From Static Inference to Dynamic Interaction: A Survey of Streaming Large Language Models

Junlong Tong[1,2], Zilong Wang[2], YuJie Ren[2], Peiran Yin[2], Hao Wu[2], Wei Zhang[2], Xiaoyu Shen[2]*

[1]Shanghai Jiao Tong University

[2]Institute of Digital Twin, Eastern Institute of Technology, Ningbo

jl-tong@sjtu.edu.cn xyshen@eitech.edu.cn

Abstract

Standard Large Language Models (LLMs) are predominantly designed for static inference with pre-defined inputs, which limits their applicability in dynamic, real-time scenarios. To address this gap, the *streaming LLM* paradigm has emerged. However, existing definitions of streaming LLMs remain fragmented, conflating streaming generation, streaming inputs, and interactive streaming architectures, while a systematic taxonomy is still lacking. This paper provides a comprehensive overview and analysis of streaming LLMs. First, we establish a unified definition of streaming LLMs based on data flow and dynamic interaction to clarify existing ambiguities. Building on this definition, we propose a systematic taxonomy of current streaming LLMs and provide an in-depth discussion of their underlying methodologies across text, speech, and video streaming scenarios. Furthermore, we explore the applications of streaming LLMs in real-world scenarios and outline promising research directions to support ongoing advances in streaming intelligence. We maintain a continuously updated repository of relevant papers at https://github.com/EIT-NLP/Awesome-Streaming-LLMs.

1 Introduction

Large Language Models (LLMs) have shown remarkable efficacy across diverse domains, exhibiting strong reasoning, generation, and cross-modal capabilities (OpenAI, 2023; Team et al., 2023; DeepSeek-AI et al., 2024). However, LLMs are predominantly pre-trained on *static and full-context corpora*, following a "read-at-once" paradigm in which the complete input is provided before any output is generated. While effective for benchmark-style tasks, this paradigm fundamentally limits their applicability in real-world environments, where information arrives incrementally, accumulates over time, and may be unbounded in length.

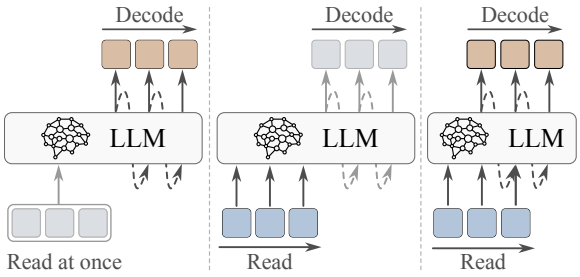


Figure 1: Illustration of three types of streaming large language models (LLMs). (*Left*) **Output-streaming LLM** performs streaming generation *after* static reading. (*Middle*) **Sequential-streaming LLM** performs streaming generation *after* streaming reading. (*Right*) **Concurrent-streaming LLM** performs streaming generation *while* streaming reading.

Such dynamic conditions are ubiquitous in tasks like real-time translation, streaming video understanding, and interactive tool agents (Agostinelli et al., 2024; Jin et al., 2025; Yang et al., 2025b). In these real-world applications, inputs such as speech and video data stream continuously, forcing systems to maintain an evolving understanding based on partial observations. In more complex scenarios, these signals may originate from *multiple concurrent streams* (Li et al., 2025k), while systems may also need to generate multiple outputs in parallel (Zhang et al., 2025b). For instance, a robot may need to act, speak, and reason simultaneously (Zhang et al., 2025e), whereas an interactive assistant may coordinate speech, visual updates, and control commands (Zhang et al., 2025a). Since the input is never fully available at any given moment, the system must dynamically decide *when to respond, when to wait for more information, and when to terminate* (Panchal et al., 2024; Zhang et al., 2025c). These requirements expose a fundamental mismatch with the offline, full-context design of standard LLMs.

Adapting LLMs to these real-world streaming scenarios presents significant challenges. Beyond architectural modifications, there is a scarcity of

*Corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 10237–10263

large-scale pre-training data that supports real-time interaction, partial-input supervision, and fine-grained temporal alignment. Motivated by this gap, recent research has begun to investigate *streaming LLMs* (Tong et al., 2025a; Chen et al., 2024a; Du et al., 2024). However, the field currently suffers from terminological ambiguity. Existing studies often conflate distinct concepts, such as *autoregressive decoding* (Kondratyuk et al., 2024), *incremental or chunk-wise encoding* (Xiao et al., 2023), and *full-duplex interaction* like GPT-4o (OpenAI, 2023), under a single "Streaming LLM" umbrella, obscuring meaningful comparisons.

In this work, we provide the first systematic review of streaming LLMs, proposing a unified definition based on data flow and interaction concurrency. As illustrated in Figure 1, we categorize these models into three distinct levels: (1) *Output-streaming LLMs*, which retain static input processing but support streaming output generation. (2) *Sequential-streaming LLMs*, which process streaming inputs incrementally but generate with full input. (3) *Concurrent-streaming LLMs*, which enable full-duplex interaction by continuously receiving inputs and generating outputs.

This taxonomy captures both conceptual distinctions and a clear progression of technical challenges: Output streaming addresses challenges in streaming and low-latency generation; sequential streaming introduces incremental encoding and context management; and concurrent streaming builds upon both to address architecture adaptation and interaction strategies required for full-duplex processing. By disentangling these paradigms, the taxonomy clarifies which challenges are shared, which are incremental, and which are unique to each category, thereby providing a structured roadmap toward the ultimate goal of fully interactive streaming LLMs. Guided by this framework, we systematically review representative methods in each category, examine emerging applications such as streaming video understanding and real-time reasoning, and highlight open problems, including trade-offs between latency and performance, to inform future research.[1]

To summarize, our main contributions include:

- To our knowledge, we are the **first systematic survey** of streaming LLMs.
- We introduce **a unified definition** of streaming LLMs, clarifying the conceptual distinctions among existing paradigms.
- We provide **a systematic taxonomy and comprehensive technical analysis**, disentangling the mechanisms of three streaming paradigms.
- We discuss **emerging applications and open research directions** for real-time and interactive streaming scenarios.

[1]We provide a detailed description of motivation, survey scope, and difference with related surveys in Appendix A.

2 Preliminaries

2.1 Background of Streaming LLMs

Current LLMs typically operate under a batch processing paradigm, where the model encodes the entire input sequence into the KV cache during the prefill phase and subsequently generates tokens autoregressively in the decoding phase. Consequently, from a data flow perspective, standard LLMs can be categorized as "streaming-output LLMs" that rely on static context availability. However, real-world data flows often exhibit dynamic and continuous characteristics (e.g., real-time speech transcription and content understanding), necessitating models capable of handling streaming inputs and executing timely output decisions; therefore, generalized streaming LLMs are defined to address such dynamic input and immediate response scenarios, aiming to transcend the limitations of static preprocessing and delayed response.

2.2 Formal Definition

To rigorously unify the diverse landscape of streaming LLMs in Figure 1, we formulate the modeling process as a conditional probability distribution $P(Y|X)$, where $X = (x_1, \ldots, x_M)$ denotes the bounded input stream and $Y = (y_1, \ldots, y_N)$ denotes the output stream. This distribution can be factorized autoregressively using the chain rule:

$$P(Y|X) = \prod_{t=1}^{N} P\big(y_t | y_{<t}, h_{1:\phi(t)}(X); \theta\big) \quad (1)$$

where θ denotes the LLM parameters, and $h_{\phi(t)}(X) = llm(x_{\phi(t)})$ represents the encoded hidden states corresponding to the input prefix $x_{\phi(t)}$. Here, $\phi(t)$ is a decision function to determine the input stream visible at generation step t. This general definition can be instantiated into three subtypes by applying varying operational constraints.

Output-streaming LLMs This paradigm imposes a static constraint where the entire input must

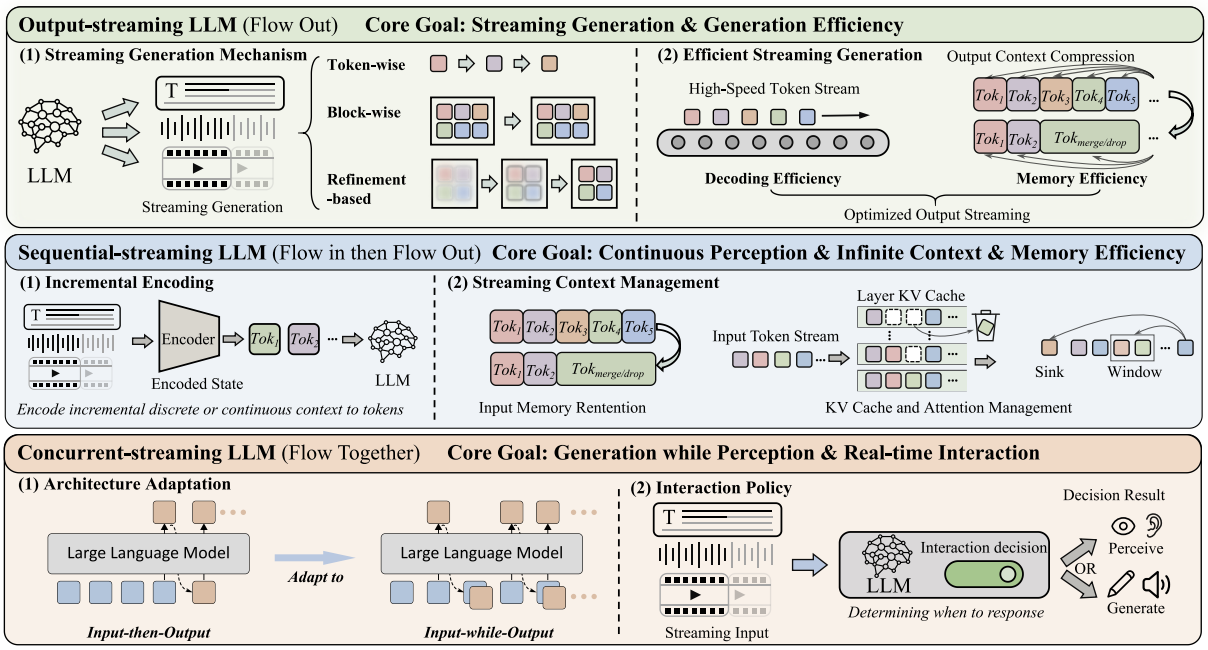


Figure 2: Overview of streaming LLM paradigms and their key challenges. The figure contrasts Output-streaming, Sequential-streaming, and Concurrent-streaming LLMs, highlighting their core goals and corresponding research components. Concurrent-streaming builds on the first two and adds extra challenges in real-time streaming architecture adaptation and interaction policy learning.

be processed before generation begins. Mathematically, the decision function is constant relative to the total input length M, i.e., $\phi(t) = M$ for all $t \in \{1, \ldots, N\}$. The hidden states are computed via a one-time global prefilling: $h_{1:\phi(t)}(X) = h_{1:M}(X) = llm(X_{1:M})$.

Sequential-streaming LLMs This paradigm processes dynamic streaming inputs but generates based on a fixed input. While the decision function mirrors the above type (i.e., $\phi(t) = M, \forall t$), the hidden states are constrained by stepwise arrival: $h_{1:M}(X) = \{llm(x_1), \ldots, llm(x_M)\}$. This represents a sequential encoding process where the context is accumulated token-by-token (or chunk-by-chunk) before the generation phase begins.

Concurrent-streaming LLMs This paradigm imposes the strictest temporal constraints, representing a dynamic process where streams unfold continuously. Mathematically, $\phi(t)$ must satisfy monotonicity and partial visibility:$1 \leq \cdots \leq \phi(t) \leq \phi(t+1) \leq \cdots \leq M$. The hidden states of input stream are computed via a dynamic or interactive process: $h_{\phi(t)}(X) = llm(X_{\phi(t)}, y_{<t})$.

The tripartite taxonomy defined above reflects a trajectory of escalating operational constraints and functional demands, shifting the paradigm from static processing to dynamic, real-time interaction.

$$1 \leq \cdots \leq \phi(t) \leq \phi(t+1) \leq \cdots \leq M.$$

2.3 Overview

This survey provides a systematical overview of research in streaming LLMs. Figure 2 illustrates the proposed taxonomy, detailing the primary research focuses and challenges within each category. Specifically, output-streaming emphasizes streaming generation mechanisms and efficient generation; sequential-streaming focuses on incremental encoding processing and context management for input streams; and concurrent-streaming integrates both tasks, additionally introducing architectural adaptations and the interactive management of simultaneous input and output streams. To navigate this comprehensive landscape, Figure 3 outlines the taxonomy structure of this survey. Guided by this taxonomy, we begin with output-streaming in Section 3, expand to the dynamic input processing of sequential-streaming in Section 4, and culminate with the interactive dynamics of concurrent-streaming in Section 5. Beyond the technical part, Section 6 reviews downstream tasks and applications, and Section 7 discusses the future directions.

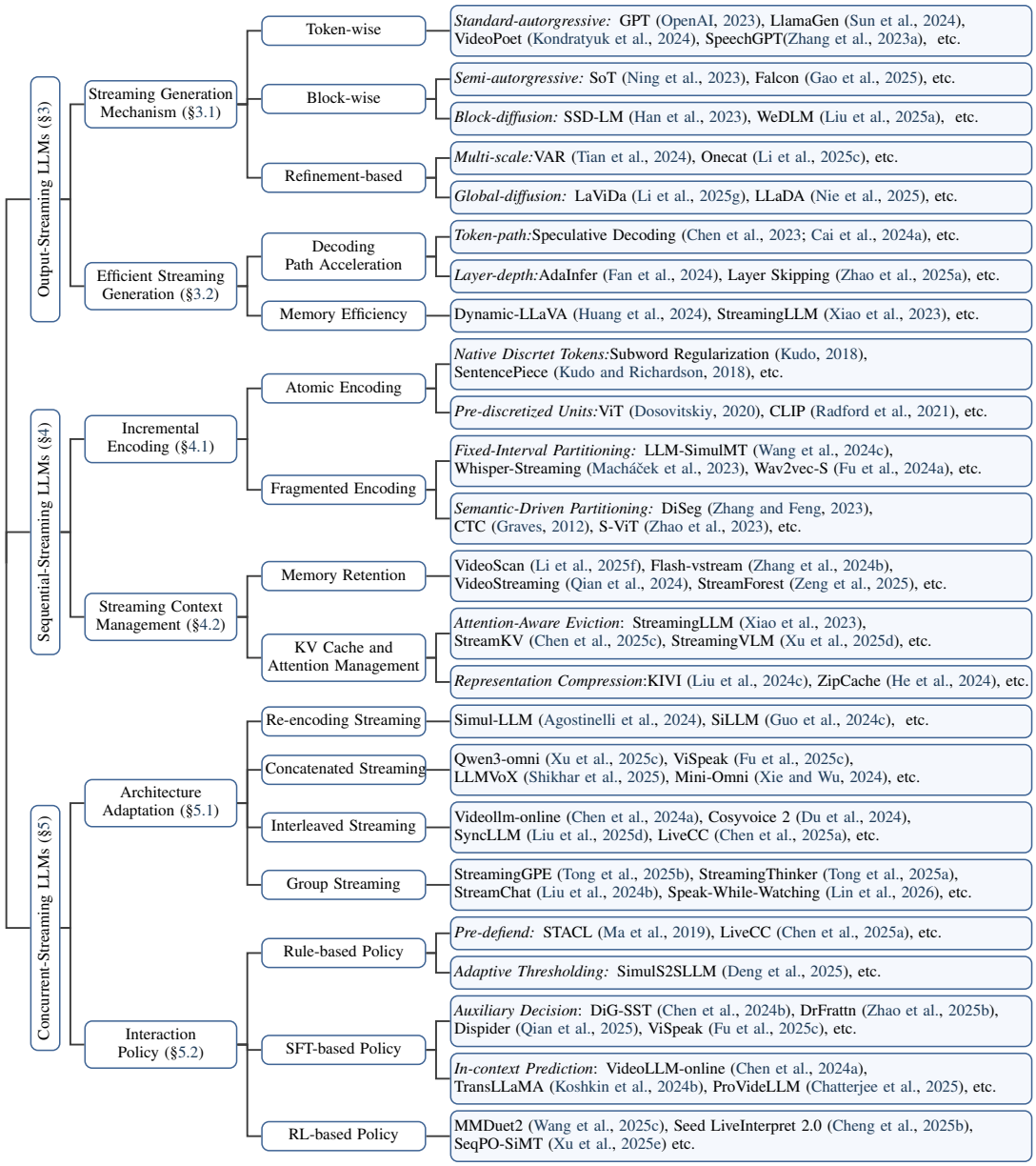


Figure 3: Taxonomy of Streaming Large Language Models.

3 Output-Streaming LLMs: Generating with Progressive Revelation

3.1 Streaming Generation Mechanism

Output streaming enables *progressive revelation* by continuously emitting intermediate results rather than waiting for completion. Based on the generation granularity and update mechanism, we categorize existing methods into: (i) **token-wise**, (ii) **block-wise**, and (iii) **refinement-based**.

Token-wise This represents the dominant generation paradigm for LLMs, employing token-wise autoregressive decoding (Team et al., 2023; DeepSeek-AI et al., 2024; Gemma Team, 2024). For multimodal outputs, systems typically extend this paradigm by aligning non-text modalities to the textual space for autoregressive streaming (Zhang et al., 2023a; Contributors, 2024).

Block-wise These methods expand the generation unit from single tokens to multi-token blocks, reducing serial depth while retaining the controllability of autoregressive modeling. We summarized them into two lines. (1) *Semi-autoregressive* relaxes intra-block dependencies to predict multiple tokens *in parallel*. (Wang et al., 2018; Hwang et al., 2025; Ning et al., 2023; Gao et al., 2025). For example, MTP (Gloeckle et al., 2024) predicts multiple tokens simultaneously for each autoregressive block step. (2) *Block-diffusion* combines diffusion-

style refinement with block-wise generation, iteratively *denoising a block* at a time and streaming blocks autoregressively (Han et al., 2023; Liu et al., 2025a; Tian et al., 2025; Arriola et al., 2025).

Refinement-based Unlike token-by-token sequential accumulation, this paradigm performs progressive refinement *from coarse to fine*, iteratively improving the semantic completeness of the entire sequence rather than merely extending its length. (1) *Multi-scale* approach decomposes generation into discrete scales (Tian et al., 2024; Li et al., 2025c; Zhuang et al., 2025). Models like VAR (Tian et al., 2024) predict the next-scale autoregressively, enabling a blur-to-clear streaming effect. (2) *Global-diffusion* refinement formulates generation as multi-step denoising over the entire sequence, starting from noise or a coarse initialization and progressively refining to a complete output. This mechanism has been successfully adapted to both text (Nie et al., 2025; Li et al., 2025g; Song et al., 2025; Li et al., 2022) and multimodal generation (Xin et al., 2025; Yang et al., 2025c).

3.2 Efficient Streaming Generation

Given the extensive scope of LLM optimization, we narrow our focus strictly to the streaming process itself, analyzing decoding and memory efficiency.[2] As *token-wise* decoding remains dominant, we focus on its optimization for efficient streaming.

Decoding Path Acceleration To mitigate autoregressive latency, optimizations modify the execution trajectory along two dimensions. (1) *Token-path* methods generate parallel candidate chains to relax strict serial dependency, including multi-path and speculative decoding (Leviathan et al., 2023; Xiao et al., 2024b). For instance, speculative decoding (Chen et al., 2023; Cai et al., 2024a; Li et al., 2024d) leverages a lightweight draft model to propose multiple candidate tokens in parallel, which are then verified and selectively accepted by a target model, reducing streaming latency. (2) *Layer-depth* methods adaptively shorten the network depth based on token difficulty (Fan et al., 2024; Del Corro et al., 2023) . For instance, by employing layer skipping (Zhao et al., 2025a), models terminate the execution path prematurely.

Memory Efficiency Since the KV cache grows linearly, optimizations aim to decouple memory cost from generated length. *Dynamic KV compression* methods limit the scope of attention targets during streaming decoding (Liu et al., 2023; Zhang et al., 2023b; Liao et al., 2025; Huang et al., 2024). Representative implementations range from sink-aware windowing (Xiao et al., 2023), which maintains a fixed budget for stability, to dynamic decision strategies (Liao et al., 2025) for KV cache management based on token importance.

4 Sequential-Streaming LLMs: Processing Dynamic Input Streams

Building upon the foundation of output-streaming, this section turns to sequential-streaming: the continuous perception of *dynamic input streams*. The core technical imperative shifts from generation latency to sustainability. Specifically, we focus on two core mechanisms: handling incremental inputs to avoid re-computation, and optimizing context management to accommodate long input streams.

4.1 Incremental Encoding

Incremental encoding processes incoming streams solely based on past states, *with historical representations remaining unchanged under subsequent streaming inputs*, avoiding quadratic re-computation. The central issue lies in *how to define encoding units*, such that the encoding of each unit is not influenced by future information. Depending on the unit construction strategy, we categorize two types: *atomic encoding* and *fragmented encoding*.

Atomic Encoding This paradigm is applicable to streams that have inherent delimiters aligned with the model's processing unit. (1) *Native Discrete Tokens*: Text is the primary example, where input is naturally segmented into discrete tokens whose representations remain unchanged as new tokens arrive (Kudo, 2018; Kudo and Richardson, 2018). (2) *Pre-defined Units*: Certain modalities admit pre-defined atomic units independent of future context. For example, video streams can be incrementally processed at the frame level, where each frame serves as a fixed encoding unit and is encoded without being influenced by subsequent frames (Dosovitskiy, 2020; Radford et al., 2021).

Fragmented Encoding Fragmented encoding handles raw continuous signals (e.g., audio waveforms and video pixel streams) without natural delimiters by introducing artificial boundaries to interface with discrete LLM architectures. Boundary

[2]We provide related survey papers on efficient LLMs in Appendix A for reference.

construction typically follows two strategies. (1) *fixed-interval partitioning*, which slices streams at uniform temporal intervals for efficiency but may disrupt semantic units (Wang et al., 2024b; Macháček et al., 2023). (2) *semantic-driven partitioning*, which leverages content-dependent cues, such as word boundaries in speech (Zhang and Feng, 2023; Graves, 2012) and shot or scene transitions in video (Zhao et al., 2023), to better preserve semantic coherence at higher computational cost.

4.2 Streaming Context Management

Streaming context management focuses on maintaining and updating contextual information during incremental processing under limited memory and computation budgets. It can be viewed through three complementary aspects: *what information to keep* over long-running streams (memory), *how to store and update* it across decoding steps (KV cache), and *how to efficiently access* it via optimized attention mechanisms (attention).

Memory Retention Memory retention concerns what historical information should be preserved or discarded during long-running input streaming. We classify these methods into two primary categories. (1) *Salient content selection and eviction* approaches focuses on identifying and retaining salient tokens or segments while discarding less informative or redundant content as the stream grows (Zhang et al., 2024b; Yao et al., 2025; Qian et al., 2024; Wang et al., 2025b). Selection criteria are typically based on importance estimation, recency, or task relevance, enabling bounded memory usage under continuous inputs. (2) Instead of outright discarding past information, *token merging and memory consolidation* compress historical representations by aggregating multiple tokens or states into more compact forms (Zhong et al., 2024; Wang et al., 2023; Zeng et al., 2025; Chen et al., 2025b). Such strategies preserve coarse-grained contextual information while reducing memory footprint, allowing long-term context to be maintained in a compressed manner.

KV Cache and Attention Management While memory retention operates at the input level, this component focuses on the *internal* maintenance of intermediate states and the optimization of attention computation. Since the attention range dictates which historical states are required for generation, attention access patterns and cache storage strategies are inherently coupled in streaming scenarios.

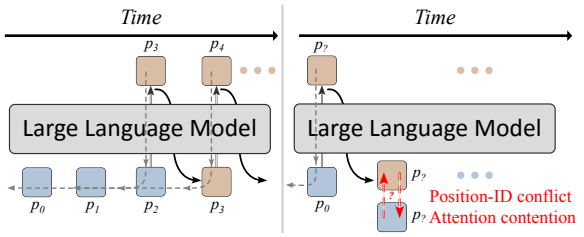


Figure 4: Illustration of structural conflicts when adapting batch-oriented LLMs (left) to concurrent streaming (right), where ⟶ indicates the token generation direction, - - - -➤ denotes attention dependencies, ▢ blocks represent the input, and ▢ blocks represent the output. (1) **Attention contention**: Ambiguous causal dependency between the newly inserted streaming input and historical outputs. (2) **Position-ID conflict**: The new streaming input and generated output compete for the identical position ID.

We categorize these strategies into two complementary directions. (1) *Attention-Aware Eviction*: These methods bound memory growth by restricting the attention mechanism to a sparse subset of historical tokens. By identifying and retaining only critical states, such as recent tokens maintained by a sliding window and high-importance attention sinks or heavy hitters, the model can safely evict unaccessed KV pairs, ensuring constant time and memory complexity without disrupting generation quality. (Xiao et al., 2023; Li et al., 2024c; Cai et al., 2024b; Yang et al., 2025e; Liao et al., 2025). (2) *Representation Compression*: Complementary to eviction, compression approaches reduce the memory footprint of the *retained* states. Techniques such as low-bit quantization or low-rank approximation compress the key–value representations, allowing the model to accommodate longer effective contexts within a fixed memory budget (Liu et al., 2024c; Hooper et al., 2024; He et al., 2024; Liu et al., 2025g).

5 Concurrent-Streaming LLMs: The Streaming of Real-Time Interaction

Concurrent-streaming represent a crucial step toward real-time interactive intelligence, requiring LLMs to simultaneously process streaming inputs and generate outputs. However, this dynamic paradigm diverges from standard static pre-training. First, regarding architecture adaptation, concurrent streaming introduce structural conflicts, as illustrated in Figure 4. Second, synchronization control governs system interactivity by dynamically deciding when to alternate between reading and writing, balancing responsiveness and coherence, as illustrated in Figure 5. Accordingly, we catego-

rize existing research into *architecture adaptation* and *interaction policy*.

5.1 Architecture Adaptation

Architecture adaptation mitigates structural conflicts inherent in concurrent processing, including attention contention and positional conflicts (Figure 4). Attention contention arises when continuously arriving inputs interleave with generation, making attention dependency ordering ambiguous, while positional conflicts occur when asynchronously injected inputs overlap with output positions. Existing work redesigns input–output interaction mechanisms, which we categorize into four representative streaming paradigms.

Re-encoded streaming The model re-encodes all historical caches whenever new input arrives (Deng et al., 2025; Agostinelli et al., 2024; Guo et al., 2024c). By recomputing representations over the entire context, this approach eliminates attention contention and positional misalignment, preserving batch-equivalent attention dependencies. However, the resulting computational overhead limits its applicability to long-context and real-time settings (Guo et al., 2024b; Raffel et al., 2024).

Concatenated streaming Concatenated streaming concatenates the newly arrived input tokens with the previously generated outputs and feeds them jointly into the model at each step (Xu et al., 2025c,b; Ding et al., 2025a; Shikhar et al., 2025). This design resolves both conflicts by unifying attention and positional ordering, but incurs growing memory and latency and requires architectural changes and retraining (Shikhar et al., 2025).

Interleaved streaming This paradigm interleaves input and output tokens within a shared sequence, assigning attention and positional encodings according to their temporal order (Chen et al., 2024a; Du et al., 2024; Liu et al., 2025d; Xu et al., 2025d; Chen et al., 2025a; Qian et al., 2025). It preserves the temporal flow of streaming interaction, enabling input and output to coexist with consistent ordering (Chen et al., 2025a). While balancing computational efficiency and real-time continuity, it requires synchronization mechanisms to prevent dependency leakage.

Grouped streaming Group streaming partitions input and output tokens into separate groups, each with independent attention relations and position IDs (Liu et al., 2024b; Tong et al., 2025a,b; Lin et al., 2026). This design eliminates attention contention while maintaining isolated positional spaces, and empirical results show that grouped positional encoding preserves streaming performance and can improve parallelism and efficiency.

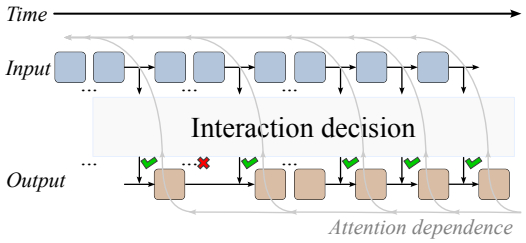


Figure 5: Illustration of interaction decision in concurrent streaming LLMs, where the model learns to dynamically schedule reading inputs and emitting outputs.

5.2 Interaction policy

Interaction policy governs read–write synchronization in concurrent LLMs, balancing latency and output quality. Existing strategies fall into three paradigms based on their optimization approach: rule-based, SFT-based, and RL-based policies.

Rule-based Interaction Rule-based approaches rely on predetermined schedules or statistical thresholds, offering interpretability and control without requiring model parameter updates. (1) *Pre-defined* strategy enforce a rigid, content-agnostic read-write rhythm (Ma et al., 2019; Chen et al., 2025a; Tong et al., 2025a). The most representative approach is the *Wait-k policy* (Ma et al., 2019). In this strategy, the model always waits for k tokens or segments of input lag before generating the corresponding output. While efficient and easy to implement, pre-defined policies lack adaptability to varying input complexity and rate fluctuations. (2) *Adaptive thresholding* methods utilize real-time inference statistics as decision signals to improve flexibility (Agostinelli et al., 2024; Yang et al., 2025g). These policies trigger read-/write actions based on metric thresholds (e.g., attention weights) rather than a fixed schedule. For instance, SimulS2S (Agostinelli et al., 2024) monitors model confidence and pauses generation to read more context whenever uncertainty exceeds a safety margin, effectively adapting to the difficulty of the incoming stream.

SFT-based Interaction Moving beyond manual rules, supervised approaches leverage labeled data to explicitly train the model to predict the opti-

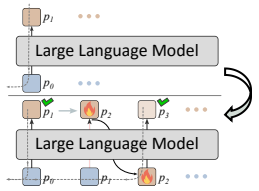
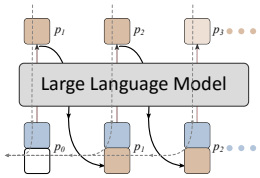
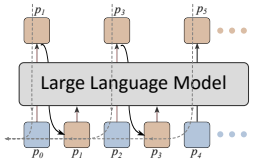
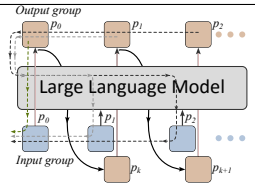
	Re-encoded streaming	Concatenated streaming	Interleaved streaming	Grouped streaming
Illustration				
Attn.	Re-encode all past caches when new input arrives to match pretraining.	Concatenate the input and output tokens into a composite token per step.	Interleave input and output tokens on the timeline.	Restrict attention within input and output groups to match pretraining.
Pos.	Reassign positions via full re-encoding.	Assign monotonic positions over concatenation.	Assign positions by interleaved time order.	Maintain separate positional spaces per group.

Table 1: Comparison of concurrent-streaming architecture adaptation methods from the perspectives of attention (Attn.) and position (Pos.). ⟶ indicates the token generation direction, while - - - -> denotes attention dependencies. ▢ blocks represent the input stream, and ▢ blocks represent the output stream. p indicates the corresponding position ID.

Streaming-In	*Bound*	*Inc.*	*Cxt.*	*Example methods*
Text	Memory	-	✓	StreamingDialogue (Macháček et al., 2023)
Audio	Causal, Memory	✓	✓	WhisperStreaming (Li et al., 2024b)
Video	Memory	-	✓	Timechat-online (Yao et al., 2025)

Table 2: Summary of **sequential streaming** tasks. Incremental encoding (*Inc*) and context management (*Ctx*) are the key technical dimensions. The checkmark (✓) indicates the scope covered by existing research.

mal interaction timing. (1) *In-context prediction* paradigm integrates decision-making directly into the autoregressive generation process (Chen et al., 2024a; Koshkin et al., 2024b). Here, the LLM is fine-tuned to emit special control tokens (e.g., <EOS> or <WAIT>) alongside standard text. This strategy unifies policy execution with language modeling, allowing the model to leverage its reasoning capabilities for control. (2) *Auxiliary decision* employ auxiliary decision modules to decouple control from generation (Zhao et al., 2025b; Chen et al., 2024b; Qian et al., 2025). This typically involves training a lightweight classifier to output a binary decision. By isolating the interaction signal, this approach allows for focused supervision on the decision boundary without interfering with the semantic distribution of the generated text.

RL-based Interaction RL-based policies model interaction control as sequential decision-making, where the LLMs selects read or write actions based on the current context (Wang et al., 2025c; Cheng et al., 2025b; Xu et al., 2025e). Optimizing quality–latency rewards enables the discovery of non-trivial interaction patterns that are difficult to encode with static rules. For example, MMDuet2 (Wang et al., 2025c) formulates proactive video interaction as an RL-driven control problem, enabling asynchronous perception and reaction under streaming video inputs.

6 Streaming Applications and Tasks

This section reviews the application-level tasks enabled by streaming LLMs, building upon the methodological taxonomy established in Sections 3–5. Notably, since output streaming is a universal property of LLM-based generation, we concentrate on task settings where streaming arises from incremental input, real-time interaction, or bidirectional coupling between input and output.

Sequential Streaming Tasks Sequential streaming tasks target long, unbounded input streams that cannot be processed in a single pass due to resource limitations. For instance, streaming long video understanding (Zhang et al., 2024b; Yao et al., 2025) requires incremental video encoding, followed by immediate decoding upon query arrival. As summarized in Table 2, different modalities emphasize distinct technical components.

Concurrent Streaming Tasks Concurrent streaming covers multimodal tasks that require simultaneous input reception and output generation. Based on processing depth, these tasks can be divided into two levels. (1) Perception-Level ($\mathcal{X} \rightarrow \mathcal{Y}$): Models focus on direct cross-modal mappings with minimal latency, including streaming translation (e.g., Seed LiveInterpret 2.0 (Cheng et al., 2025b)), ASR/TTS (e.g., CosyVoice (Du et al., 2024)), real-time video captioning (e.g., LiveCC (Chen et al., 2025a)), and streaming QA (e.g., Qwen3-Omni (Xu et al., 2025c)). (2) Cognition-Level ($\mathcal{X} \rightarrow \mathcal{Z} \rightarrow \mathcal{Y}$): Tasks require maintaining and updating a latent

Task type	*Level*	*Modality*		*Paradigm*				*Interaction policy*			*Example methods*
		In	*Out*	*R.*	*C.*	*I.*	*G.*	*Rule*	*SFT*	*RL*	
Translation	$\mathcal{X} \to \mathcal{Y}$	T/S	T/S	✓	✓	✓	✓	✓	✓	✓	Seed LiveInterpret 2.0 (Cheng et al., 2025b)
Detection	$\mathcal{X} \to \mathcal{Y}$	T/S/V	T	-	-	✓	-	-	✓	-	FineHarm (Li et al., 2025i)
ASR	$\mathcal{X} \to \mathcal{Y}$	S	T	-	-	✓	✓	✓	✓	-	ReaLLM (Seide et al., 2024), Llama-omni (Fang et al., 2024)
TTS	$\mathcal{X} \to \mathcal{Y}$	T	S	-	✓	-	✓	✓	✓	-	Cosyvoice (Du et al., 2024), DSM (Zeghidour et al., 2025)
QA	$\mathcal{X} \to \mathcal{Y}$	T/S/V	T/S	-	✓	✓	-	-	✓	✓	Qwen3-omni (Xu et al., 2025c), VideoLLM-online (Chen et al., 2024a)
Description	$\mathcal{X} \to \mathcal{Y}$	V	T	-	-	✓	✓	✓	✓	-	LiveCC (Chen et al., 2025a), StreamMind (Ding et al., 2025b)
VLA	$\mathcal{X} \to \mathcal{Y}$	V	T	-	-	✓	-	✓	-	-	StreamVLN (Wei et al., 2025b), ActiveVLN (Zhang et al., 2025e)
Reasoning	$\mathcal{X} \to \mathcal{Z}$	T/S/V	T/S	-	-	✓	✓	✓	-	-	StreamingThinker (Tong et al., 2025a)
	$\mathcal{Z} \to \mathcal{Y}$	T/S/V	T/S	-	-	✓	-	✓	-	-	AsyncReasoning (Yakushev et al., 2025)
Tool usage	$\mathcal{X} \to \mathcal{Z}$	T/S/V	T	-	-	✓	-	✓	-	-	AViLA (Zhang et al., 2025a), StreamRAG (Arora et al., 2025)
	$\mathcal{Z} \to \mathcal{Y}$	T/S/V	T	-	-	✓	-	✓	-	-	Conveyor (Xu et al., 2024), AsyncLM (Gim et al., 2024)

Table 3: Summary of **concurrent streaming** tasks and representative methods. Tasks are categorized by processing depth (*Level*), where $\mathcal{X} \to \mathcal{Y}$ denotes direct mapping (perception) and $\mathcal{X} \to \mathcal{Z} \to \mathcal{Y}$ denotes intermediate processing with a latent state $\mathcal{Z}$ (cognition). Modality: text (*T*), speech (*S*), vision (*V*). Streaming Paradigm: re-encoding (*R*), Concatenated (*C*), Interactive (*I*), Group (*G*). Interaction Policy: Rule-based (*Rule*), SFT-based (*SFT*), and RL-based (*RL*). The checkmark (✓) indicates the scope covered by existing research.

state $\mathcal{Z}$ to support complex behaviors such as streaming reasoning (e.g., StreamingThinker (Tong et al., 2025a)) and streaming tool usage (e.g., AViLA (Zhang et al., 2025a)). Here, the latent state decouples immediate perception from final output generation. We summarize the corresponding technical categories of these tasks in Table 3.

7 Future Directions

To provide a comprehensive roadmap, we categorize future research into two complementary perspectives: the *technical level* (i.e., how to build better streaming models) and the *application level* (i.e., how to apply streaming models).

Technical Level (1) ***Efficient Streaming LLMs***. Efficiency under strict latency and memory constraints remains a core challenge, involving incremental encoding, decoding acceleration, and long-term context management. (2) ***Alternative Concurrent Streaming Paradigms***. Beyond interleaved and group-based strategies, more effective streaming paradigms remain to be explored. In particular, extending streaming interaction to semi-autoregressive or block-wise generation frameworks presents a promising yet underexplored direction. (3) ***Proactive Interaction Policies***. Designing interaction policies that adaptively balance reading and generation is essential for real-time streaming performance. (4) ***Interpretability***. The behavioral dynamics of LLMs in interactive streaming settings remain largely unexplored, calling for greater interpretability.

Application Level (1) ***Expansion of Streaming Modalities***. Current streaming LLMs primarily focus on text, audio, and basic video interactions. Extending streaming LLMs to additional modalities requires transcending these limitations toward complex, omni-modal continuous streams (e.g., parallel video-audio streams) to achieve real-time streaming multimodal understanding and generation in highly dynamic environments. (2) ***Expansion of Concurrency Levels***. A promising direction is to expand current streaming LLMs from two-level perceptual concurrency (e.g., "listen-while-speaking" and "read-while-thinking") to deeper, multi-level asynchronous processing. This includes 3-level streaming (introducing streaming "perceiving, reasoning, and generation") and 4-level streaming (introducing concurrent "perceiving, reasoning, tool-using, and generation") to achieve true *multi-stream intelligence*. (3) ***Expansion of Streaming Tasks***. The application of streaming LLMs is expected to shift from simple, passive responses toward complex proactive interactions and long-context engagements. Advancing these capabilities involves empowering models to actively initiate interventions and maintain long-term memory, ultimately achieving brain-like streaming intelligence.

8 Conclusion

This survey presents a unified view of streaming LLMs by clarifying their definitions and organizing existing approaches into output-streaming, sequential-streaming, and concurrent-streaming paradigms based on data flow and interaction concurrency. We review representative methodologies and application scenarios, and discuss the fundamental challenges posed by real-time and interactive settings. We hope this work serves as a concise reference and a conceptual foundation for future research on streaming intelligence.

Limitations

This survey focuses on clarifying the conceptual landscape of Streaming Large Language Models through unified definitions, paradigms, and representative methods. As a result, it does not aim to provide an exhaustive comparison of all existing implementations or a comprehensive empirical evaluation across tasks and systems. Moreover, our discussion primarily centers on high-level design principles and paradigms, leaving detailed system-level optimizations and deployment-specific considerations for future studies.

References

Victor Agostinelli, Max Wild, Matthew Raffel, Kazi Fuad, and Lizhong Chen. 2024. Simul-llm: A framework for exploring high-quality simultaneous translation with large language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 10530–10541.

Elad Amrani, Leonid Karlinsky, and Alex Bronstein. 2025. Sample-and parameter-efficient autoregressive image models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30127–30136.

Chenxin An, Fei Huang, Jun Zhang, Shansan Gong, Xipeng Qiu, Chang Zhou, and Lingpeng Kong. 2024. Training-free long-context scaling of large language models. *arXiv preprint arXiv:2402.17463*.

Siddhant Arora, Haidar Khan, Kai Sun, Xin Luna Dong, Sajal Choudhary, Seungwhan Moon, Xinyuan Zhang, Adithya Sagar, Surya Teja Appini, Kaushik Patnaik, et al. 2025. Stream rag: Instant and accurate spoken dialogue systems with streaming tool usage. *arXiv preprint arXiv:2510.02044*.

Marianne Arriola, Aaron Gokaslan, Justin T Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sekhar Sahoo, and Volodymyr Kuleshov. 2025. Block diffusion: Interpolating between autoregressive and diffusion language models. *arXiv preprint arXiv:2503.09573*.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Aaron van den Oord. 2021. Structured denoising diffusion models in discrete state-spaces. *arXiv preprint arXiv:2107.03006*.

Richard He Bai, Zijin Gu, Tatiana Likhomanenko, and Navdeep Jaitly. 2025. Speakstream: Streaming text-to-speech with interleaved data. *arXiv preprint arXiv:2505.19206*.

Richard He Bai, Tatiana Likhomanenko, Ruixiang Zhang, Zijin Gu, Zakaria Aldeneh, and Navdeep Jaitly. 2024. dmel: Speech tokenization made simple. *arXiv preprint arXiv:2407.15835*.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. 2023a. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533.

Zalán Borsos, Matt Sharifi, Damien Vincent, Eugene Kharitonov, Neil Zeghidour, and Marco Tagliasacchi. 2023b. Soundstorm: Efficient parallel audio generation. *arXiv preprint arXiv:2305.09636*.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The revolution of multimodal large language models: a survey. *arXiv preprint arXiv:2402.12451*.

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D Lee, Deming Chen, and Tri Dao. 2024a. Medusa: Simple llm inference acceleration framework with multiple decoding heads. *arXiv preprint arXiv:2401.10774*.

Zefan Cai, Yichi Zhang, Bofei Gao, Yuliang Liu, Yucheng Li, Tianyu Liu, Keming Lu, Wayne Xiong, Yue Dong, Junjie Hu, and Xiao Wen. 2024b. Pyramidkv: Dynamic kv cache compression based on pyramidal information funneling. *arXiv preprint arXiv:2406.02069*.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. 2023. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. 2022. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11315–11325.

Dibyadip Chatterjee, Edoardo Remelli, Yale Song, Bugra Tekin, Abhay Mittal, Bharat Bhatnagar, Necati Cihan Camgoz, Shreyas Hampali, Eric Sauser, Shugao Ma, et al. 2025. Streaming videollms for real-time procedural video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22586–22598.

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. 2024a. Videollm-online: Online video large language model for streaming video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18407–18418.

Joya Chen, Ziyun Zeng, Yiqi Lin, Wei Li, Zejun Ma, and Mike Zheng Shou. 2025a. Livecc: Learning video llm with streaming speech transcription at scale. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29083–29095.

Xinjie Chen, Kai Fan, Wei Luo, Linlin Zhang, Libo Zhao, Xinggao Liu, and Zhongqiang Huang. 2024b. Divergence-guided simultaneous speech translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17799–17807.

Xueyi Chen, Keda Tao, Kele Shao, and Huan Wang. 2025b. Streamingtom: Streaming token compression for efficient video understanding. *arXiv preprint arXiv:2510.18269*.

Yilong Chen, Xiang Bai, Zhibin Wang, Chengyu Bai, Yuhan Dai, Ming Lu, and Shanghang Zhang. 2025c. Streamkv: Streaming video question-answering with segment-based kv cache retrieval and compression. *arXiv preprint arXiv:2511.07278*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Zhuohan Chen, Yizhe Zhang, Ziyang Chen, Yuhao Lin, Songlin Wang, Hao Li, Kurt Keutzer, Joseph E Gonzalez, Michael W Mahoney, and Ion Stoica. 2023. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*.

Jian Cheng, Haidong Kang, Yuxin Shao, Nan Li, Pengjun Chen, Rui Wang, Saiqin Long, Xiaochun Yang, and Lianbo Ma. 2025a. Survey on efficient large language models: Principles, algorithms, applications, and open issues. *IEEE Transactions on Neural Networks and Learning Systems*.

Shanbo Cheng, Yu Bao, Zhichao Huang, Yu Lu, Ningxin Peng, Lu Xu, Runsheng Yu, Rong Cao, Yujiao Du, Ting Han, et al. 2025b. Seed liveinterpret 2.0: End-to-end simultaneous speech-to-speech translation with your voice. *arXiv preprint arXiv:2507.17527*.

Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. 2024. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. *arXiv preprint arXiv:2407.06135*.

Cheng-Han Chiang, Xiaofei Wang, Linjie Li, Chung-Ching Lin, Kevin Lin, Shujie Liu, Zhendong Wang, Zhengyuan Yang, Hung-yi Lee, and Lijuan Wang. 2025a. Shanks: Simultaneous hearing and thinking for spoken language models. *arXiv preprint arXiv:2510.06917*.

Cheng-Han Chiang, Xiaofei Wang, Linjie Li, Chung-Ching Lin, Kevin Lin, Shujie Liu, Zhendong Wang, Zhengyuan Yang, Hung-yi Lee, and Lijuan Wang. 2025b. Stitch: Simultaneous thinking and talking with chunked reasoning for spoken language models. *arXiv preprint arXiv:2507.15375*.

Marco Comunità, Zhi Zhong, Akira Takahashi, Shiqi Yang, Mengjie Zhao, Koichi Saito, Yukara Ikemiya, Takashi Shibuya, Shusuke Takahashi, and Yuki Mitsufuji. 2024. Specmaskgit: Masked generative modeling of audio spectrograms for efficient audio synthesis and beyond. *arXiv preprint arXiv:2406.17672*.

Anonymous Contributors. 2024. Llamagen: Large language model for continuous image generation. Open-source implementation; inspired by VQ-less LLM architecture for image generation.

Trung Dang, David Aponte, Dung Tran, and Kazuhito Koishida. 2024. Livespeech: Low-latency zero-shot text-to-speech via autoregressive modeling of audio discrete codes. *arXiv preprint arXiv:2406.02897*.

Pierre V Dantas, Lucas C Cordeiro, and Waldir SS Junior. 2025. A review of state-of-the-art techniques for large language model compression. *Complex & Intelligent Systems*, 11(9):407.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.

Luciano Del Corro, Allie Del Giorno, Sahaj Agarwal, Bin Yu, Ahmed Awadallah, and Subhabrata Mukherjee. 2023. Skipdecode: Autoregressive skip decoding with batching and caching for efficient llm inference. *arXiv preprint arXiv:2307.02628*.

Keqi Deng, Wenxi Chen, Xie Chen, and Phil Woodland. 2025. Simuls2s-llm: Unlocking simultaneous inference of speech llms for speech-to-speech translation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16718–16734.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. 2025a. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*.

Xin Ding, Hao Wu, Yifan Yang, Shiqi Jiang, Qianxi Zhang, Donglin Bai, Zhibo Chen, and Ting Cao. 2025b. Streammind: Unlocking full frame rate streaming video dialogue through event-gated cognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13448–13459.

Haotian Dong, Ye Li, Rongwei Lu, Chen Tang, Shu-Tao Xia, and Zhi Wang. 2025. Vvs: Accelerating speculative decoding for visual autoregressive generation via partial verification skipping. *arXiv preprint arXiv:2511.13587*.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

Siqi Fan, Xin Jiang, Xiang Li, Xuying Meng, Peng Han, Shuo Shang, Aixin Sun, Yequan Wang, and Zhongyuan Wang. 2024. Not all layers of llms are necessary during inference. *arXiv preprint arXiv:2403.02181*.

Yingqi Fan, Anhao Zhao, Jinlan Fu, Junlong Tong, Hui Su, Yijie Pan, Wei Zhang, and Xiaoyu Shen. 2025. Visipruner: Decoding discontinuous cross-modal dynamics for efficient multimodal llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18896–18913.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2024. Llama-omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*.

Qingkai Fang, Yan Zhou, Shoutao Guo, Shaolei Zhang, and Yang Feng. 2025. Llama-omni2: Llm-based real-time spoken chatbot with autoregressive streaming speech synthesis. *arXiv preprint arXiv:2505.02625*.

Yuan Feng, Junlin Lv, Yukun Cao, Xike Xie, and S Kevin Zhou. 2024. Ada-kv: Optimizing kv cache eviction by adaptive budget allocation for efficient llm inference. *arXiv preprint arXiv:2407.11550*.

Yuan Feng, Junlin Lv, Yukun Cao, Xike Xie, and S Kevin Zhou. 2025. Identify critical kv cache in llm inference from an output perturbation perspective. *arXiv preprint arXiv:2502.03805*.

Markus Frohmann, Igor Sterner, Ivan Vulić, Benjamin Minixhofer, and Markus Schedl. 2024. Segment any text: A universal approach for robust, efficient and adaptable sentence segmentation. *arXiv preprint arXiv:2406.16678*.

Biao Fu, Kai Fan, Minpeng Liao, Yidong Chen, Xiaodong Shi, and Zhongqiang Huang. 2024a. wav2vec-s: Adapting pre-trained speech models for streaming. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11465–11480.

Biao Fu, Minpeng Liao, Kai Fan, Chengxi Li, Liang Zhang, Yidong Chen, and Xiaodong Shi. 2025a. Llms can achieve high-quality simultaneous machine translation as efficiently as offline. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20372–20395.

Biao Fu, Donglei Yu, Minpeng Liao, Chengxi Li, Yidong Chen, Kai Fan, and Xiaodong Shi. 2025b. Efficient and adaptive simultaneous speech translation with fully unidirectional architecture. *arXiv preprint arXiv:2504.11809*.

Shenghao Fu, Qize Yang, Yuan-Ming Li, Yi-Xing Peng, Kun-Yu Lin, Xihan Wei, Jian-Fang Hu, Xiaohua Xie, and Wei-Shi Zheng. 2025c. Vispeak: Visual instruction feedback in streaming videos. *arXiv preprint arXiv:2503.12769*.

Yu Fu, Zefan Cai, Abedelkadir Asi, Wayne Xiong, Yue Dong, and Wen Xiao. 2024b. Not all heads matter: A head-level kv cache compression method with integrated retrieval and reasoning. *arXiv preprint arXiv:2410.19258*.

Xiangxiang Gao, Weisheng Xie, Yiwei Xiang, and Feng Ji. 2025. Falcon: Faster and parallel inference of large language models through enhanced semi-autoregressive drafting and custom-designed decoding tree. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23933–23941.

Gemma Team. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. In *Proceedings of EMNLP-IJCNLP*.

In Gim, Seung-seob Lee, and Lin Zhong. 2024. Asynchronous llm function calling. *arXiv preprint arXiv:2412.07017*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Fabian Gloeckle, Badr Youbi Idrissi, Baptiste Rozière, David Lopez-Paz, and Gabriel Synnaeve. 2024. Better & faster large language models via multi-token prediction. *arXiv preprint arXiv:2404.19737*.

Chengyue Gong, Xuezhe Feng, Guanyi Qin, Yixin Liu, et al. 2022. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*.

Alex Graves. 2012. Connectionist temporal classification. In *Supervised sequence labelling with recurrent neural networks*, pages 61–93. Springer.

Jiatao Gu, Changhan Wang, and Jake Zhao. 2019. Levenshtein transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Dake Guo, Jixun Yao, Linhan Ma, He Wang, and Lei Xie. 2025. Streamflow: Streaming flow matching with block-wise guided attention mask for speech token decoding. *arXiv preprint arXiv:2506.23986*.

Hao-Han Guo, Yao Hu, Kun Liu, Fei-Yu Shen, Xu Tang, Yi-Chen Wu, Feng-Long Xie, Kun Xie, and Kai-Tuo Xu. 2024a. Fireredtts: A foundation text-to-speech framework for industry-level generative speech applications. *arXiv preprint arXiv:2409.03283*.

Shoutao Guo, Shaolei Zhang, and Yang Feng. 2024b. Decoder-only streaming transformer for simultaneous translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8851–8864.

Shoutao Guo, Shaolei Zhang, Zhengrui Ma, Min Zhang, and Yang Feng. 2024c. Sillm: Large language models for simultaneous machine translation. *arXiv preprint arXiv:2402.13036*.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. 2025. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15733–15744.

Xiaochuang Han, Sachin Kumar, and Yulia Tsvetkov. 2023. Ssd-lm: Semi-autoregressive simplex-based diffusion language model for text generation and modular control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11575–11596.

Jiadong Hao, Bohan Zhang, Yuchen Lu, Chengcheng Zhang, and Kunda Yang. Stylle: Style learning and latent editing for stylized text and speech generation.

Yefei He, Luoming Zhang, Weijia Wu, Jing Liu, Hong Zhou, and Bohan Zhuang. 2024. Zipcache: Accurate and efficient kv cache quantization with salient token identification. *Advances in Neural Information Processing Systems*, 37:68287–68307.

Coleman Hooper, Sehoon Kim, Hiva Mohammadzadeh, Michael W Mahoney, Yakun S Shao, Kurt Keutzer, and Amir Gholami. 2024. Kvquant: Towards 10 million context length llm inference with kv cache quantization. *Advances in Neural Information Processing Systems*, 37:1270–1303.

Chihan Huang and Hao Tang. 2025. Ctrldiff: Boosting large diffusion language models with dynamic block prediction and controllable generation. *arXiv preprint arXiv:2505.14455*.

Kuan-Po Huang, Shu-wen Yang, Huy Phan, Bo-Ru Lu, Byeonggeun Kim, Sashank Macha, Qingming Tang, Shalini Ghosh, Hung-yi Lee, Chieh-Chi Kao, et al. 2025a. Impact: Iterative mask-based parallel decoding for text-to-audio generation with diffusion modeling. *arXiv preprint arXiv:2506.00736*.

Wenxuan Huang, Zijie Zhai, Yunhang Shen, Shaosheng Cao, Fei Zhao, Xiangfeng Xu, Zheyu Ye, Yao Hu, and Shaohui Lin. 2024. Dynamic-llava: Efficient multimodal large language models via dynamic vision-language context sparsification. *arXiv preprint arXiv:2412.00876*.

Xiaohu Huang, Hao Zhou, and Kai Han. 2025b. Prunevid: Visual token pruning for efficient video large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19959–19973.

Hyeonbin Hwang, Byeongguk Jeon, Seungone Kim, Jiyeon Kim, Hoyeon Chang, Sohee Yang, Seungpil Won, Dohaeng Lee, Youbin Ahn, and Minjoon Seo. 2025. Let's predict sentence by sentence. *arXiv preprint arXiv:2505.22202*.

Javier Iranzo-Sánchez, Jorge Iranzo-Sánchez, Adrià Giménez, Jorge Civera, and Alfons Juan. 2024. Segmentation-free streaming machine translation. *Transactions of the Association for Computational Linguistics*, 12:1104–1121.

Doohyuk Jang, Sihwan Park, June Yong Yang, Yeonsung Jung, Jihun Yun, Souvik Kundu, Sung-Yub Kim, and Eunho Yang. 2024. Lantern: Accelerating visual autoregressive models with relaxed speculative decoding. *arXiv preprint arXiv:2410.03355*.

Dongya Jia, Zhuo Chen, Jiawei Chen, Chenpeng Du, Jian Wu, Jian Cong, Xiaobin Zhuang, Chumin Li, Zhen Wei, Yuping Wang, and Yuxuan Wang. 2025. Ditar: Diffusion transformer autoregressive modeling for speech generation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 27255–27270.

Xinqi Jin, Hanxun Yu, Bohan Yu, Kebin Liu, Jian Liu, Keda Tao, Yixuan Pei, Huan Wang, Fan Dang, Jiangchuan Liu, et al. 2025. Streamingassistant: Efficient visual token pruning for accelerating online video understanding. *arXiv preprint arXiv:2512.12560*.

Sehoon Kim, Karttikeya Mangalam, Suhong Moon, Jitendra Malik, Michael W Mahoney, Amir Gholami, and Kurt Keutzer. 2023. Speculative decoding with big little decoder. *Advances in Neural Information Processing Systems*, 36:39236–39256.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, Jose Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. 2024. Videopoet: A large language model for zero-shot video generation. In *International Conference on Machine Learning*, pages 25105–25124. PMLR.

Roman Koshkin, Katsuhito Sudoh, and Satoshi Nakamura. 2024a. Llms are zero-shot context-aware simultaneous translators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1207.

Roman Koshkin, Katsuhito Sudoh, and Satoshi Nakamura. 2024b. Transllama: Llm-based simultaneous translation system. *arXiv preprint arXiv:2402.04636*.

Pin-Jui Ku, He Huang, Jean-Marie Lemercier, Subham Sekhar Sahoo, Zhehuai Chen, and Ante Jukić. 2025. Discrete diffusion for generative modeling of text-aligned speech tokens. *arXiv preprint arXiv:2509.20060*.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Chenyang Le, Bing Han, Jinshun Li, Songyong Chen, and Yanmin Qian. 2025. Simulmega: Moe routers are advanced policy makers for simultaneous speech translation. *arXiv preprint arXiv:2509.01200*.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in neural information processing systems*, 36:14005–14034.

Haodong Lei, Hongsong Wang, Xin Geng, Liang Wang, and Pan Zhou. 2025. Fast inference of visual autoregressive model with adjacency-adaptive dynamical draft trees. *arXiv preprint arXiv:2512.21857*.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. Fast inference from transformers via speculative decoding. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

Bohan Li, Zhihan Li, Haoran Wang, Hanglei Zhang, Yiwei Guo, Hankun Wang, Xie Chen, and Kai Yu. 2025a. Robust and efficient autoregressive speech synthesis with dynamic chunk-wise prediction policy. *arXiv preprint arXiv:2506.22023*.

Bohan Li, Hankun Wang, Situo Zhang, Yiwei Guo, and Kai Yu. 2025b. Fast and high-quality autoregressive speech synthesis via speculative decoding. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Han Li, Xinyu Peng, Yaoming Wang, Zelin Peng, Xin Chen, Rongxiang Weng, Jingang Wang, Xunliang Cai, Wenrui Dai, and Hongkai Xiong. 2025c. Onecat: Decoder-only auto-regressive model for unified understanding and generation. *arXiv preprint arXiv:2509.03498*.

Haoyang Li, Yiming Li, Anxin Tian, Tianhao Tang, Zhanchao Xu, Xuejia Chen, Nicole Hu, Wei Dong, Qing Li, and Lei Chen. 2024a. A survey on large language model acceleration based on kv cache management. *arXiv preprint arXiv:2412.19442*.

Jia-Nan Li, Quan Tu, Cunli Mao, Zhengtao Yu, Ji-Rong Wen, and Rui Yan. 2024b. Streamingdialogue: Prolonged dialogue learning via long context compression with minimal losses. *Advances in Neural Information Processing Systems*, 37:86074–86101.

Jiajun Li, Yue Ma, Xinyu Zhang, Qingyan Wei, Songhua Liu, and Linfeng Zhang. 2025d. Skipvar: Accelerating visual autoregressive modeling via adaptive frequency-aware skipping. *arXiv preprint arXiv:2506.08908*.

Kunjun Li, Zigeng Chen, Cheng-Yen Yang, and Jenq-Neng Hwang. 2025e. Memory-efficient visual autoregressive modeling with scale-aware kv cache compression. *arXiv preprint arXiv:2505.19602*.

Ruanjun Li, Yuedong Tan, Yuanming Shi, and Jiawei Shao. 2025f. Videoscan: Enabling efficient streaming video understanding via frame-level semantic carriers. *arXiv preprint arXiv:2503.09387*.

Shufan Li, Konstantinos Kallidromitis, Hritik Bansal, Akash Gokul, Yusuke Kato, Kazuki Kozuka, Jason Kuen, Zhe Lin, Kai-Wei Chang, and Aditya Grover. 2025g. Lavida: A large diffusion language model for multimodal understanding. *arXiv preprint arXiv:2505.16839*.

Wei Li, Bing Hu, Rui Shao, Leyang Shen, and Liqiang Nie. 2025h. Lion-fs: Fast & slow video-language thinker as online video assistant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3240–3251.

Xiang Lisa Li, John Thickstun Zhao, James Diffenderfer, Xuezhe He, Percy Liang, and Graham Neubig. 2022. Diffusion-LM improves controllable text generation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yang Li, Qiang Sheng, Yehan Yang, Xueyao Zhang, and Juan Cao. 2025i. From judgment to interference: Early stopping llm harmful outputs via streaming content monitoring. *arXiv preprint arXiv:2506.09996*.

Ying Li, chengfei lv, and Huan Wang. 2025j. Freqexit: Enabling early-exit inference for visual autoregressive models via frequency-aware guidance. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. 2024c. Snapkv: Llm knows what you are looking for before generation. *Advances in Neural Information Processing Systems*, 37:22947–22970.

Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. 2024d. Eagle-2: Faster inference of language models with dynamic draft trees. *arXiv preprint arXiv:2406.16858*.

Zinuo Li, Xian Zhang, Yongxin Guo, Mohammed Bennamoun, Farid Boussaid, Girish Dwivedi, Luqi Gong, and Qiuhong Ke. 2025k. Watch and listen: Understanding audio-visual-speech moments with multimodal llm. *arXiv preprint arXiv:2505.18110*.

Mengqi Liao, Lu Wang, Chaoyun Zhang, Zekai Shen, Xiaowei Mao, Si Qin, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Huaiyu Wan. 2025. G-KV: Decoding-time KV cache eviction with global attention. *arXiv preprint arXiv:2512.00504*.

Junyan Lin, Junlong Tong, Hao Wu, Jialiang Zhang, Jinming Liu, Xin Jin, and Xiaoyu Shen. 2026. Speak while watching: Unleashing true real-time video understanding capability of multimodal large language models. *arXiv preprint arXiv:2601.06843*.

Zijian Lin, Yang Zhang, Yougen Yuan, Yuming Yan, Jinjiang Liu, Zhiyong Wu, Pengfei Hu, and Qun Yu. 2025. Accelerating autoregressive speech synthesis inference with speech speculative decoding. *arXiv preprint arXiv:2505.15380*.

Aiwei Liu, Minghua He, Shaoxun Zeng, Sijun Zhang, Linhao Zhang, Chuhan Wu, Wei Jia, Yuan Liu, Xiao Zhou, and Jie Zhou. 2025a. Wedlm: Reconciling diffusion language models with standard causal attention for fast inference. *arXiv preprint arXiv:2512.22737*.

Jiahao Liu, Qifan Wang, Jingang Wang, and Xunliang Cai. 2024a. Speculative decoding via early-exiting for faster llm inference with thompson sampling control mechanism. *arXiv preprint arXiv:2406.03853*.

Jiaheng Liu, Dawei Zhu, Zhiqi Bai, Yancheng He, Huanxuan Liao, Haoran Que, Zekun Wang, Chenchen Zhang, Ge Zhang, Jiebin Zhang, et al. 2025b. A comprehensive survey on long context language modeling. *arXiv preprint arXiv:2503.17407*.

Jiesong Liu, Brian Park, and Xipeng Shen. 2025c. A drop-in solution for on-the-fly adaptation of speculative decoding in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9778–9794.

Jihao Liu, Zhiding Yu, Shiyi Lan, Shihao Wang, Rongyao Fang, Jan Kautz, Hongsheng Li, and Jose M Alvare. 2024b. Streamchat: Chatting with streaming video. *arXiv preprint arXiv:2412.08646*.

Shang Liu, Yao Lu, Wenji Fang, Jing Wang, and Zhiyao Xie. 2025d. Sync-llm: Generation of large-scale synthetic circuit code with hierarchical language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17361–17376.

Tianqiao Liu, Xueyi Li, Hao Wang, Haoxuan Li, Zhichao Chen, Weiqi Luo, and Zitao Liu. 2025e. From text to talk: Audio-language model needs non-autoregressive joint training. *arXiv preprint arXiv:2509.20072*.

Wenrui Liu, Qian Chen, Wen Wang, Guanrou Yang, Weiqin Li, Minghui Fang, Jialong Zuo, Xiaoda Yang, Tao Jin, Jin Xu, et al. 2025f. Speech token prediction via compressed-to-fine language modeling for speech generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10632–10641.

Xiang Liu, Zhenheng Tang, Peijie Dong, Zeyu Li, Yue Liu, Bo Li, Xuming Hu, and Xiaowen Chu. 2025g. Chunkkv: Semantic-preserving kv cache compression for efficient long-context llm inference. *arXiv preprint arXiv:2502.00299*.

Yiheng Liu, Liao Qu, Huichao Zhang, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Xian Li, Shuai Wang, Daniel K Du, et al. 2025h. Detailflow: 1d coarse-to-fine autoregressive image generation via next-detail prediction. *arXiv preprint arXiv:2505.21473*.

Yuxuan Liu et al. 2024c. Kivi: A tuning-free asymmetric 2-bit quantization for KV cache. *arXiv preprint arXiv:2402.02750*.

Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. 2023. Scissorhands: Exploiting the persistence of importance hypothesis for llm kv cache compression at test time. *Advances in Neural Information Processing Systems*, 36:52342–52364.

Yen-Ju Lu, Yashesh Gaur, Wei Zhou, Benjamin Muller, Jesus Villalba, Najim Dehak, Luke Zettlemoyer, Gargi Ghosh, Mike Lewis, Srinivasan Iyer, et al. 2025. Latent speech-text transformer. *arXiv preprint arXiv:2510.06195*.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, et al. 2019. Stacl: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3036.

Dominik Macháček, Raj Dabre, and Ondřej Bojar. 2023. Turning whisper into real-time transcription system. *arXiv preprint arXiv:2307.14743*.

Benjamin Minixhofer, Jonas Pfeiffer, and Ivan Vulić. 2023. Where's the point? self-supervised multilingual punctuation-agnostic sentence segmentation. *arXiv preprint arXiv:2305.18893*.

Tan Dat Nguyen, Ji-Hoon Kim, Jeongsoo Choi, Shukjae Choi, Jinseok Park, Younglo Lee, and Joon Son Chung. 2025. Accelerating codec-based speech synthesis with multi-token prediction and speculative decoding. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. Large language diffusion models. *arXiv preprint arXiv:2502.09992*.

Xuefei Ning, Zinan Lin, Zixuan Zhou, Zifu Wang, Huazhong Yang, and Yu Wang. 2023. Skeleton-of-thought: Large language models can do parallel decoding. *Proceedings ENLSP-III*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Siqi Ouyang, Xi Xu, and Lei Li. 2025. Infinisst: Simultaneous translation of unbounded speech with large language model. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3032–3046.

Sunny Panchal, Apratim Bhattacharyya, Guillaume Berger, Antoine Mercier, Cornelius Böhm, Florian Dietrichkeit, Reza Pourreza, Xuanlin Li, Pulkit Madan, Mingu Lee, et al. 2024. What to say and when to say it: Live fitness coaching as a testbed for situated interaction. *Advances in Neural Information Processing Systems*, 37:75853–75882.

William Peebles and Saining Xie. 2022. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*.

Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. 2025. Dispider: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24045–24055.

Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. 2024. Streaming long video understanding with large language models. *Advances in Neural Information Processing Systems*, 37:119336–119360.

Zhen Qin, Weigao Sun, Dong Li, Xuyang Shen, Weixuan Sun, and Yiran Zhong. 2024. Lightning attention-2: A free lunch for handling unlimited sequence lengths in large language models. *arXiv preprint arXiv:2401.04658*.

Ziran Qin, Yuchen Cao, Mingbao Lin, Wen Hu, Shixuan Fan, Ke Cheng, Weiyao Lin, and Jianguo Li. 2025a. Cake: Cascading and adaptive kv cache eviction with layer preferences. *arXiv preprint arXiv:2503.12491*.

Ziran Qin, Youru Lv, Mingbao Lin, Zeren Zhang, Chanfan Gan, Tieyuan Chen, and Weiyao Lin. 2025b. Autoregressive image generation needs only a few lines of cached tokens. *arXiv preprint arXiv:2512.04857*.

Ziran Qin, Youru Lv, Mingbao Lin, Zeren Zhang, Danping Zou, and Weiyao Lin. 2025c. Head-aware kv cache compression for efficient visual autoregressive modeling. *arXiv preprint arXiv:2504.09261*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Matthew Raffel, Victor Agostinelli, and Lizhong Chen. 2024. Simultaneous masking, not prompting optimization: A paradigm shift in fine-tuning llms for simultaneous translation. *arXiv preprint arXiv:2405.10443*.

Aditya Ramesh et al. 2021. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*.

Shuhuai Ren, Shuming Ma, Xu Sun, and Furu Wei. 2025. Next block prediction: Video generation via semi-autoregressive modeling. *arXiv preprint arXiv:2502.07737*.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Laura Ruis, Mitchell Stern, Julia Proskurnia, and William Chan. 2020. Insertion-deletion transformer. *arXiv preprint arXiv:2001.05540*.

Frank Seide, Morrie Doulaty, Yangyang Shi, Yashesh Gaur, Junteng Jia, and Chunyang Wu. 2024. Speech reallm–real-time streaming speech recognition with multimodal llms by teaching the flow of time. *arXiv preprint arXiv:2406.09569*.

Zhengyan Sheng, Zhihao Du, Shiliang Zhang, Zhijie Yan, Yexin Yang, and Zhenhua Ling. 2025. Syncspeech: Low-latency and efficient dual-stream text-to-speech based on temporal masked transformer. *arXiv preprint arXiv:2502.11094*.

Mohan Shi, Yuchun Shu, Lingyun Zuo, Qian Chen, Shiliang Zhang, Jie Zhang, and Li-Rong Dai. 2023. Semantic vad: Low-latency voice activity detection for speech interaction. *arXiv preprint arXiv:2305.12450*.

Sambal Shikhar, Mohammed Irfan Kurpath, Sahal Shaji Mullappilly, Jean Lahoud, Fahad Shahbaz Khan, Rao Muhammad Anwer, Salman Khan, and Hisham Cholakkal. 2025. Llmvox: Autoregressive streaming text-to-speech model for any llm. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20481–20493.

Junhyuk So, Juncheol Shin, Hyunho Kook, and Eunhyeok Park. 2025. Grouped speculative decoding for autoregressive image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15375–15384.

Yuxuan Song, Zheng Zhang, Cheng Luo, Pengyang Gao, Fan Xia, Hao Luo, Zheng Li, Yuehang Yang, Hongli Yu, Xingwei Qu, et al. 2025. Seed diffusion: A large-scale diffusion language model with high-speed inference. *arXiv preprint arXiv:2508.02193*.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. 2024. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*.

Hanlin Tang, Yang Lin, Jing Lin, Qingsen Han, Shikuan Hong, Yiwu Yao, and Gongyi Wang. 2024. Razorattention: Efficient kv cache compression through retrieval heads. *arXiv preprint arXiv:2407.15891*.

Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. 2025. Dycoke: Dynamic compression of tokens for fast video large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18992–19001.

Chameleon Team. 2024. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Yao Teng, Han Shi, Xian Liu, Xuefei Ning, Guohao Dai, Yu Wang, Zhenguo Li, and Xihui Liu. 2024. Accelerating auto-regressive text-to-image generation with training-free speculative jacobi decoding. *arXiv preprint arXiv:2410.01699*.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. 2024. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865.

Yuchuan Tian, Yuchen Liang, Jiacheng Sun, Shuo Zhang, Guangwen Yang, Yingte Shu, Sibo Fang, Tianyu Guo, Kai Han, Chao Xu, et al. 2025. From next-token to next-block: A principled adaptation path for diffusion llms. *arXiv preprint arXiv:2512.06776*.

Junlong Tong, Yingqi Fan, Anhao Zhao, Yunpu Ma, and Xiaoyu Shen. 2025a. Streamingthinker: Large language models can think while reading. *arXiv preprint arXiv:2510.17238*.

Junlong Tong, Jinlan Fu, Zixuan Lin, Yingqi Fan, Anhao Zhao, Hui Su, and Xiaoyu Shen. 2025b. Llm as effective streaming processor: Bridging streaming-batch mismatches with group position encoding. *arXiv preprint arXiv:2505.16983*.

Genshun Wan, Wenhui Zhang, Jing-Xuan Zhang, Shifu Xiong, Jianqing Gao, and Zhongfu Ye. 2026. Streaming speech recognition with decoder-only large language models and latency optimization. *arXiv preprint arXiv:2601.22779*.

Ao Wang, Hui Chen, Jiaxin Li, Jianchao Tan, Kefeng Zhang, Xunliang Cai, Zijia Lin, Jungong Han, and Guiguang Ding. 2024a. Prefixkv: Adaptive prefix kv cache is what vision instruction-following models need for efficient generation. *arXiv preprint arXiv:2412.03409*.

Chunqi Wang, Ji Zhang, Haiqing Chen, Chenghao Tao, et al. 2018. Semi-autoregressive neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ArXiv:1808.08583.

Haibo Wang, Bo Feng, Zhengfeng Lai, Mingze Xu, Shiyu Li, Weifeng Ge, Afshin Dehghan, Meng Cao, and Ping Huang. 2025a. Streambridge: Turning your offline video large language model into a proactive streaming assistant. *arXiv preprint arXiv:2505.05467*.

Haoyu Wang, Guoqiang Hu, Guodong Lin, Wei-Qiang Zhang, and Jian Li. 2024b. Simul-whisper: Attention-guided streaming whisper with truncation detection. *arXiv preprint arXiv:2406.10052*.

Minghan Wang, Thuy Vu, Jinming Zhao, Fatemeh Shiri, Ehsan Shareghi, and Gholamreza Haffari. 2024c. Simultaneous machine translation with large language models. In *Proceedings of the 22nd Annual Workshop of the Australasian Language Technology Association*, pages 89–103.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2023. Augmenting language models with long-term memory. *Advances in Neural Information Processing Systems*, 36:74530–74543.

Xindi Wang, Mahsa Salmani, Parsa Omidi, Xiangyu Ren, Mehdi Rezagholizadeh, and Armaghan Eshaghi. 2024d. Beyond the limits: A survey of techniques to extend the context length in large language models. *arXiv preprint arXiv:2402.02244*.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. 2024e. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*.

Yiyu Wang, Xuyang Liu, Xiyan Gui, Xinying Lin, Boxue Yang, Chenfei Liao, Tailai Chen, and Linfeng Zhang. 2025b. Accelerating streaming video large language models via hierarchical token compression. *arXiv preprint arXiv:2512.00891*.

Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. 2024f. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. *arXiv preprint arXiv:2409.00750*.

Yueqian Wang, Songxiang Liu, Disong Wang, Nuo Xu, Guanglu Wan, Huishuai Zhang, and Dongyan Zhao. 2025c. Mmduet2: Enhancing proactive interaction of video mllms with multi-turn reinforcement learning. *arXiv preprint arXiv:2512.06810*.

Yuhao Wang, Heyang Liu, Ziyang Cheng, Ronghua Wu, Qunshan Gu, Yanfeng Wang, and Yu Wang. 2025d. Vocalnet: Speech llm with multi-token prediction for faster and high-quality generation. *arXiv preprint arXiv:2504.04060*.

Yuqing Wang, Shuhuai Ren, Zhijie Lin, Yujin Han, Haoyuan Guo, Zhenheng Yang, Difan Zou, Jiashi

Feng, and Xihui Liu. 2025e. Parallelized autoregressive visual generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12955–12965.

Yuxuan Wang, Yiqi Song, Cihang Xie, Yang Liu, and Zilong Zheng. 2025f. Videollamb: Long streaming video understanding with recurrent memory bridges. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24170–24181.

Zili Wang, Robert Zhang, Kun Ding, Qi Yang, Fei Li, and Shiming Xiang. 2024g. Continuous speculative decoding for autoregressive image generation. *arXiv preprint arXiv:2411.11925*.

Chiyue Wei, Cong Guo, Junyao Zhang, Haoxuan Shan, Yifan Xu, Ziyue Zhang, Yudong Liu, Qinsi Wang, Changchun Zhou, Hai Li, et al. 2025a. Focus: A streaming concentration architecture for efficient vision-language models. *arXiv preprint arXiv:2512.14661*.

Meng Wei, Chenyang Wan, Xiqian Yu, Tai Wang, Yuqiang Yang, Xiaohan Mao, Chenming Zhu, Wenzhe Cai, Hanqing Wang, Yilun Chen, et al. 2025b. Streamvln: Streaming vision-and-language navigation via slowfast context modeling. *arXiv preprint arXiv:2507.05240*.

Zhuofan Wen, Shangtong Gui, and Yang Feng. 2024. Speculative decoding with ctc-based draft model for llm inference acceleration. *Advances in Neural Information Processing Systems*, 37:92082–92100.

Haibin Wu, Naoyuki Kanda, Sefik Emre Eskimez, and Jinyu Li. 2024a. Ts3-codec: Transformer-based simple streaming single codec. *arXiv preprint arXiv:2411.18803*.

Hao Wu, Yingqi Fan, Jinyang Dai, Junlong Tong, Yunpu Ma, and Xiaoyu Shen. 2026a. Hidrop: Hierarchical vision token reduction in mllms via late injection, concave pyramid pruning, and early exit. *arXiv preprint arXiv:2602.23699*.

Hao Wu, Junlong Tong, Xudong Wang, Yang Tan, Changyu Zeng, Anastasia Antsiferova, and Xiaoyu Shen. 2026b. From data to model: A survey of the compression lifecycle in mllms. *TechRxiv preprint TechRxiv:177220375.55495124*.

Shiwei Wu, Joya Chen, Kevin Qinghong Lin, Qimeng Wang, Yan Gao, Qianli Xu, Tong Xu, Yao Hu, Enhong Chen, and Mike Zheng Shou. 2024b. Videollm-mod: Efficient video-language streaming with mixture-of-depths vision computation. *Advances in Neural Information Processing Systems*, 37:109922–109947.

Yecheng Wu, Han Cai, Junyu Chen, Zhuoyang Zhang, Enze Xie, Jincheng Yu, Junsong Chen, Jinyi Hu, Yao Lu, and Song Han. 2025. Dc-ar: Efficient masked autoregressive image generation with deep compression hybrid tokenizer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18034–18045.

Jiaer Xia, Peixian Chen, Mengdan Zhang, Xing Sun, and Kaiyang Zhou. 2025a. Streaming video instruction tuning. *arXiv preprint arXiv:2512.21334*.

Yinfeng Xia, Huiyan Li, Chenyang Le, Manhong Wang, Yutao Sun, Xingyang Ma, and Yanmin Qian. 2025b. Mfla: Monotonic finite look-ahead attention for streaming speech recognition. *arXiv preprint arXiv:2506.03722*.

Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. 2024a. Duoattention: Efficient long-context llm inference with retrieval and streaming heads. *arXiv preprint arXiv:2410.10819*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*.

Zilin Xiao, Hongming Zhang, Tao Ge, Siru Ouyang, Vicente Ordonez, and Dong Yu. 2024b. Parallelspec: Parallel drafter for efficient speculative decoding. *arXiv preprint arXiv:2410.05589*.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. 2024. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*.

Roy Xie, David Qiu, Deepak Gopinath, Dong Lin, Yanchao Sun, Chong Wang, Saloni Potdar, and Bhuwan Dhingra. 2025. Interleaved reasoning for large language models via reinforcement learning. *arXiv preprint arXiv:2505.19640*.

Zhifei Xie and Changqiao Wu. 2024. Mini-omni: Language models can hear, talk while thinking in streaming. *arXiv preprint arXiv:2408.16725*.

Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, et al. 2025. Lumina-dimoo: An omni diffusion large language model for multimodal generation and understanding. *arXiv preprint arXiv:2510.06308*.

Z Xin, Z Dong, L Shimin, Z Yaqian, and Q Xipeng. 2024. Speechtokenizer: Unified speech tokenizer for speech language models. In *Proc. Int. Conf. Learn. Representations*, pages 1–21.

Boxun Xu, Yu Wang, Zihu Wang, and Peng Li. 2025a. Ams-kv: Adaptive kv caching in multi-scale visual autoregressive transformers. *arXiv preprint arXiv:2511.16047*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. 2025b. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. 2025c. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

Ruyi Xu, Guangxuan Xiao, Yukang Chen, Liuning He, Kelly Peng, Yao Lu, and Song Han. 2025d. Streamingvlm: Real-time understanding for infinite video streams. *arXiv preprint arXiv:2510.09608*.

Ting Xu, Zhichao Huang, Jiankai Sun, Shanbo Cheng, and Wai Lam. 2025e. Seqpo-simt: Sequential policy optimization for simultaneous machine translation. *arXiv preprint arXiv:2505.20622*.

Yechen Xu, Xinhao Kong, Tingjun Chen, and Danyang Zhuo. 2024. Conveyor: Efficient tool-aware llm serving with tool partial execution. *arXiv preprint arXiv:2406.00059*.

George Yakushev, Nataliia Babina, Masoud Vahid Dastgerdi, Vyacheslav Zhdanovskiy, Alina Shutova, and Denis Kuznedelev. 2025. Asynchronous reasoning: Training-free interactive thinking llms. *arXiv preprint arXiv:2512.10931*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Dongjie Yang, XiaoDong Han, Yan Gao, Yao Hu, Shilin Zhang, and Hai Zhao. 2024a. Pyramidinfer: Pyramid kv cache compression for high-throughput llm inference. *arXiv preprint arXiv:2405.12532*.

Haolin Yang, Feilong Tang, Lingxiao Zhao, Xiang An, Ming Hu, Huifa Li, Xinlin Zhuang, Yifan Lu, Xiaofeng Zhang, Abdalla Swikir, et al. 2025b. Streamagent: Towards anticipatory agents for streaming video understanding. *arXiv preprint arXiv:2508.01875*.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. 2025c. Mmada: Multimodal large diffusion language models. *arXiv preprint arXiv:2505.15809*.

Shang Yang, Junxian Guo, Haotian Tang, Qinghao Hu, Guangxuan Xiao, Jiaming Tang, Yujun Lin, Zhijian Liu, Yao Lu, and Song Han. 2025d. Lserve: Efficient long-sequence llm serving with unified sparse attention. *arXiv preprint arXiv:2502.14866*.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. 2023. Gated linear attention transformers with hardware-efficient training. *arXiv preprint arXiv:2312.06635*.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. 2024b. Parallelizing linear transformers with the delta rule over sequence length. *Advances in neural information processing systems*, 37:115491–115522.

Yanlai Yang, Zhuokai Zhao, Satya Narayan Shukla, Aashu Singh, Shlok Kumar Mishra, Lizhu Zhang, and Mengye Ren. 2025e. Streammem: Query-agnostic kv cache memory for streaming video understanding. *arXiv preprint arXiv:2508.15717*.

Yifan Yang, Shujie Liu, Jinyu Li, Yuxuan Hu, Haibin Wu, Hui Wang, Jianwei Yu, Lingwei Meng, Haiyang Sun, Yanqing Liu, et al. 2025f. Pseudo-autoregressive neural codec language models for efficient zero-shot text-to-speech synthesis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9316–9325.

Yifan Yang, Ziyang Ma, Shujie Liu, Jinyu Li, Hui Wang, Lingwei Meng, Haiyang Sun, Yuzhe Liang, Ruiyang Xu, Yuxuan Hu, et al. 2024c. Interleaved speech-text language models are simple streaming text to speech synthesizers. *arXiv preprint arXiv:2412.16102*.

Zeyu Yang, Lai Wei, Roman Koshkin, Xi Chen, and Satoshi Nakamura. 2025g. Sasst: Leveraging syntax-aware chunking and llms for simultaneous speech translation. *arXiv preprint arXiv:2508.07781*.

Zhenyu Yang, Yuhang Hu, Zemin Du, Dizhan Xue, Shengsheng Qian, Jiahong Wu, Fan Yang, Weiming Dong, and Changsheng Xu. 2025h. Svbench: A benchmark with temporal multi-turn dialogues for streaming video understanding. *arXiv preprint arXiv:2502.10810*.

Zhenyu Yang, Kairui Zhang, Yuhang Hu, Bing Wang, Shengsheng Qian, Bin Wen, Fan Yang, Tingting Gao, Weiming Dong, and Changsheng Xu. 2025i. Livestar: Live streaming assistant for real-world online video understanding. *arXiv preprint arXiv:2511.05299*.

Moran Yanuka, Paul Dixon, Eyal Finkelshtein, Daniel Rotman, and Raja Giryes. 2025. Principled coarse-grained acceptance for speculative decoding in speech. *arXiv preprint arXiv:2511.13732*.

Linli Yao, Yicheng Li, Yuancheng Wei, Lei Li, Shuhuai Ren, Yuanxin Liu, Kun Ouyang, Lean Wang, Shicheng Li, Sida Li, et al. 2025. Timechat-online: 80% visual tokens are naturally redundant in streaming videos. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10807–10816.

Yao Yao, Zuchao Li, and Hai Zhao. 2024. Sirllm: Streaming infinite retentive llm. *arXiv preprint arXiv:2405.12528*.

Zhen Ye, Peiwen Sun, Jiahe Lei, Hongzhan Lin, Xu Tan, Zheqi Dai, Qiuqiang Kong, Jianyi Chen, Jiahao Pan, Qifeng Liu, et al. 2025. Codec does matter: Exploring the semantic shortcoming of codec for audio

language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25697–25705.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. 2025. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22963–22974.

Wenyi Yu, Siyin Wang, Xiaoyu Yang, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Guangzhi Sun, Lu Lu, Yuxuan Wang, and Chao Zhang. 2024. Salmonn-omni: A codec-free LLM for full-duplex speech understanding and generation. *arXiv preprint arXiv:2411.18138*.

Neil Zeghidour, Eugene Kharitonov, Manu Orsini, Václav Volhejn, Gabriel de Marmiesse, Edouard Grave, Patrick Pérez, Laurent Mazaré, and Alexandre Défossez. 2025. Streaming sequence-to-sequence learning with delayed streams modeling. *arXiv preprint arXiv:2509.08753*.

Xiangyu Zeng, Kefan Qiu, Qingyu Zhang, Xinhao Li, Jing Wang, Jiaxin Li, Ziang Yan, Kun Tian, Meng Tian, Xinhai Zhao, et al. 2025. Streamforest: Efficient online video understanding with persistent event memory. *arXiv preprint arXiv:2509.24871*.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023a. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024a. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

Gengyuan Zhang, Tanveer Hannan, Hermine Kleiner, Beste Aydemir, Xinyu Xie, Jian Lan, Thomas Seidl, Volker Tresp, and Jindong Gu. 2025a. Avila: Asynchronous vision-language agent for streaming multimodal data interaction. *arXiv preprint arXiv:2506.18472*.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. 2024b. Flash-vstream: Memory-based real-time understanding for long video streams. *arXiv preprint arXiv:2406.08085*.

Jialiang Zhang, Junlong Tong, Junyan Lin, Hao Wu, , Yirong Sun, Yunpu Ma, and Xiaoyu Shen. 2026. Think-as-you-see: Streaming chain-of-thought reasoning for large vision-language models. *arXiv preprint arXiv:2603.02872*.

Shaolei Zhang, Qingkai Fang, Shoutao Guo, Zhengrui Ma, Min Zhang, and Yang Feng. 2024c. Streamspeech: Simultaneous speech-to-speech translation with multi-task learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8964–8986.

Shaolei Zhang and Yang Feng. 2023. End-to-end simultaneous speech translation with differentiable segmentation. *arXiv preprint arXiv:2305.16093*.

Shaolei Zhang, Shoutao Guo, Qingkai Fang, Yan Zhou, and Yang Feng. 2025b. Stream-omni: Simultaneous multimodal interactions with large language-vision-speech model. *arXiv preprint arXiv:2506.13642*.

Xuan Zhang, Cunxiao Du, Chao Du, Tianyu Pang, Wei Gao, and Min Lin. 2024d. Simlayerkv: A simple framework for layer-level kv cache reduction.

Yanqi Zhang, Yuwei Hu, Runyuan Zhao, John Lui, and Haibo Chen. 2024e. Unifying kv cache compression for large language models with leankv. *arXiv preprint arXiv:2412.03131*.

Yichi Zhang, Xin Luna Dong, Zhaojiang Lin, Andrea Madotto, Anuj Kumar, Babak Damavandi, Joyce Chai, and Seungwhan Moon. 2025c. Proactive assistant dialogue generation from streaming egocentric videos. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12055–12079.

Yulin Zhang, Cheng Shi, Yang Wang, and Sibei Yang. 2025d. Eyes wide open: Ego proactive video-llm for streaming video. *arXiv preprint arXiv:2510.14560*.

Zekai Zhang, Weiye Zhu, Hewei Pan, Xiangchen Wang, Rongtao Xu, Xing Sun, and Feng Zheng. 2025e. Activevln: Towards active exploration via multi-turn rl in vision-and-language navigation. *arXiv preprint arXiv:2509.12618*.

Zeyu Zhang, Shuning Chang, Yuanyu He, Yizeng Han, Jiasheng Tang, Fan Wang, and Bohan Zhuang. 2025f. Blockvid: Block diffusion for high-quality and consistent minute-long video generation. *arXiv preprint arXiv:2511.22973*.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, et al. 2023b. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in Neural Information Processing Systems*, 36:34661–34710.

Anhao Zhao, Fanghua Ye, Yingqi Fan, Junlong Tong, Zhiwei Fei, Hui Su, and Xiaoyu Shen. 2025a. Skipgpt: Dynamic layer pruning reinvented with token awareness and module decoupling. *arXiv preprint arXiv:2506.04179*.

Libo Zhao, Jing Li, and Ziqian Zeng. 2024. Psfuture: A pseudo-future-based zero-shot adaptive policy for simultaneous machine translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1869–1881.

Libo Zhao, Jing Li, and Ziqian Zeng. 2025b. Drfrattn: Directly learn adaptive policy from attention for simultaneous machine translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34881–34894.

Yucheng Zhao, Chong Luo, Chuanxin Tang, Dongdong Chen, Noel Codella, and Zheng-Jun Zha. 2023. Streaming video model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14602–14612.

W Zhong et al. 2024. Enhancing large language models with long-term memory. *AAAI Conference on Artificial Intelligence*.

Xiabin Zhou, Wenbin Wang, Minyan Zeng, Jiaxian Guo, Xuebo Liu, Li Shen, Min Zhang, and Liang Ding. 2024. Dynamickv: Task-aware adaptive kv cache compression for long context llms. *arXiv preprint arXiv:2412.14838*.

Qianchao Zhu, Jiangfei Duan, Chang Chen, Siran Liu, Xiuhong Li, Guanyu Feng, Xin Lv, Xiao Chuanfu, Dahua Lin, and Chao Yang. 2025. Sampleattention: Near-lossless acceleration of long context llm inference with adaptive structured sparse attention. *Proceedings of Machine Learning and Systems*, 7.

Xianwei Zhuang, Yuxin Xie, Yufan Deng, Liming Liang, Jinghan Ru, Yuguo Yin, and Yuexian Zou. 2025. Vargpt: Unified understanding and generation in a visual autoregressive multimodal large language model. *arXiv preprint arXiv:2501.12327*.

A Survey Scope and Positioning

A.1 Motivation and Necessity of This Survey

The motivation for this survey stems from three key observations regarding the current landscape of Large Language Models (LLMs): the paradigm shift to streaming scenarios, ambiguity in "streaming" terminology, and absence of comprehensive reviews in streaming LLMs domain.

The Paradigm Shift to Streaming Scenarios While LLMs have demonstrated remarkable capabilities across various static inputs, real-world deployment increasingly demands streaming interaction. Applications such as digital human assistants, real-time simultaneous interpretation, and embodied robotics require models to process continuous input streams and generate low-latency responses. The transition from "static batch processing" to "dynamic streaming interaction" presents unique challenges in memory management, temporal coherency, and inference efficiency that traditional LLM research overlooks.

Ambiguity in "Streaming" Terminology There is currently a significant semantic ambiguity in the usage of the term "streaming" within the community. It is often conflated across three distinct dimensions: streaming generation (token-by-token output), streaming processing (handling dynamic input context), and streaming interaction (dynamic generate with partial and dynamic input). This survey aims to disambiguate these concepts and provide a rigorous taxonomy.

Absence of Comprehensive Reviews Despite the surge in related research, there is a notable lack of a systematic survey dedicated to Streaming LLMs.

A.2 Focus and Scope Delimitation

To ensure depth and coherence, we delineate the scope of this survey as follows: We primarily focus on decoder-only LLMs, and we structure the survey by tracing the evolution of streaming capabilities: from static input / streaming output (standard generation), to streaming Input / streaming output (infinite context processing), and finally to dynamic interaction (duplex/omni-streaming). We conducted a systematic literature review of top-tier venues in AI, NLP, CV, and speech, **with a cutoff date of December 2025**.

A.3 Comparison with Existing Surveys

While our survey establishes a unified taxonomy for *Streaming LLMs* centered on dynamic data flow and real-time interaction, it is crucial to delineate its scope from other prominent research directions in the LLM landscape. Below, we contrast our focus with three major categories of existing surveys: Efficient LLMs, Multimodal LLMs, and Long-Context LLMs.

Efficient LLMs. The technologies surveyed under the field of Efficient LLMs, including model compression and KV cache management, are foundational technology for efficient, accurate and intelligent Streaming LLMs (Li et al., 2024a; Dantas et al., 2025; Cheng et al., 2025a). However, existing typical surveys in this category, such as the survey on KV cache management for acceleration (Li et al., 2024a), the comprehensive survey on efficient LLMs (Dantas et al., 2025), and the review on compression techniques (Cheng et al., 2025a), predominantly analyze these methods from an offline and static perspective. Central questions of these surveys is mostly on how to reduce the computational or memory footprint of a model that is operating on a complete, existing context to acquire higher throughput or enable development on hardware with constrained resources. In comparison, this survey re-contextualizes these optimizations within a streaming paradigm. Techniques such as dynamic KV cache management and lightweight model adaptation under the overarching imperative of online, real-time interaction are unified. The key challenge shifts from static resource reduction to dynamic runtime budgeting under the strict latency constraints of streaming, where inputs are incrementally available, as is concurrent streaming defined, and outputs must also be generated incremenntally. Thus, while efficient LLM research only casts light on how can we run the model more efficiently, this research also asks how can it read, listen, see and respond efficiently as the world unfolds.

Multimodal LLMs. The field of MLLMs (Zhang et al., 2024a) focuses on augmenting language models with the ability to process and generate content across diverse modalities like vision, audio, and video. Key challenges include cross-modal alignment, fusion strategies, and the design of modality-specific encoders and decoders. Although some MM-LLM applications (e.g., real-time video anal-

ysis or speech-to-speech translation) are inherently streaming, the primary goal of MM-LLM research is to achieve strong performance on multimodal understanding and generation benchmarks. Our survey, however, abstracts away from the specifics of any single modality. We treat the input and output as generic token streams and instead concentrate on the *temporal dynamics of the interaction*. A streaming LLM architecture, as defined in our work, can serve as the backbone for a multimodal system, but the core innovations we survey—such as concurrent perception-generation loops and infinite context processing—are orthogonal to the problem of modality grounding. Our focus is on *how* information flows over time, not *what* the information represents.

Long-Context LLMs. Surveys in this category, such as (Liu et al., 2025b) and (Wang et al., 2024d), primarily focus on expanding the model's static capacity to process extremely long, finite input sequences (e.g., long documents or multi-turn histories). Their core goal is to extend the usable context window and make inference over long sequences efficient, covering key technologies like positional encoding extrapolation, efficient attention architectures (e.g., sparse attention), and sophisticated KV-cache management. While these advances in long-context modeling provide a crucial foundational capability for processing extensive information, their perspective is largely centered on a "read-then-write" inference paradigm for offline, bounded inputs. In stark contrast, our survey on Streaming LLMs investigates the dynamic interaction paradigm required for unbounded, real-time token streams. We focus on the unique challenges of concurrent reading and writing, incremental processing of growing states, and online context/KV budgeting under strict latency constraints. Therefore, while long-context techniques are often essential enabling components, our work shifts the focus from merely enlarging a fixed context window to orchestrating continuous, low-latency reasoning and generation within an ever-flowing data stream.

B Supplementary Literature

Due to space limitations, we defer a broader collection of related work to this appendix. Following the taxonomy in Figure 3, we organize additional literature into three paradigms: (1) Output-streaming LLMs, (2) Sequential-streaming LLMs, and (3) Concurrent-streaming LLMs. This appendix complements the main text by summarizing representative yet less-discussed threads and implementations, rather than aiming for an exhaustive bibliography.

Table 2 presents additional methods for output-streaming LLMs, organized by streaming generation mechanisms and efficiency techniques.

Table 3 summarizes methods for sequential-streaming LLMs, focusing on incremental encoding and streaming context management.

Typical Surveys	Primary Focus	Typical Technologies Covered	Differentiation in This Survey
		Survey Category: Efficient LLMs	
(Li et al., 2024a) (Dantas et al., 2025) (Cheng et al., 2025a) (Wu et al., 2026b)	Compression/adaptation and memory bottlenecks.	1) Compression: quantization, pruning, distillation, low-rank; and 2) KV-cache management: selection / eviction, cache compression, offloading, sliding-window / hierarchical cache.	Prior surveys treat compression and KV-cache optimization as separate threads; we unify them under streaming interaction, highlighting online constraints and dynamic runtime budgeting.
		Survey Category: Multimodal LLMs	
(Zhang et al., 2024a) (Caffagni et al., 2024)	Architectures, training recipes, and benchmarks for MLLMs.	1) Encoder + Projector + LLM, alignment module, tokenizer; and 2) multimodal pretraining & instruction tuning.	Prior MLLM surveys assume fixed inputs and emphasize alignment and benchmarked capabilities. We focus on streaming interaction with token stream abstraction, concurrent IO, incremental perception, and online memory and budget control.
		Survey Category: Long-Context LLMs	
(Wang et al., 2024d) (Liu et al., 2025b)	Long-context modeling: extending usable context windows and making long-sequence inference efficient.	1) Position extrapolation / interpolation; 2) efficient long-sequence attention and architectures; 3) KV-cache management (compression, eviction, and offloading); and 4) workflow-level augmentation (prompt compression, retrieval/external memory).	Prior surveys focus on enlarging a fixed context window for offline inputs or read then write inference. We study streaming token streams with concurrent read and write, incremental inputs, growing states, and online context and KV budgeting for unbounded streams.

Table 1: Comparison between this survey and existing related surveys. We highlight the unique positioning of our work in the context of streaming interaction.

Streaming Generation				
Mechanism			*Modality-Out*	*Methods*
Token	*Block*	*Refinement*		
✓	-	-	T	GPT (OpenAI, 2023), Gemini (Team et al., 2023), Qwen3 (Yang et al., 2025a), DeepSeek-V3 (DeepSeek-AI et al., 2024), InternVL (Chen et al., 2024c), ChatGLM (GLM et al., 2024), Gemma (Gemma Team, 2024)
✓	-	-	S	AudioLM (Borsos et al., 2023a), SpeechGPT (Zhang et al., 2023a), AudioPaLM (Rubenstein et al., 2023), FireRedTTS (Guo et al., 2024a), Moshi (Défossez et al., 2024), Llama-omni2 (Fang et al., 2025), Qwen3-Omni (Xu et al., 2025c), StyLLE (Hao et al.), Llmvox (Shikhar et al., 2025), SpeakStream (Bai et al., 2025)
✓	-	-	V	DALLE (Ramesh et al., 2021), VideoPoet (Kondratyuk et al., 2024), Chameleon (Team, 2024), Emu3 (Wang et al., 2024e), Anole (Chern et al., 2024), Lumina-mGPT2.0 (Xin et al., 2025), Infinity (Han et al., 2025)
-	✓	-	T	SAT (Wang et al., 2018), SoT (Ning et al., 2023), CtrlDiff (Huang and Tang, 2025), PredSent (Hwang et al., 2025), Falcon (Gao et al., 2025), SSD-LM (Han et al., 2023), WeDLM (Liu et al., 2025a), Next-Block (Tian et al., 2025), Block Diffusion (Arriola et al., 2025)
-	✓	-	S	PALLE (Yang et al., 2025f), SyncSpeech (Sheng et al., 2025), DCAR (Li et al., 2025a), StreamFlow (Guo et al., 2025), TtT (Liu et al., 2025e), DiTAR (Jia et al., 2025)
-	✓	-	V	show-o (Xie et al., 2024), XTRA (Amrani et al., 2025), NTP (Ren et al., 2025), CausVid (Yin et al., 2025), BlockVid (Zhang et al., 2025f), NBP (Ren et al., 2025)
-	-	✓	T	Mask-Predict (Ghazvininejad et al., 2019), LevT (Gu et al., 2019), Insertion-Deletion (Ruis et al., 2020), Diffusion-LM (Li et al., 2022), DiffuSeq (Gong et al., 2022), D3PM (Austin et al., 2021)
-	-	✓	S	SoundStorm (Borsos et al., 2023b), Voicebox (Le et al., 2023), Specmaskgit (Comunità et al., 2024), IMPACT (Huang et al., 2025a), Maskgct (Wang et al., 2024f), DDM-TASTE (Ku et al., 2025)
-	-	✓	V	MaskGIT (Chang et al., 2022), Muse (Chang et al., 2023), DiT (Peebles and Xie, 2022), VAR (Tian et al., 2024), DetailFlow (Liu et al., 2025h), DC-AR (Wu et al., 2025)

Streaming Efficiency			
Efficient		*Modality-Out*	*Methods*
Decode	*Memory*		
✓	-	T	Speculative Sampling (Chen et al., 2023), Medusa (Cai et al., 2024a), EAGLE2 (Li et al., 2024d), BiLd (Kim et al., 2023), CTC-based Drafting (Wen et al., 2024), FLY (Liu et al., 2025c), SkipDecode (Del Corro et al., 2023), SkipGPT (Zhao et al., 2025a), EESD (Liu et al., 2024a), HiDrop (Wu et al., 2026a), Visipruner (Fan et al., 2025)
✓	-	S	LiveSpeech (Dang et al., 2024), MTP-SpecDec (Nguyen et al., 2025), SSD (Lin et al., 2025), VocalNet (Wang et al., 2025d), VADUSA (Li et al., 2025b), PCG (Yanuka et al., 2025)
✓	-	V	SJD (Teng et al., 2024), CSpD (Wang et al., 2024g), GSD (So et al., 2025), VVS (Dong et al., 2025), FreqExit (Li et al., 2025j), SkipVAR (Li et al., 2025d), PAR (Wang et al., 2025e), ADT-Tree (Lei et al., 2025), Lantern (Jang et al., 2024)
-	✓	T	StreamingLLM (Xiao et al., 2023), H2O (Zhang et al., 2023b), Scissorhands (Liu et al., 2023), Snapkv (Li et al., 2024c), Dynamickv (Zhou et al., 2024), Chunkkv (Liu et al., 2025g)
-	✓	S	wu2024ts3 (Wu et al., 2024a), LST (Lu et al., 2025), SpeechTokenPrediction (Liu et al., 2025f)
-	✓	V	HACK (Qin et al., 2025c), ScaleKV (Li et al., 2025e), AMS-KV (Xu et al., 2025a), LineAR (Qin et al., 2025b)

Table 2: Summary of additional literature on output-streaming LLMs, complementing the discussion in Sec. 3.

Incremental Encoding			
Type		*Modality-In*	*Methods*
Fragmented Encoding	*Atomic Encoding*		
✓	-	T	SimulMT (Wang et al., 2024c), Moshi (Défossez et al., 2024), Codec (Ye et al., 2025), dmel (Bai et al., 2024), Lightweight Audio Segmentation (Frohmann et al., 2024), Semantic VAD (Shi et al., 2023)
✓	-	S	Whisper-Streaming (Macháček et al., 2023), SimulST (Zhang and Feng, 2023), CTC (Graves, 2012), Speechtokenizer (Xin et al., 2024), Moshi (Défossez et al., 2024), Codec (Ye et al., 2025), dmel (Bai et al., 2024), Lightweight Audio Segmentation (Frohmann et al., 2024), Semantic VAD (Shi et al., 2023)
✓	-	V	S-ViT (Zhao et al., 2023)
-	✓	T	SaT (Frohmann et al., 2024), SegFree (Iranzo-Sánchez et al., 2024), WtP (Minixhofer et al., 2023), subword regularization (Kudo, 2018), SentencePiece (Kudo and Richardson, 2018),
-	✓	V	ViT (Dosovitskiy, 2020), CLIP (Radford et al., 2021)
Streaming Context Management			
Type			*Methods*
Mem.	*KV*	*Attn.*	
✓	-	-	StreamingTOM (Chen et al., 2025b), MemoryBank (Zhong et al., 2024), LongMem (Wang et al., 2023), VideoStreaming (Qian et al., 2024), Timechat-online (Yao et al., 2025), Prunevid (Huang et al., 2025b), DyCoke (Tao et al., 2025), ProVideLLM (Chatterjee et al., 2025), VideoLLaMB (Wang et al., 2025f), STREAMMIND (Ding et al., 2025b), VideoStreaming (Qian et al., 2024), StreamingAssistant (Jin et al., 2025), Focus (Wei et al., 2025a), StreamForest (Zeng et al., 2025), Flash-vstream (Zhang et al., 2024b)
-	✓	-	H2o (Zhang et al., 2023b), PyramidKV (Cai et al., 2024b), SnapKV (Li et al., 2024c), StreamKV (Chen et al., 2025c), STC (Wang et al., 2025b), Streammem (Yang et al., 2025e), AViLA (Zhang et al., 2025a), StreamingVLM (Xu et al., 2025d), PyramidInfer (Yang et al., 2024a), DynamicKV (Zhou et al., 2024), PrefixKV (Wang et al., 2024a), CAKE (Qin et al., 2025a), SimLayerKV (Zhang et al., 2024d), AdaKV (Feng et al., 2024), CriticalKV (Feng et al., 2025), LeanKV (Zhang et al., 2024e), RazorAttention (Tang et al., 2024), HeadKV (Fu et al., 2024b), DuoAttention (Xiao et al., 2024a)
-	-	✓	Attention Sink (Xiao et al., 2023), Sirllm (Yao et al., 2024), GLA (Yang et al., 2023), DeltaNet (Yang et al., 2024b), Lightning attention-2 (Qin et al., 2024), SAMPLEATTENTION (Zhu et al., 2025), Lserve (Yang et al., 2025d), DCA (An et al., 2024)

Table 3: Summary of additional literature on sequential-streaming LLMs, complementing the discussion in Sec. 4.

Streaming Paradigm						
Paradigm				*Modality*		*Methods*
R.	*C.*	*I.*	*G.*	*In*	*Out*	
✓	-	-	-	T	T	Simul-LLM (Agostinelli et al., 2024), SiLLM (Guo et al., 2024c), TransLLaMA (Koshkin et al., 2024b), CAST (Koshkin et al., 2024a), RALCP (Wang et al., 2024c)
✓	-	-	-	S	T	CAST (Koshkin et al., 2024a), TransLLaMA (Koshkin et al., 2024b)
-	✓	-	-	T	S	LLMVoX (Shikhar et al., 2025), Mini-Omni (Xie and Wu, 2024)
-	✓	-	-	S	S	Mini-Omni (Xie and Wu, 2024)
-	✓	-	-	V	T	ViSpeak (Fu et al., 2025c)
-	-	✓	-	T	T	EAST (Fu et al., 2025a), Shanks (Chiang et al., 2025a)
-	-	✓	-	T	S	STITCH (Chiang et al., 2025b)
-	-	✓	-	S	T	EASiST (Fu et al., 2025b), InfiniSST (Ouyang et al., 2025), SASST (Yang et al., 2025g), StreamingASR (Wan et al., 2026)
-	-	✓	-	S	S	SALMONN-omni (Yu et al., 2024)
-	-	✓	-	V	T	Videollm-online (Chen et al., 2024a), LiveCC (Chen et al., 2025a), ProVideLLM (Chatterjee et al., 2025), StreamBridge (Wang et al., 2025a), LiveStar (Yang et al., 2025i), SVBench (Yang et al., 2025h), ProASIST (Zhang et al., 2025c)
-	-	-	✓	T	T	StreamingGPE (Tong et al., 2025b), StreamingThinker (Tong et al., 2025a), DST (Guo et al., 2024b)
-	-	-	✓	S	T	StreamingGPE (Tong et al., 2025b)
-	-	-	✓	V	T	StreamChat (Liu et al., 2024b), Speak-While-Watching (Lin et al., 2026), TaYS (Zhang et al., 2026)
Interaction Policy						
Policy				*Modality*		*Methods*
Rule	*SFT*	*RL*		*In*	*Out*	
✓	-	-		T	T	Simul-LLM (Agostinelli et al., 2024; Raffel et al., 2024), StreamingGPE (Tong et al., 2025b), STACL (Ma et al., 2019), AsyncReasoning (Yakushev et al., 2025), StreamingThinker (Tong et al., 2025a), Conveyor (Xu et al., 2024), AsyncLM (Gim et al., 2024)
✓	-	-		T	S	CosyVoice 2 (Du et al., 2024), IST-LM (Yang et al., 2024c), DSM (Zeghidour et al., 2025)
✓	-	-		S	T	MFLA (Xia et al., 2025b), InfiniSST (Ouyang et al., 2025), LLM as Processor (Tong et al., 2025b), SASST (Yang et al., 2025g), SimulS2S-LLM (Deng et al., 2025), ReaLLM (Seide et al., 2024), Llama-omni (Fang et al., 2024)
✓	-	-		S	S	StreamRAG (Arora et al., 2025)
✓	-	-		V	T	LiveCC (Chen et al., 2025a), StreamVLN (Wei et al., 2025b), ActiveVLN (Zhang et al., 2025e), AViLA (Zhang et al., 2025a)
-	✓	-		T	T	SiLLM (Guo et al., 2024c), TransLLaMa (Koshkin et al., 2024b), EAST (Fu et al., 2025a), DrFrattn (Zhao et al., 2025b), FineHarm (Li et al., 2025i), PsFuture (Zhao et al., 2024)
-	✓	-		T	S	SimulMEGA (Le et al., 2025), Cosyvoice (Du et al., 2024), DSM (Zeghidour et al., 2025)
-	✓	-		S	T	Divergence (Chen et al., 2024b), SimulMEGA (Le et al., 2025), ReaLLM (Seide et al., 2024), Llama-omni (Fang et al., 2024)
-	✓	-		S	S	StreamSpeech (Zhang et al., 2024c), EASiST (Fu et al., 2025b), SimulMEGA (Le et al., 2025)
-	✓	-		V	T	Videollm-online (Chen et al., 2024a), ProVideLLM (Chatterjee et al., 2025), EyesWO (Zhang et al., 2025d), Streamo (Xia et al., 2025a), ProASIST (Zhang et al., 2025c), Videollm-MOD (Wu et al., 2024b), DisPider (Qian et al., 2025), Stream-VLM (Panchal et al., 2024), Lion-FS (Li et al., 2025h), ProVideLLM (Chatterjee et al., 2025), StreamBridge (Wang et al., 2025a)
-	-	✓		T	T	SeqPO-SiMT (Xu et al., 2025e), Interleaved Reasoning (Xie et al., 2025)
-	-	✓		T	S	Seed LiveInterpret 2.0 (Cheng et al., 2025b)
-	-	✓		S	T	Seed LiveInterpret 2.0 (Cheng et al., 2025b)
-	-	✓		S	S	Seed LiveInterpret 2.0 (Cheng et al., 2025b)
-	-	✓		V	T	MMDuet2 (Wang et al., 2025c)

Table 4: Summary of additional literature on concurrent-streaming LLMs, complementing the discussion in Sec. 5.