experiments on a single NVIDIA DGX H800 server with one H800-80GB GPU available. It has 192-core Intel Xeon Platinum CPUs@2.4GHz and 2TB DRAM with 2 hyperthreading. We use Ubuntu 20.04 with Linux kernel 5.16.7 and CUDA 12.2.

6.2 Workloads

We construct the following two kinds of workflows.

6.2.1 Synchronous Workload. To address RQ1, we create a synchronous workload based on the selected datasets, where one request (prompt) is processed at a time. Each request is completed before initiating the next, providing a controlled environment to accurately measure the accuracy-latency trade-off without interference from concurrent requests. Latency under conditions of request interference will be studied using asynchronous workloads in RQ2.

For each request, we split its contexts into chunks using *KVSplit* and pre-generate the corresponding KV cache. Then, we send the request to the EPIC with the cache_ids of the pre-cached chunks and dynamic tokens. We record the *TTFT* and accuracy (*F1 score* or *Rouge-L score*) for each request.

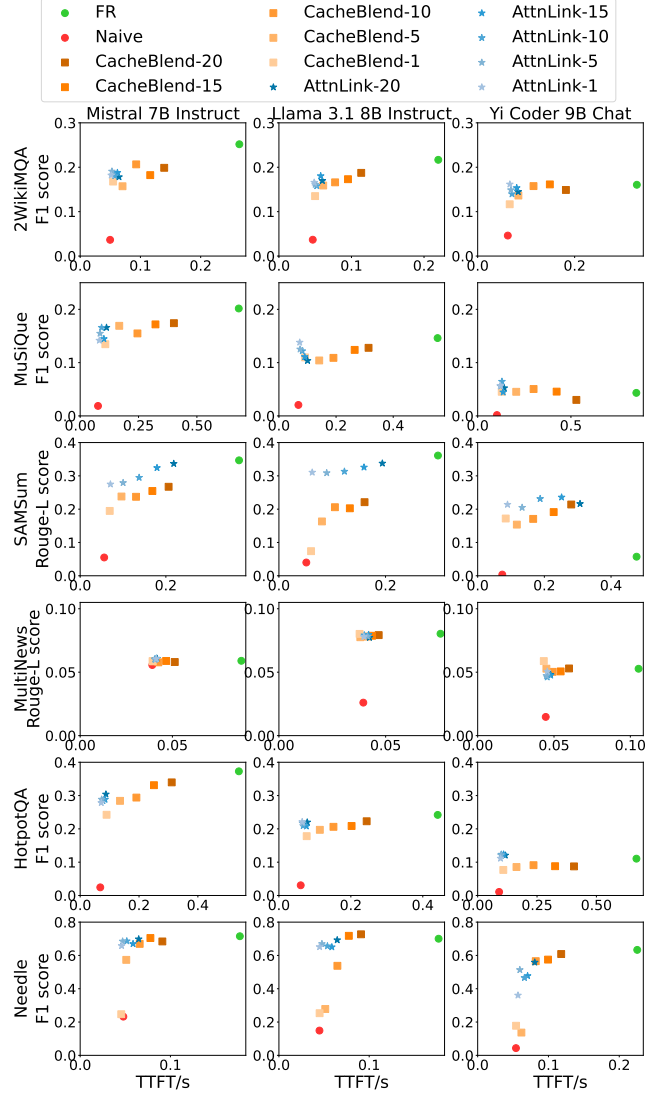


Figure 6. Trade-off between accuracy (*F1 score* or *Rouge-L score*) on y-axis and latency (*TTFT*) on x-axis. Each point indicates the average *TTFT* and accuracy for running synchronous workloads of one dataset (each row) on one model (each column) using one specific algorithm (each legend label), where the top-left corner of each subfigure represents the Pareto frontier. The *FR* baseline (green circle) achieves the highest accuracy but suffers from the worst latency, while the *Naive* baseline (red circle) provides the lowest latency but at the cost of the poorest accuracy. *AttnLink* (a series of gradient blue stars) establishes a new Pareto frontier, surpassing *CacheBlend* (a series of gradient orange rectangles). The *k* in *AttnLink-k* denotes the number of recomputed tokens on each chunk boundary, while the *r* in *CacheBlend-r* represents the ratio of recomputed tokens.

6.2.2 Asynchronous Workload. Context caching, introduced in late June 2024⁵, is a relatively new technique, and as such, lacks publicly available traces or request arrival patterns. To answer RQ2 and mitigate potential bias, we simulate context caching scenarios using the following approach. First, we select d test cases from *2WikiMQA* to simulate d users querying the same inference instance within a limited time window (40 seconds). Second, each user splits their contexts into chunks and pre-generates the corresponding KV cache. The parameter d primarily affects the ratio r of GPU HBM used to store context KV cache, referred to as context cache ratio (CCR). Third, we repeatedly send the d requests containing cache IDs and dynamic tokens at a constant request rate over the 40-second period, measuring the latency and throughput of completed requests. We simulate different questions on the same set of documents by sending requests without prefix caching. Additionally, request arrival times are simulated by sampling from a Poisson distribution to model varying request rates.

6.3 RQ1. Accuracy-latency trade-off of *AttnLink*

To address RQ1, we employ the synchronous workloads described in the preceding section, yielding two key insights from the experimental results (Figure 6). First, different variants of *AttnLink* (represented by a series of gradient blue stars) establish a new Pareto frontier, outperforming different variants of *CacheBlend* (represented by gradient orange rectangles). Second, *AttnLink*-20 is sufficient to limit accuracy drops within 0 - 7% and reduces up to 300% *TTFT* (this number increases as chunk size increases), compared to the default *CacheBlend*-15 configuration as reported in their paper. On the contrary, *CacheBlend*-1 or *CacheBlend*-5 recompute a similar number of tokens to *AttnLink*-20, but they have very bad accuracy (up to 80% drop compared to *FR*).

However, there are five unusual observations that warrant further explanation:

- All approaches, including *FR*, *CacheBlend*, and *AttnLink*, exhibit anomalous behavior in the third column (Yi model) due to the Yi Coder model’s poor handling of document understanding. This observation suggests that to ensure both algorithms perform optimally, robust models well-suited to the task are required.
- *CacheBlend* performs particularly poorly on the *SAMSum* dataset. This is likely because *CacheBlend* is overoptimized to deal with cross-attention scenarios, but *SAMSum* mainly consists of few-shot examples that do not need cross-attention.

- All approaches using all models perform poorly on the *MultiNews* dataset. This phenomenon can be attributed to the inherent difficulty of summarizing long documents with small models.
- *CacheBlend* and *AttnLink* exhibit similar *TTFT* in the *MultiNews* dataset. This is because each document in *MultiNews* is relatively short (around a few hundred tokens), making the number of tokens recomputed (k tokens in *AttnLink*- k) equivalent to the percentage of tokens recomputed ($r\%$ in *CacheBlend*- r).
- Unlike *CacheBlend*, increasing the number of recomputed tokens in *AttnLink* does not necessarily lead to higher accuracy. This is because the recomputation of only the "initial" tokens at each chunk boundary is critical for accuracy recovery. Recomputing additional tokens beyond these yields minimal improvement and can even result in accuracy degradation, as observed in datasets like *MuSiQue* and *Llama 3.1*.

6.4 RQ2. Latency and throughput of EPIC under asynchronous workloads

To address RQ2, we employ the asynchronous workloads on *AttnLink*-16 and *CacheBlend*-15. Additionally, due to the expiration of our H800 GPU, we used an A100 40G for this experiment (all other experiments in other RQs are carried out with H800 GPU), which we believe has a minimal impact on the validity of our conclusions. The results for latency (*TTFT*) and throughput (tokens/s) are discussed in the following two paragraphs and shown in Figure 7.

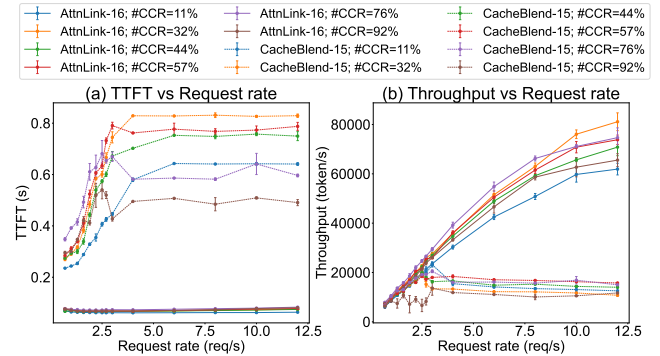


Figure 7. Latency and throughput comparison of EPIC-16 and *CacheBlend*-15 under asynchronous workloads with varying request rates and context cache ratios (CCR). Each data point represents the average and standard deviation from five experiments. EPIC-16 is shown using solid lines, while *CacheBlend*-15 is represented with dashed lines. Two algorithms with the same CCR are shown as the same color. (a) *TTFT* vs. request rates. (b) Throughput vs. request rates.

EPIC achieves throughput that is 3x to 8x higher than *CacheBlend*. Additionally, with higher CCR, EPIC’s throughput continues to improve until it reaches a threshold (about

⁵<https://medium.com/@priyanshu.pansari/geminis-game-changing-context-caching-feature-saving-money-and-time-in-llms-03e56c141bae>

30%), after which further increases in CCR have a reverse effect.

Regarding *TTFT* versus request rates (Figure 7 (a)), we observe three key trends. First, EPIC-16 exhibits *TTFT* values that are up to 8x smaller than CacheBlend-15, largely because EPIC-16 recomputes significantly fewer tokens. Second, as the context cache ratio (CCR) increases, EPIC-16’s *TTFT* remains stable, while CacheBlend-15’s *TTFT* fluctuates around 0.5 seconds. This is likely due to EPIC-16 generating fewer intermediate results, leading to reduced interference amplification. Third, as the request rates increase, *TTFT* plateaus instead of growing exponentially. This behavior can be attributed to vLLM’s predetermined limit on the number of requests in the running queue. When memory capacity is insufficient, additional requests are not added to the running queue. If we were to include the *TTFT* of all waiting requests, it would approach infinity.

Regarding throughput versus request rates (Figure 7 (b)), we observe two notable facts. First, EPIC-16 achieves a throughput that is also up to 7x higher than CacheBlend-15, as it precomputes fewer tokens, allowing more requests to be processed simultaneously. Second, as CCR increases, EPIC-16’s throughput continues to improve until the CCR reaches a threshold (approximately 30%), beyond which further increases in CCR lead to a reverse effect because requests start to interfere with each other severely. In contrast, CacheBlend-15’s throughput remains constant as it becomes incapable of handling additional requests.

The observed throughput improvement may appear unusually high due to our asynchronous workload settings, where all context caches are pre-generated, and only recomputation occurs on the GPU. This allows the GPU memory to handle up to 7x more requests using *AttnLink* (a few tokens to recompute per request) compared to CacheBlend (7x tokens to recompute per request), resulting in a corresponding 7x throughput increase. However, this number should be interpreted cautiously when considering real-world scenarios.

6.5 RQ3. EPIC performance under long context

To address RQ3, we send one request of varying context lengths with a fixed chunk size (1024 tokens) and observe two behaviors of *FR*, CacheBlend-15, and *AttnLink*-16, as shown in Figure 8. First, as context length increases, the *TTFT* of both *FR* and CacheBlend-15 grows quadratically, while EPIC-16 exhibits nearly linear growth. This difference arises because *FR* and CacheBlend-15 have time and resource complexities of $O(N^2)$, while EPIC-16 operates with a complexity of $O(kN)$, where k represents the number of recomputed tokens at chunk boundaries. Since $k \ll N$ grows slowly relative to N , EPIC-16 scales more efficiently. Second, EPIC-16 supports longer context length compared to CacheBlend-15. Specifically, CacheBlend-15 encounters an out-of-memory (OOM) error at approximately 35,000 tokens, while EPIC-16 avoids OOM until the context length reaches 50,000 tokens.

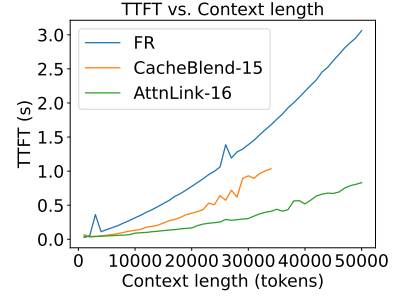


Figure 8. TTFT vs. context length of *FR*, CacheBlend-15 and *AttnLink*-16, using a fixed chunk size of 1024 tokens. For *FR*, we do not pre-generate context cache to display its full quadratic time complexity trend, as it would otherwise run out of memory earlier than CacheBlend-15 and *AttnLink*-16.

This difference is due to CacheBlend-15’s need to process more tokens and generate additional intermediate results, leading to higher GPU memory usage. *FR* potentially lasts the longest before encountering OOM because we do not pre-generate context cache to display its full quadratic time complexity trend, as it would otherwise run out of memory earlier than CacheBlend-15 and *AttnLink*-16.

6.6 RQ4. The impact of splitting methods on accuracy and performance

To address RQ4, we use the synchronous workloads constructed from *SAMSum* and *HotpotQA* on the Mistral model, comparing two splitting methods: semantic-based (ours) and fixed-token-based (512-token) chunks. We have two key observations from the experimental results (Figure 9). First, the choice of splitting method has a significant impact on accuracy for both CacheBlend and *AttnLink*. For instance, using 512-token chunks improves accuracy on *SAMSum* but reduces it on *HotpotQA*, compared to semantic-based splitting. We observe that the semantic-based splitting method suits long documents such as those in *HotpotQA*, *2WikiMQA*, *Needle in a haystack*, while the fixed-token-based splitting method suits short documents such as those in *SAMSum* (short few-shot examples). Second, the choice of splitting method has a significant impact on performance for *AttnLink*. Since the number of recomputed tokens in *AttnLink* increases with chunk count, semantic-based splitting on *HotpotQA* leads to larger chunks, reducing the number of recomputed tokens and resulting in lower *TTFT*, while the opposite occurs in *SAMSum*.

In RQ1, we compared *AttnLink* and CacheBlend using the same semantic-based splitting method. In this figure, we are able to extend the comparison by evaluating *AttnLink* with semantic-based splitting (blue stars in the first column) against CacheBlend with its original 512-token splitting method (orange squares in the second column). The results demonstrate that our algorithm remains on the Pareto

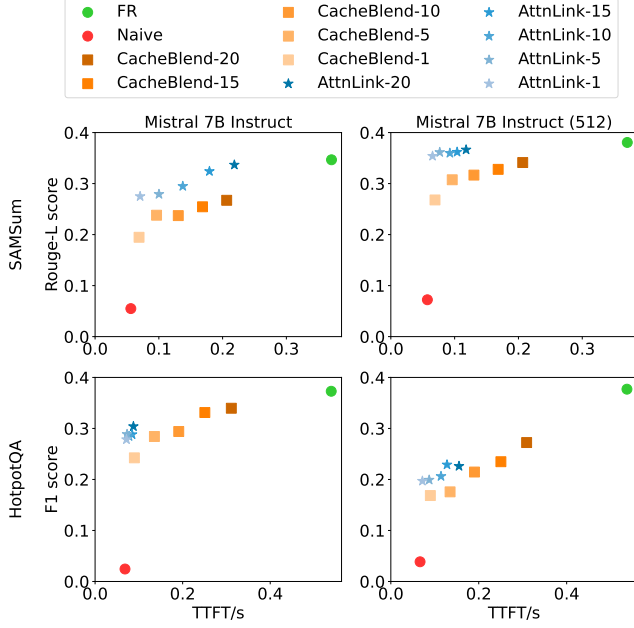


Figure 9. The axes, legends, and data points in this figure follow the same definitions as those in Figure 6, with one key difference: the second column represents results from the Mistral model using fixed-size token chunks (512 tokens per chunk). We only show the results of one model and two datasets due to the page limitation.

frontier. The trend of *HotpotQA* is consistent across datasets that are not shown in the figure, such as *2WikiMQA*, *Needle in a Haystack*, and *MuSiQue*, which all involve long documents. For *MultiNews*, accuracy remains relatively low as before.

These observations highlight the need to offer a component like *KVSplit*, allowing users to tailor the splitting method to specific tasks and models, utilizing their task-specific knowledge. Importantly, this component is only functional when explicit context caching is employed.

7 Related Work and Discussion

Our work is unique in proposing position-independent context caching and advancing the SOTA results in this field. We briefly navigate the whole design space.

LLM Serving Optimizations. Several serving systems emerged in the past year. vLLM [11] is a pioneering work in this space featuring PagedAttention for higher throughput. SGLang [32] is another serving system featuring a novel frontend language and a backend runtime. SGLang proposes three novel techniques: RadixAttention, a compressed finite state machine, and API speculative execution. Aside from full-systems, there are also many scheduling optimizations such as disaggregated prefill and decode [9, 10, 19, 33], continuous batching [29], multi-lora [12, 22], etc.

Context Caching (CC). Prompt Cache [5] was one of the first approaches to reuse attention states for frequently occurring text segments, such as system messages, by introducing a schema called prompt modules. This schema allows Prompt Cache to store and reuse KV caches while maintaining positional accuracy. In late 2023, several prefix-based CC solutions emerged, including Pensieve [30], CacheGen [14], and SGLang [32]. By mid-2024, vendors like Kimi [17] and Gemini [8] began offering explicit CC features.

However, prefix-based CC has clear limitations, as it only works when there is an exact prefix match across requests. To overcome this, a concurrent work called CacheBlend [27] introduced a position-independent approach that selectively recomputes a small subset of tokens, enabling KV cache reuse regardless of token location. CacheBlend is a pioneering effort in PIC. Building on this foundation, our work makes two new contributions. First, we propose a more efficient PIC linking algorithm based on static sparsity, significantly improving TTFT and accuracy across most datasets. Second, we address the limitations of rigid, fixed-length chunking by developing a more flexible and adaptive chunking strategy.

Sparsity. Sparsity is essential for improving long-context inference and can be divided into two types: dynamic and static. Dynamic sparsity (e.g., H2O [31], Quest [23]) adapts in real-time by identifying and filtering out less important query-key connections as sequences are processed. In contrast, static sparsity (e.g., Longformer [2]) relies on predefined sparse patterns, which simplifies implementation but reduces flexibility. Most previous works apply sparsity to reduce memory footprint in the decode stage. In contrast, both CacheBlend and *AttnLink* leverages sparsity to reduce computation in the prefill stage, but CacheBlend leverages dynamic sparsity while our *AttnLink* leverages static sparsity to enable efficient position-independent caching.

Discussion on Context Caching Abstraction. In explicit CC, users manage the process of splitting static tokens into chunks, invoking CC APIs, and indexing pre-generated KV caches using cache IDs [8, 17]. This method gives users control over the semantic structure of static tokens, allowing for accuracy, storage, and latency/throughput optimizations. Since users have a deeper understanding of the content, they can organize static tokens to improve linking efficiency and preserve semantic independence. Implicit CC, on the other hand, shifts the responsibility for splitting and caching to the system, which indexes pre-generated KV caches using methods like text matching (e.g., hash-based or radix-tree-based approaches) [9, 11, 32]. Although this reduces the user’s workload, it introduces indexing overhead and may lead to suboptimal token splitting, potentially affecting accuracy. Overall, two CC abstractions present different trade-offs between usability and performance.

8 Conclusion

In this paper, we presented EPIC, a large language model (LLM) serving system that overcomes the limitations of traditional prefix-based context caching by enabling position-independent caching (PIC) to accelerate LLM inference. EPIC introduces AttnLink, a novel linking algorithm that leverages static attention sparsity, alongside KVSplit, a customizable chunking strategy. These modules significantly improve time-to-first-token (TTFT) and throughput while maintaining accuracy. Our evaluation across a range of datasets and LLM models demonstrates that EPIC delivers up to 8× improvements in TTFT and 7× in throughput compared to existing systems, with minimal or no accuracy loss. We believe that position-independent caching is still in its early stages, and future work could explore more advanced sparse attention algorithms tailored for PIC and further optimize the interaction between PIC-enabled prefill and decode phases.

References

- [1] Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. Longbench: A bilingual, multitask benchmark for long context understanding, 2024.
- [2] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. 2020.
- [3] Devendra Singh Chaplot. Albert q. jiang, alexandre sablayrolles, arthur mensch, chris bamford, devendra singh chaplot, diego de las casas, florian bressand, gianna lengyel, guillaume lample, lucile saulnier, l  lio renard lavaud, marie-anne lachaux, pierre stock, teven le scao, thibaut lavril, thomas wang, timoth  e lacroix, william el sayed. *arXiv preprint arXiv:2310.06825*, 2023.
- [4] fastforwardlabs. Evaluating qa: Metrics, predictions, and the null response. https://github.com/fastforwardlabs/ff14_blog/blob/master/_notebooks/2020-06-09-Evaluating_BERT_on_SQuAD.ipynb, 2020.
- [5] In Gim, Guojun Chen, Seung-seob Lee, Nikhil Sarda, Anurag Khandelwal, and Lin Zhong. Prompt cache: Modular attention reuse for low-latency inference. *Proceedings of Machine Learning and Systems*, 6:325–338, 2024.
- [6] gkamradt. Llmtest needle in a haystack - pressure testing llms. https://github.com/gkamradt/LLMTest_NeedleInAHaystack, 2023.
- [7] Google. Gemini. <https://gemini.google.com/>.
- [8] Google. Gemini Context Caching. <https://ai.google.dev/gemini-api/docs/caching?lang=python>.
- [9] Cunchen Hu, Heyang Huang, Junhao Hu, Jiang Xu, Xusheng Chen, Tao Xie, Chenxi Wang, Sa Wang, Yungang Bao, Ninghui Sun, et al. Memserve: Context caching for disaggregated llm serving with elastic memory pool. *arXiv preprint arXiv:2406.17565*, 2024.
- [10] Cunchen Hu, Heyang Huang, Liangliang Xu, Xusheng Chen, Jiang Xu, Shuang Chen, Hao Feng, Chenxi Wang, Sa Wang, Yungang Bao, et al. Inference without interference: Disaggregate llm inference for mixed downstream workloads. *arXiv preprint arXiv:2401.11181*, 2024.
- [11] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626, 2023.
- [12] Suyi Li, Hanfeng Lu, Tianyuan Wu, Minchen Yu, Qizhen Weng, Xusheng Chen, Yizhou Shan, Binhang Yuan, and Wei Wang. Caraserve: Cpu-assisted and rank-aware lora serving for generative llm inference. *arXiv preprint arXiv:2401.11240*, 2024.
- [13] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
- [14] Yuhao Liu, Hanchen Li, Kuntai Du, Jiayi Yao, Yihua Cheng, Yuyang Huang, Shan Lu, Michael Maire, Henry Hoffmann, Ari Holtzman, et al. CacheGen: Fast context loading for language model applications. *arXiv preprint arXiv:2310.07240*, 2023.
- [15] LMCACHE. LMCACHE. <https://github.com/LMCACHE/LMCACHE>.
- [16] Meta. Introducing Llama 3.1: Our most capable models to date. <https://ai.meta.com/blog/meta-llama-3-1/>.
- [17] Moonshot AI. Kimi Context Caching. <https://platform.moonshot.cn/docs/guide/use-context-caching-feature-of-kimi-api>.
- [18] OpenAI. OpenAI Prompt Caching. <https://platform.openai.com/docs/guides/prompt-caching>.
- [19] Pratyush Patel, Esha Choukse, Chaojie Zhang, Aashaka Shah,   igo Goiri, Saeed Maleki, and Ricardo Bianchini. Splitwise: Efficient generative llm inference using phase splitting. In *2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)*, pages 118–132. IEEE, 2024.
- [20] Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. Efficiently scaling transformer inference. *Proceedings of Machine Learning and Systems*, 5:606–624, 2023.
- [21] Ruoyu Qin, Zheming Li, Weiran He, Mingxing Zhang, Yongwei Wu, Weimin Zheng, and Xinran Xu. Mooncake: A kv-cache-centric disaggregated architecture for llm serving, 2024. URL <https://arxiv.org/abs/2407.00079>.
- [22] Ying Sheng, Shiyi Cao, Dacheng Li, Coleman Hooper, Nicholas Lee, Shuo Yang, Christopher Chou, Banghua Zhu, Lianmin Zheng, Kurt Keutzer, et al. S-lora: Serving thousands of concurrent lora adapters. *arXiv preprint arXiv:2311.03285*, 2023.
- [23] Jiaming Tang, Yilong Zhao, Kan Zhu, Guangxuan Xiao, Baris Kasikci, and Song Han. Quest: Query-aware sparsity for efficient long-context llm inference. 2024.
- [24] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.
- [25] Wikipedia. Position-independent code. https://en.wikipedia.org/wiki/Position-independent_code.
- [26] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.
- [27] Jiayi Yao, Hanchen Li, Yuhao Liu, Siddhant Ray, Yihua Cheng, Qizheng Zhang, Kuntai Du, Shan Lu, and Junchen Jiang. Cacheblend: Fast large language model serving with cached knowledge fusion. *arXiv preprint arXiv:2405.16444*, 2024.
- [28] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024.
- [29] Gyeong-In Yu, Joo Seong Jeong, Geon-Woo Kim, Soojeong Kim, and Byung-Gon Chun. Orca: A distributed serving system for {Transformer-Based} generative models. In *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)*, pages 521–538, 2022.
- [30] Lingfan Yu and Jinyang Li. Stateful large language model serving with pensieve. *arXiv preprint arXiv:2312.05516*, 2023.
- [31] Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher R  , Clark Barrett, et al. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in Neural Information Processing Systems*, 36, 2024.
- [32] Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Jeff Huang, Chuyue Sun, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E Gonzalez, et al. Efficiently programming large language models using sglang. *arXiv preprint arXiv:2312.07104*, 2023.

- [33] Yinmin Zhong, Shengyu Liu, Junda Chen, Jianbo Hu, Yibo Zhu, Xuanzhe Liu, Xin Jin, and Hao Zhang. Distserve: Disaggregating prefill and decoding for goodput-optimized large language model serving. *arXiv preprint arXiv:2401.09670*, 2024.
- [34] Zixuan Zhou, Xuefei Ning, Ke Hong, Tianyu Fu, Jiaming Xu, Shiyao Li, Yuming Lou, Luning Wang, Zhihang Yuan, Xiuhong Li, et al. A survey on efficient inference for large language models. *arXiv preprint arXiv:2404.14294*, 2024.