Token Reduction Should Go Beyond Efficiency in Generative Models – From Vision, Language to Multimodality

Zhenglun Kong[1]*, Yize Li[2]*, Fanhu Zeng[3], Lei Xin[4,6], Shvat Messica[1], Xue Lin[2], Pu Zhao[2], Manolis Kellis[5], Hao Tang[6], Marinka Zitnik[1]
[1]Harvard University, [2]Northeastern University, [3]Chinese Academy of Sciences,
[4]Wuhan University, [5]Massachusetts Institute of Technology, [6]Peking University,
{zhenglun_kong,marinka}@hms.harvard.edu, li.yize@northeastern.edu

Abstract

In Transformer architectures, tokens—discrete units derived from raw data—are formed by segmenting inputs into fixed-length chunks. Each token is then mapped to an embedding, enabling parallel attention computations while preserving the input's essential information. Due to the quadratic computational complexity of transformer self-attention mechanisms, token reduction has primarily been used as an efficiency strategy. This is especially true in single vision and language domains, where it helps balance computational costs, memory usage, and inference latency. Despite these advances, this paper argues that token reduction should transcend its traditional efficiency-oriented role in the era of large generative models. Instead, we position it as a fundamental principle in generative modeling, critically influencing both model architecture and broader applications. Specifically, we contend that across vision, language, and multimodal systems, token reduction can: (i) facilitate deeper multimodal integration and alignment, (ii) mitigate "overthinking" and hallucinations, (iii) maintain coherence over long inputs, and (iv) enhance training stability, etc. We reframe token reduction as more than an efficiency measure. By doing so, we outline promising future directions, including algorithm design, reinforcement learning-guided token reduction, token optimization for in-context learning, and broader ML and scientific domains. We highlight its potential to drive new model architectures and learning strategies that improve robustness, increase interpretability, and better align with the objectives of generative modeling. [2]

1 Introduction

Transformer-based generative models [10, 19, 31, 87] have emerged as dominant deep learning architectures across vision, language, and multimodal tasks, due to their ability to process long sequences of tokens, which are the fundamental representational units derived from raw data such as subwords in language or image patches in vision. As these models are applied to increasingly complex real-world tasks, the input sequence lengths of both the models and their training datasets continue to grow. However, the quadratic computational complexity of the attention mechanism results in high memory usage and slow inference, which hinders the practical deployment of generative models at scale. Token reduction addresses this challenge by reducing the number of tokens processed during inference. By pruning or merging tokens, token reduction decreases computational cost and accelerates runtime, offering a practical solution for improving the efficiency of generative models.

*Equal contribution.
[2]We collected a list of token reduction papers at Github.

Token reduction has been widely adopted in computer vision, language processing, and multimodal tasks. In vision transformers, it has primarily been used to reduce computational cost by removing visually redundant tokens [8, 48, 57, 71]. While effective in decreasing token-level computations and accelerating inference, this approach can discard subtle but important visual features, especially in dense prediction tasks such as segmentation and object detection, leading to significant performance degradation. In language models, token reduction has commonly been implemented through early-exit mechanisms and token-skipping strategies [59, 96], which reduce the number of intermediate tokens processed and thus lower computational overhead. Similarly, multimodal large language models (MLLMs) apply visual token pruning primarily during the prefill stage [17], where adaptive attention patterns are learned in the early layers to prune tokens in later stages. Despite progress in existing work, token reduction is still predominantly viewed as a means of improving computational efficiency, primarily by reducing the number of tokens to minimize associated computations and accelerate inference. Such an efficiency-only mindset has critical limitations. Naive pruning methods may discard informative tokens, thereby degrading model understanding and performance [57, 110, 111]. Furthermore, token reduction is commonly treated as a post hoc optimization, rather than being integrated into the core design and training of the model [17].

In this paper, **we argue that viewing token reduction purely from an efficiency perspective is fundamentally limited.** Instead, **we position token reduction as a core design principle in generative modeling**, deeply integrated with both training and inference to prioritize tokens that maximize downstream task performance and semantic integrity.

Modern generative tasks present numerous challenges that highlight the need for thoughtful token selection: (i) Ultra-long contexts in language modeling require selective retention of relevant segments to preserve coherence. (ii) LLMs frequently exhibit overthinking, repeatedly attending to low-value tokens and producing redundant or contradictory outputs. (iii) Multimodal generation tasks often face issues of visual redundancy, where background tokens overshadow salient visual features critical for accurate understanding. (iv) Noisy or irrelevant tokens introduced during training slow down convergence and harm model stability. By learning to intelligently select, merge, or compress tokens based on their contribution to generation objectives-rather than solely on raw redundancy-models can simultaneously reduce computational load, improve robustness, and enhance interpretability and alignment. This paper makes the following three key contributions:

- We provide a comprehensive analysis of the limitations inherent in current efficiency-centric token pruning methods across vision, language, and multimodal generative models, highlighting critical shortcomings in these widely adopted strategies.
- We identify core challenges faced by modern generative models including insufficient visual representation, semantic misalignment, overthinking in reasoning, and training instability. We then demonstrate how principled token reduction strategies can effectively mitigate these issues.
- We propose a roadmap for future research on token reduction, including directions for method design, reinforcement learning-guided token selection, adaptive in-context compression, and hardware-algorithm co-design, etc. These directions aim to support the development of next-generation generative architectures that are both robust and efficient.

This position paper is organized as follows: Sec. 2 reviews prior token reduction methods across various modalities. Sec. 3 introduces the problem formulation, Sec. 4 formalizes the identified challenges and demonstrates how informed token reduction strategies can address them. Sec. 5 proposes promising research directions for advancing token reduction as well as broader implications.

2 Related Work

2.1 Token Reduction in Vision Models

Image Classification. Classification serves as a fundamental task for vision models and token reduction techniques have been widely applied in it due to its simplicity and versatility. It has been widely explored from various aspects [8, 57, 71, 99, 108]. Specifically, DynamicViT [71] devises a lightweight module to predict the importance score of each token, thereby pruning unimportant tokens. E-ViT [57] identifies attentive tokens from feed-forward network, enabling token pruning without additional parameters. ToMe [8] merges tokens with similarity based on bipartite matching

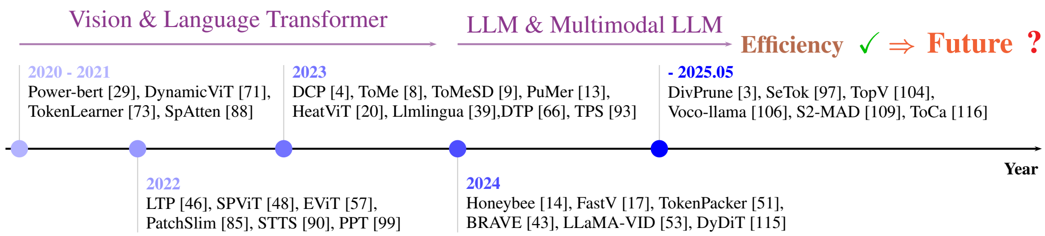

Figure 1: Timeline of notable method developments for token reduction methods with modality shifts (Vision & Language $\rightarrow$ Multimodal LLMs). All these strategies aim to speed up inference with neglected performance drops. Conversely, we ask: **What is the next token reduction paradigm in generative model design that goes beyond test-time accelerations?**

to maintain information utility. PPT [99] analyzes the statistic data between layers and adaptively employs token pruning and merging within layers to achieve higher acceleration performance.

Video Compression. Unlike token reduction in image classification, video compression focuses more on the temporal redundancy within videos, and algorithms are developed to reduce the number of tokens with less computational overhead. Various token reduction methods have been investigated for different tasks including video understanding [73, 90], video editing [52], video-text retrieval [62, 76], video action detection [16] and so on. Specifically, STTS [90] introduces a lightweight framework that dynamically selects the most informative spatial-temporal tokens in video transformers. Tokenlearner [73] proposes an adaptive tokenization module that learns a handful of informative spatial-temporal tokens, significantly reducing computational costs. EVAD [16] selectively drops irrelevant spatial-temporal tokens in non-keyframes while preserving keyframe and motion-relevant tokens, and then refines actor features using a context-aware decoder to maintain accuracy with reduced computations.

Generative Tasks. Token reduction in generative tasks [41] aims to accelerate generative models through the efficient utilization of tokens. It can be applied to both diffusion models [9, 89] and diffusion transformers [116]. Specifically, ToMeSD [9] exploits natural redundancy in generated images by merging redundant tokens, successfully extending token merging to stable diffusion with simple unmerging. DyDiT [115] reduces redundancy with a Timestep-wise Dynamic Width approach to adopt model width conditioned on the generation timesteps, and a Spatial-wise Dynamic Token strategy to avoid redundant computations at unnecessary spatial locations.

2.2 Token Reduction in Language Models

Token reduction strategies in language modeling have evolved from early optimizations for BERT [37, 44, 46, 47, 105] to techniques specifically designed for LLMs. PoWER-BERT [29] introduces progressive word-vector elimination by removing redundant token representations based on self-attention dynamics, improving inference efficiency. Learned token pruning [46] extends this approach by learning attention-based thresholds to adaptively prune uninformative tokens, thereby reducing computational costs while preserving model performance. In LLMs, token reduction must account for the constraints of autoregressive decoding across diverse downstream tasks. Dynamic pooling methods [4, 86] adjust token representations on the fly during inference to reduce redundancy. Prompt compression techniques [24, 27, 39] aim to reduce computational overhead by compressing the input prompt before generation. Selective decoding approaches [24, 95] reduce per-step inference costs by computing key-value pairs only for tokens critical to predicting the next token. In multi-agent systems, S^2-MAD [109] proposes a sparsification mechanism that limits unnecessary token exchanges between agents, reducing communication costs and improving the efficiency of collaborative reasoning.

2.3 Token Reduction in Multimodal LLMs

Recent work has explored visual token pruning by addressing attention inefficiencies of deep transformer layers in MLLMs [3, 6, 59, 74]. Specifically, FastV [17] shows that deeper vision-language layers expend significant computations on redundant image tokens. To address this, a lightweight module is adopted to adaptively prune these tokens, reducing inference overheads in subsequent stages. A complementary approach modifies vision feature extractors or projectors to output a smaller set of highly informative image tokens, effectively distilling the input into a compressed representation [11, 14, 43, 51, 54, 106]. However, efficiency gains from these prefill stages often fade during the decoding phase, where per-token computations dominate. To overcome this, recent methods jointly optimize token reduction during both prefill and decoding stages, ensuring sustained speedups throughout inference [35, 78]. In Fig. 1, we present a timeline of notable developments in token reduction methods, illustrating the shift from early applications in ViT and BERT-based models to more recent advances in LLMs and MLLMs.

3 Problem Formulation

In modern generative models [7, 30, 60, 67, 70], a token denotes one fundamental unit of input or representation, typically encoded as a vector. For example, a token might correspond to a subword in language, a patch in an image, or an embedding of a time step in audio. We denote a sequence of N input tokens as $X = [x_1, \ldots, x_N] \in \mathbb{R}^d$. Token reduction refers to any operation that compresses the token sequence to M tokens (with $M < N$) by removing or consolidating tokens while aiming to preserve the original information.

Broadly, token reduction methods fall into the following categories: 1) Token pruning methods [8] that remove entire unimportant tokens, simply dropping them from the sequence; 2) Token merging methods [8, 9] which fuse information from multiple tokens into fewer tokens, effectively compressing the sequence by merging similar or related tokens; 3) Hybrid strategies [13, 45, 99] that combine pruning and merging within a unified framework; 4) Token distillation approaches [11, 65] which integrate rich information across longer input sequences or multiple modalities into fewer condensed tokens, enabling efficient cross-modal interactions and long-context reasoning in LLMs and MLLMs.

A core challenge in token reduction is the determination of tokens to be pruned or merged. There are various importance criteria and scoring mechanisms to rank token significance, including attention-based heuristics [57], gradient or loss-based criteria [35], clustering [32], and learned predictors [71].

From a purely efficiency-oriented perspective, token reduction delivers substantial computational efficiency gains by reducing the quadratic computation cost from $\mathcal{O}(N^2)$ to $\mathcal{O}(M^2)$ in attention mechanisms. By eliminating redundant tokens and processing fewer computations during inference, it effectively accelerates the inference speed and improves the model throughput, which is crucial for latency-sensitive tasks or real-time applications. Furthermore, it also shrinks the memory footprint for activations and gradients (e.g., key/value caches), alleviating memory usage for both inference and training, which is particularly beneficial for wide deployments on resource-limited platforms.

However, as stated in this position paper, token reduction can benefit models in multiple ways beyond efficiency, which will be introduced in detail in the following sections.

4 Challenges and Positions

In this section, we establish token reduction as a foundational mechanism for addressing critical challenges in modern generative systems. We claim five core challenges across modalities: visual representation sparsity, semantic misalignment, reasoning redundancy, training instability, and long-context overload. We demonstrate how principled token reduction strategies intrinsically address these issues through dynamic token-semantic co-optimization. **We position token reduction not only as an efficiency tool, but as an essential paradigm for enhancing semantic coherence and enabling sustainable scaling of generative systems.**

4.1 Obtain Informative Visual Representation

MLLMs often suffer from noisy visual inputs that impede fine-grained understanding. We outline key challenges in MLLM visual reasoning: ① *Text-Visual Attention Shift:* Due to the rotary positional embeddings in LLM decoders, later text tokens disproportionately attend to spatially lower image regions [33], shifting attention away from semantically important areas (e.g., objects at the top of an image); ② *Visual Redundancy:* Empirical studies [59, 96] show that beyond the first few layers, many image tokens contribute little new information, ③ *Task-Guided Focus in VQA:* In multimodal question answering, the question itself pinpoints relevant image regions (e.g., "kitten color" directs focus to the kitten patch), implying that many image tokens are unnecessary for correct answers [78].

Therefore, we position token reduction as a representation-learning optimization: selecting the subset of tokens that preserves informative visual representation. For example, VisPruner [113] identifies high-value tokens using visual-encoder attention and removes duplicates via clustering to ensure diversity. VTW [59] observes that visual information migrates into text tokens within early layers; it therefore withdraws all visual tokens after a chosen layer based on KL-divergence criteria. TRIM [78] leverages the CLIP metric and IQR scoring function to adaptively select image tokens that are crucial for answering questions, while an aggregated token is used to retain additional image information.

4.2 Better Multimodal Token Alignment

Despite their impressive capabilities, MLLMs continue to face challenges in semantic alignment. Standard vision tokenizers typically split images into fixed-size patches, which can fragment coherent visual entities (e.g., objects or regions) across multiple tokens. This fragmentation weakens the alignment between visual and linguistic representations. Token reduction offers a promising solution by selecting visual tokens based on semantic importance, thereby producing a compact set of tokens that better align with language representations. Specifically, SeTok [97] dynamically clusters visual features into semantically meaningful tokens using a density-peak algorithm, which determines both the number and structure of token groupings per image. This approach preserves both high- and low-frequency semantics, substantially improving concept-level alignment and downstream task performance. M3 [11] introduces a hierarchical token structure that captures coarse-to-fine semantic granularity, allowing different levels of abstraction to be selectively retained depending on task needs.

4.3 Reduce Overthinking in Reasoning

LLM reasoning. In the context of language models, overthinking refers to generating excessively long or convoluted chains of reasoning that go beyond what is necessary to reach a correct answer. An LLM may produce verbose, repetitive, or even self-contradictory explanations when it fails to converge on a solution-often due to uncertainty [82, 91]. Such extended reasoning trajectories are inefficient and recent studies show that state-of-the-art reasoners can consume over 15,000 tokens to solve math problems that could be addressed with a concise chain-of-thought (CoT) of just a few hundred tokens [34]. Mitigating overthinking is thus crucial for improving both accuracy and efficiency in generative modeling. By reducing unnecessary tokens during reasoning, LLMs can focus on salient steps, aligning generation with a more concise and logical trajectory.

CoT-Influx [36] introduces a CoT pruning strategy in which concise reasoning examples are included in the prompt. By pruning unimportant tokens from these examples, more reasoning demonstrations can fit into the context window, surprisingly leading to improved math reasoning accuracy. TokenSkip [100] enables LLMs to skip less important tokens within CoT sequences and learn shortcuts between critical reasoning steps. This allows for controllable CoT compression with adjustable compression ratios, enabling models to automatically trim redundant tokens during reasoning.

MLLM reasoning. MLLMs, which reason over text and other modalities, face similar overthinking issues. In vision-language tasks, overthinking often manifests as excessive processing of visual tokens or overly detailed image descriptions, resulting in inefficiency and potential confusion [15]. Token reduction techniques in MLLMs aim to promote more focused and sparse reasoning over multimodal inputs. For example, FAST [101] rewards shorter-than-average token sequences for correct answers, while allowing longer reasoning for more complex tasks. It also adjusts policy optimization constraints to tighten output exploration for simple tasks (thus reducing unnecessary tokens) and loosen it for harder ones to allow deeper reasoning.

Together, these strategies reduce overthinking in straightforward cases, boosting efficiency while preserving effective reasoning depth for complex scenarios.

4.4 Improve Training Stability

While token reduction has traditionally been employed as a post-training optimization to enhance inference efficiency, recent research indicates its potential to significantly improve training stability when integrated into the pre-training phase [25, 55, 58], suggesting that selective token utilization during training can lead to more robust and effective model learning.

One notable approach is Rho-1 [58], which involves scoring tokens based on their alignment with a desired distribution using a reference model and then focusing the training loss on tokens with higher scores. Therefore, it effectively filters out noisy or less informative tokens, leading to faster convergence and improved performance. UPFT [38] emphasizes the importance of initial reasoning steps in training. By reducing the number of training tokens, UPFT encourages the model to focus on the initial prefix substrings of reasoning trajectories, which are often more stable and contain crucial information. This focus helps the model avoid being influenced by subsequent complex or potentially erroneous information, thereby improving training stability. However, how to design token reduction algorithms that are tightly integrated with training procedures such as in modern methods like GRPO [75], remains underexplored. Future research should investigate specialized approaches that incorporate token reduction directly into training objectives, enabling models to learn to prioritize or discard tokens in a task-aware and gradient-aligned manner.

4.5 Enhance Long Context & Video Understanding

Long-context LLMs. Long-context language modeling presents unique challenges: ① Long texts often contain raw tokens that exhibit repetitive descriptions and irrelevant details that strain the attention mechanism; ② LLM-based agent systems use input data as sequential prompts for reasoning or for switching between multiple tasks, which can lead to overload when the prompt grows too large; ③ It is very difficult to scale up to even longer content for learning more information. Token reduction techniques directly address these issues by distilling extensive input sequences into compact summary vectors or representative tokens. By doing so, models preserve core information such as key events, central themes, or task-specific facts, while significantly decreasing cognitive load. For example, AutoCompressors [18] trains pre-trained LLMs to compress long contexts into compact summary tokens, reducing token length by orders of magnitude to extend context windows and speed up inference. TokenSwift [98] reduces the effective number of tokens that the model dynamically processes during generation by using multi-token parallel generation and n-gram retrieval for token reutilization, therefore enabling efficient ultra-long sequence generation (up to 100K tokens).

Video-based MLLMs. The necessity of token reduction primarily lies in enhancing the model's effective understanding of video content through: ① *Instruction-guided information filtering*: token reduction prioritizes selecting visual information relevant to user instructions over raw data volume. ② *preserving spatiotemporal structure*: token reduction strategically compresses massive spatiotemporal information to retain spatiotemporal dependencies, ensuring the model can capture dynamic semantics, as well as prevent redundant tokens interfere with long temporal reasoning. ③ *Preserving semantic integrity*: it facilitates feasible processing of extremely long sequences in learning while preserving semantic integrity. ④ *Multi-modal alignment*: token reduction distills visual information into a compact, semantically aligned form, thereby efficiently bridging the gap between language and vision [63]. By doing so, it effectively addresses the challenges posed by the low abstractness and lack of guidance inherent in raw visual inputs, which are the root causes of semantic misalignment and optimization ambiguity in multi-modal models. Recent works illustrate these principles: HICom [63] conducts conditional token compression at local and global levels using user instructions as guidance to retain instruction-relevant visual information while reducing computational burden. Video-XL-Pro [61] employs reconstructive token compression with a dynamic token synthesizer and semantic-guided masking to generate compact yet comprehensive video tokens for improved MLLM performance and efficiency.

5 Future Directions

In this section, we propose eight promising directions for token reduction beyond the efficiency benefits, organized into three categories: (i) Algorithmic Innovations (Sec. 5.1∼5.4), (ii) Application Innovations (Sec. 5.5∼ 5.7), and (iii) Hardware-Algorithm Co-Design (Sec. 5.8).

5.1 Design of New Algorithms

Future research on algorithm design should explore holistic and adaptive token reduction strategies. Building on recent advances, we outline six promising directions:

Better Token Importance Metrics. It is critical to re-evaluate how token importance is defined and measured. More robust and unbiased scoring mechanisms can be developed, such as predictors [2] or meta-learning frameworks that go beyond attention-based proxies. These models should capture downstream utility with minimal supervision, enabling adaptive pruning across tasks and domains.

Constructive Token Compression. Token reduction can shift from purely eliminative pruning to strategies that merge spatially or semantically similar tokens into compact summary vectors [51].

Mitigating Position Bias. In MLLMs, attention-based pruning methods (e.g., FastV) often rely on attention scores from a fixed query token, leading to retained tokens concentrating in specific image regions (e.g., lower corner) [94] with potential position bias. Future methods should preserve spatial diversity by enforcing structural uniformity in retained tokens to improve robustness on visual tasks.

Cross-Modal Guided Pruning. Pruning decisions in MLLMs should be guided by inter-modality dependencies, rather than made independently for each modality. For example, text-guided pruning of visual tokens can improve alignment between modalities [12]. The design should account for joint representations and semantic correspondence across all relevant inputs.

End-to-End Sparsification. Token reduction should consider both the prefill stage and decoding phase for LLMs. This includes dynamically managing the sparsity of KV caches and selectively updating generated tokens, sustaining efficiency gains throughout the entire inference process [35].

Hardware-Algorithm Co-Design. Token pruning can explore custom hardware and compiler optimizations that take advantage of dynamic token sparsity patterns (e.g., irregular memory access and conditional computation) to maximize throughput and energy efficiency as detailed in Sec. 5.8.

5.2 Token Reduction for Reinforcement Learning

Reinforcement learning (RL)-driven token reduction technology demonstrates significant potential in multimodal large models, with its core lying in balancing computational efficiency and reasoning accuracy through dynamic reward mechanisms and sparsity constraints. Under the "Fast-Slow Thinking" framework [101], RL can guide hierarchical screening of high-value tokens: The fast path employs sparsity strategies (e.g., rule-based rewards or information density scoring) to filter redundant visual or semantic features, while the slow path focuses on refined reasoning. This hierarchical mechanism not only reduces computational overhead by up to 30-50%, but also continuously optimizes token selection strategies via RL's online feedback. For instance, integrating self-play mechanisms with the GRPO algorithm (Group Relative Policy Optimization) enables the generation of diverse candidate paths and selects optimal outputs.

Furthermore, the sparsity reinforcement paradigm proposed in ZipR1 [15] highlights RL's unique advantage in cross-modal generalization. By designing sparsity rewards (e.g., token quantity limits or step compression incentives), models can autonomously generate concise intermediate reasoning chains and even distill such capabilities into lightweight models. For example, in visual question answering tasks, models can dynamically extract key image region features while enhancing answer reliability through multi-candidate contrast mechanisms (e.g., parallel samplings), achieving the goal of "low computational cost, high task accuracy." Looking ahead, this technology could advance toward adaptive sparsity and cross-modal alignment, providing efficient inference foundations for edge computing and real-time interaction scenarios. Simultaneously, it will drive MLLMs toward co-evolution of lightweight design and strong robustness.

5.3 From Prompt Tuning to In-context-learning

Current token reduction efforts for prompts have primarily aimed at compressing prompts for efficiency, often with impressive results [39, 65]. Looking forward, token reduction should evolve into enhancing reasoning and maximizing utility per token in context. Instead of focusing solely on making prompts shorter, future research should explore how each remaining token can carry more information or trigger more complex inference during in-context learning. One direction is to alter the generation paradigm itself, for example, training language models to predict multiple tokens per step [28]. Another idea is to enable deeper internal reasoning without increasing prompt length. CoD [1] introduces an iterative approach using token reduction within fixed-length constraints: Each step incorporates new salient entities into the summary by compressing/abstracting existing content, forcing the model to enhance density without expanding length.

In summary, the next phase of token reduction research should shift focus from simple prompt compression to reasoning-centric compression. Rather than just trimming prompts, we should ask: *How can we make each token in the prompt or context do more work for us?* This involves training models with objectives that reward higher-level inference per token, developing architectures that recycle tokens for multi-step thinking, or dynamically selecting the most salient tokens to keep at each stage of dialogue.

5.4 Complementary to Other Methods

Token reduction can complement other efficiency techniques, such as quantization. By selectively reducing the number of tokens processed during inference, models can improve both performance and efficiency, particularly when paired with quantization strategies [50]. Traditional key-value cache quantization methods often suffer from accuracy loss due to their inability to handle outlier tokens that carry distinct or rare features. To mitigate this issue, Outlier Token Tracking [81] identifies outlier tokens during decoding and excludes them from quantization, preserving full-precision necessary representations and improving key-value cache quantization accuracy. Similarly, Agile-Quant [77] incorporates token pruning as a preprocessing step to reduce the impact of activation outliers. It prunes tokens based on their attention to the start-of-sequence token, discarding those with low attentiveness, which often appear in adjacent channels and contribute to quantization noise. This targeted pruning reduces interaction distances between salient tokens and helps maintain model accuracy under low-bit quantization settings.

5.5 Towards Dense Prediction Tasks for Vision

Existing works primarily concentrate on compressing the backbone of models to ensure their generalization ability, and few works explore recovering all tokens for dense prediction tasks [9, 56]. It is necessary to develop custom token reduction methods for various downstream dense prediction applications like autonomous driving and robotic control with specific settings and requirements. Lacking these specialized designs would lead to a mismatch and performance drop when deployed in real-world settings. For example, autonomous driving [112] would require displacement and velocity based on occupancy prediction, and robotic control [92] would demand rotation angle according to the grid map. Therefore, how to develop fast and specialized token reduction strategies tailored for downstream dense prediction tasks is crucial for deployment in practical scenarios.

5.6 Towards Long Video Applications

Exploiting long videos holds great potential, as processing hours of footage is significantly more labor-intensive and time-consuming than working with short clips. Due to the inherent complexity and resource demands, most current research on long video learning focuses on discriminative tasks such as video understanding [49, 72]. In contrast, broader applications including long video editing [114], long video-text retrieval, and narrative-level generation remain largely underexplored. Progress in these areas could have a significant impact on scene editing in video clips, character rendering in movies, and retrieving useful information from numerous videos. Moreover, token reduction offers a path toward interpretability and efficiency in long video processing [40]. This mimics the human visual system, which does not attend to every frame in detail but instead focuses on salient spatiotemporal changes, such as actions or object movement, while filtering out static, redundant content like backgrounds and stationary objects. Future models should similarly prioritize informative frames and temporal segments, allowing them to reason over extended video sequences with greater efficiency and interpretability.

5.7 Towards AI for Broader ML and Scientific Domains

Token reduction methods can also offer powerful opportunities to reshape broader machine learning and scientific applications. In particular, domains such as medicine, biology, chemistry, and temporal data analysis frequently encounter complex data structures, heterogeneous data sources, and intricate domain-specific relationships. Informed tokenization approaches promise to address these challenges by transforming complex and rich scientific data into concise, informative, and flexible representations, significantly enhancing the utility of transformer-based foundation models across these domains.

Building Biomedical Tokenizers. Recent works exemplify the transformative potential of advanced tokenization methods in the biomedical domain, including protein [107, 26, 83], genomic [22], and chemical structure [102] tokenizers. Collectively, these methods illustrate how informed reduction and condensation of input tokens can lead to more effective and interpretable scientific models. For example, traditional tokenizers in EHR foundation models typically treat medical codes as isolated textual units, neglecting their inherent structured and relational context, such as hierarchical relationships, disease co-occurrences, and drug-treatment associations found within biomedical ontologies. To solve this issue, MedTok [80] integrates textual descriptions and graph-based relational data into a unified tokenization framework. It first uses a language model encoder to extract embeddings from medical code descriptions and employs a graph encoder to capture relational structures from biomedical ontologies. These embeddings are combined into a compact token space through vector quantization, preserving both modality-specific and cross-modality information.

To enhance informativeness and reduce redundancy, MedTok employs a token packing mechanism. It optimizes shared tokens and modality-specific tokens, ensuring that the final tokens encode both shared semantic meaning and modality-specific structure. This process drastically reduces effective vocabulary size, addressing the scalability challenge of 600,000+ medical codes by collapsing redundant representations while preserving critical clinical context. Inspired by adaptive tokenization methods for vision [21, 103, 42], future EHR tokenization would be adaptive, enabling dynamic representation of patients' medical histories, where the length of the token series of each patient history would be directly correlated to length and complexity. Such adaptive tokenization can significantly improve training and inference efficiency across diverse healthcare systems.

Time-Series Data and Clinical Reasoning. Temporal dynamics form an essential component of clinical reasoning, particularly through longitudinal patient data like lab tests and vital signs. However, current large language models struggle to effectively incorporate time-series inputs due to challenges in temporal tokenization [79, 5, 23, 79, 64]. Future tokenization methods should not only dynamically adjust the number of tokens according to temporal complexity but also selectively focus on time segments most relevant to the clinical context, prompt, or task at hand [84]. This could enhance training effectiveness and inference accuracy, helping create the next generation of EHR foundation models, which are flexible not only over different tasks or prompts, but also over different data sources, patients, and populations. The complexity and richness of EHR data offer opportunities for AI-driven advancements in patient health outcomes. Future EHR models should support comprehensive reasoning capabilities, encompassing complete patient histories, such as vitals, lab results, diagnoses, and procedures over time. They could facilitate timely disease predictions, accurately forecast chronic disease trajectories, and anticipate patient responses to treatments.

5.8 Algorithm-Hardware Co-Design

While algorithmic advancements in token reduction have achieved impressive computational savings, the next crucial step is to integrate these techniques with hardware-aware design principles. We posit that algorithm-hardware co-design is essential for holistic optimization across the compute stack, considering the interplay between algorithmic choices, hardware architectures (specialized data paths, memory hierarchies, communication fabrics, control logic, etc.), and compiler/runtime support (efficient sparse mapping, dynamic scheduling, irregular-data management, etc.) [20, 69].

Currently, co-design efforts targeted at token reduction lag significantly behind pure algorithmic research. This gap is problematic because hardware design needs to balance PPA (power, performance, and area), platform specifics, data movement costs, control overhead, and scalability/reusability [68]. Algorithms developed in isolation often generate sparse or irregular compute patterns that general-purpose hardware cannot exploit effectively. Therefore, future research should aim to: 1) Design parameterizable, reconfigurable accelerator modules-such as on-the-fly importance-scoring units and sparse-data pipelines-that natively support token-reduced Transformers. 2) Explore Processing-in-Memory (PIM) architectures to alleviate severe memory bottlenecks caused by dynamic token pruning. By executing scoring operations or partial attention mechanisms within or near memory arrays, PIM can drastically reduce data movement costs and improve end-to-end efficiency.

6 Conclusion

In this position paper, we have argued that token reduction must evolve beyond a mere efficiency optimization to become a core design principle in generative modeling. We have shown how principled token reduction can address key challenges such as enhancing semantic fidelity in vision-language alignment, curbing verbose reasoning trajectories, preserving long-range coherence, and stabilizing learning dynamics. Looking ahead, we outlined a roadmap of promising directions. We anticipate that embracing token reduction as a holistic, task-aware mechanism will yield the next generation of generative architectures that are not only more efficient but also more robust, interpretable, and aligned with real-world demands.

References

[1] Griffin Adams, Alexander R Fabbri, Faisal Ladhak, Eric Lehman, and Noémie Elhadad. From sparse to dense: Gpt-4 summarization with chain of density prompting. *EMNLP, 4th New Frontier Summarization Workshop*, page 68, 2023.

[2] Yash Akhauri, Ahmed F AbouElhamayed, Yifei Gao, Chi-Chih Chang, Nilesh Jain, and Mohamed S Abdelfattah. Tokenbutler: Token importance is predictable. *arXiv preprint arXiv:2503.07518*, 2025.

[3] Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. *CVPR*, 2025.

[4] Sotiris Anagnostidis, Dario Pavllo, Luca Biggio, Lorenzo Noci, Aurelien Lucchi, and Thomas Hofmann. Dynamic context pruning for efficient and interpretable autoregressive transformers. *NeurIPS*, 36:65202–65223, 2023.

[5] Md Fahim Anjum. Lipcot: Linear predictive coding based tokenizer for self-supervised learning of time series data via language models. *arXiv preprint arXiv:2408.07292*, 2024.

[6] Kazi Hasan Ibn Arif, JinYi Yoon, Dimitrios S Nikolopoulos, Hans Vandierendonck, Deepu John, and Bo Ji. Hired: Attention-guided token dropping for efficient inference of high-resolution vision-language models. In *AAAI*, volume 39, pages 1773–1781, 2025.

[7] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[8] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *ICLR*, 2023.

[9] Daniel Bolya and Judy Hoffman. Token merging for fast stable diffusion. In *CVPR*, pages 4599–4603, 2023.

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020.

[11] Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. Matryoshka multimodal models. In *ICLR*, 2025.

[12] Jianjian Cao, Peng Ye, Shengze Li, Chong Yu, Yansong Tang, Jiwen Lu, and Tao Chen. Madtp: Multimodal alignment-guided dynamic token pruning for accelerating vision-language transformer. In *CVPR*, pages 15710–15719, 2024.

[13] Qingqing Cao, Bhargavi Paranjape, and Hannaneh Hajishirzi. Pumer: Pruning and merging tokens for efficient vision language models. *ACL*, 2023.

[14] Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. In *CVPR*, pages 13817–13827, 2024.

[15] Feng Chen, Yefei He, Lequan Lin, Jing Liu, Bohan Zhuang, and Qi Wu. Zipr1: Reinforcing token sparsity in mllms. *arXiv preprint arXiv:2504.18579*, 2025.

[16] Lei Chen, Zhan Tong, Yibing Song, Gangshan Wu, and Limin Wang. Efficient video action detection with token dropout and context refinement. In *ICCV*, pages 10388–10399, 2023.

[17] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *ECCV*, pages 19–35. Springer, 2024.

[18] Alexis Chevalier, Alexander Wettig, Anirudh Ajith, and Danqi Chen. Adapting language models to compress contexts. *EMNLP*, 2023.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[20] Peiyan Dong, Mengshu Sun, Alec Lu, Yanyue Xie, Kenneth Liu, Zhenglun Kong, Xin Meng, Zhengang Li, Xue Lin, Zhenman Fang, et al. Heatvit: Hardware-efficient adaptive token pruning for vision transformers. In *2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 442–455. IEEE, 2023.

[21] Shivam Duggal, Phillip Isola, Antonio Torralba, and William T Freeman. Adaptive length image tokenization via recurrent allocation. In *First Workshop on Scalable Optimization for Efficient and Adaptive Foundation Models*, 2024.

[22] Bell Raj Eapen. Genomic tokenizer: Toward a biology-driven tokenization in transformer models for dna sequences. *bioRxiv*, pages 2025–04, 2025.

[23] Liri Fang, Yuncong Chen, Wenchao Yu, Yanchi Liu, Lu-an Tang, Vetle I Torvik, and Haifeng Chen. Tsla: A multi-task time series language model. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[24] Qichen Fu, Minsik Cho, Thomas Merth, Sachin Mehta, Mohammad Rastegari, and Mahyar Najibi. Lazyllm: Dynamic token pruning for efficient long context llm inference. *arXiv preprint arXiv:2407.14057*, 2024.

[25] Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Masked diffusion transformer is a strong image synthesizer. In *ICCV*, pages 23164–23173, 2023.

[26] Zhangyang Gao, Cheng Tan, Jue Wang, Yufei Huang, Lirong Wu, and Stan Z Li. Foldtoken: Learning protein language via vector quantization and beyond. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 219–227, 2025.

[27] Tao Ge, Jing Hu, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. In-context autoencoder for context compression in a large language model. *ICLR*, 2024.

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

[29] Saurabh Goyal, Anamitra Roy Choudhury, Saurabh Raje, Venkatesan Chakaravarthy, Yogish Sabharwal, and Ashish Verma. Power-bert: Accelerating bert inference via progressive word-vector elimination. In *ICML*, pages 3690–3699. PMLR, 2020.

[30] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[31] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing XU, and Yunhe Wang. Transformer in transformer. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *NeurIPS*, volume 34, pages 15908–15919. Curran Associates, Inc., 2021.

[32] Joakim Bruslund Haurum, Sergio Escalera, Graham W Taylor, and Thomas B Moeslund. Agglomerative token clustering. In *European Conference on Computer Vision*, pages 200–218. Springer, 2024.

[33] Xiangyu Hong, Che Jiang, Biqing Qi, Fandong Meng, Mo Yu, Bowen Zhou, and Jie Zhou. On the token distance modeling ability of higher rope attention dimension. *EMNLP Findings*, 2024.

[34] Bairu Hou, Yang Zhang, Jiabao Ji, Yujian Liu, Kaizhi Qian, Jacob Andreas, and Shiyu Chang. Thinkprune: Pruning long chain-of-thought of llms via reinforcement learning. *arXiv preprint arXiv:2504.01296*, 2025.

[35] Wenxuan Huang, Zijie Zhai, Yunhang Shen, Shaosheng Cao, Fei Zhao, Xiangfeng Xu, Zheyu Ye, Yao Hu, and Shaohui Lin. Dynamic-llava: Efficient multimodal large language models via dynamic vision-language context sparsification. *ICLR*, 2025.

[36] Xijie Huang, Li Lyna Zhang, Kwang-Ting Cheng, Fan Yang, and Mao Yang. Fewer is more: Boosting llm reasoning with reinforced context pruning. *EMNLP*, 2024.

[37] Xin Huang, Ashish Khetan, Rene Bidart, and Zohar Karnin. Pyramid-bert: Reducing complexity via successive core-set based token selection. *ACL*, 2022.

[38] Ke Ji, Jiahao Xu, Tian Liang, Qiuzhi Liu, Zhiwei He, Xingyu Chen, Xiaoyuan Liu, Zhijie Wang, Junying Chen, Benyou Wang, et al. The first few tokens are all you need: An efficient and effective unsupervised prefix fine-tuning method for reasoning models. *arXiv preprint arXiv:2503.02875*, 2025.

[39] Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. Llmlingua: Compressing prompts for accelerated inference of large language models. *EMNLP*, 2023.

[40] Jindong Jiang, Xiuyu Li, Zhijian Liu, Muyang Li, Guo Chen, Zhiqi Li, De-An Huang, Guilin Liu, Zhiding Yu, Kurt Keutzer, et al. Token-efficient long video understanding for multimodal llms. *arXiv preprint arXiv:2503.04130*, 2025.

[41] Chen Ju, Haicheng Wang, Haozhe Cheng, Xu Chen, Zhonghua Zhai, Weilin Huang, Jinsong Lan, Shuai Xiao, and Bo Zheng. Turbo: Informativity-driven acceleration plug-in for vision-language large models. In *ECCV*, pages 436–455. Springer, 2024.

[42] Minseo Kang and Byunghan Lee. Tictok: Time-series anomaly detection with contrastive tokenization. *IEEE Access*, 11:81011–81020, 2023.

[43] Oğuzhan Fatih Kar, Alessio Tonioni, Petra Poklukar, Achin Kulshrestha, Amir Zamir, and Federico Tombari. Brave: Broadening the visual encoding of vision-language models. In *ECCV*, pages 113–132. Springer, 2024.

[44] Gyuwan Kim and Kyunghyun Cho. Length-adaptive transformer: Train once with length drop, use anytime with search. *ACL*, 2021.

[45] Minchul Kim, Shangqian Gao, Yen-Chang Hsu, Yilin Shen, and Hongxia Jin. Token fusion: Bridging the gap between token pruning and token merging. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1383–1392, 2024.

[46] Sehoon Kim, Sheng Shen, David Thorsley, Amir Gholami, Woosuk Kwon, Joseph Hassoun, and Kurt Keutzer. Learned token pruning for transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 784–794, 2022.

[47] Yeachan Kim, Junho Kim, Jun-Hyung Park, Mingyu Lee, and SangKeun Lee. Leap-of-thought: Accelerating transformers via dynamic token routing. In *EMNLP*, pages 15757–15769, 2023.

[48] Zhenglun Kong, Peiyan Dong, Xiaolong Ma, Xin Meng, Wei Niu, Mengshu Sun, Xuan Shen, Geng Yuan, Bin Ren, Hao Tang, et al. Spvit: Enabling faster vision transformers via latency-aware soft token pruning. In *ECCV*, pages 620–640. Springer, 2022.

[49] Seon-Ho Lee, Jue Wang, Zhikang Zhang, David Fan, and Xinyu Li. Video token merging for long video understanding. *NeurIPS*, 37:13851–13871, 2024.

[50] Siyuan Li, Luyuan Zhang, Zedong Wang, Juanxi Tian, Cheng Tan, Zicheng Liu, Chang Yu, Qingsong Xie, Haonan Lu, Haoqian Wang, et al. Mergevq: A unified framework for visual generation and representation with disentangled token merging and quantization. *arXiv preprint arXiv:2504.00999*, 2025.

[51] Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. Tokenpacker: Efficient visual projector for multimodal llm. *arXiv preprint arXiv:2407.02392*, 2024.

[52] Xirui Li, Chao Ma, Xiaokang Yang, and Ming-Hsuan Yang. Vidtome: Video token merging for zero-shot video editing. In *CVPR*, pages 7486–7495, 2024.

[53] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *ECCV*, pages 323–340. Springer, 2024.

[54] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024.

[55] Yize Li, Yihua Zhang, Sijia Liu, and Xue Lin. Pruning then reweighting: Towards data-efficient training of diffusion models. In *IEEE ICASSP*, 2025.

[56] Weicong Liang, Yuhui Yuan, Henghui Ding, Xiao Luo, Weihong Lin, Ding Jia, Zheng Zhang, Chao Zhang, and Han Hu. Expediting large-scale vision transformer for dense prediction without fine-tuning. *NeurIPS*, 35:35462–35477, 2022.

[57] Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. *ICLR*, 2022.

[58] Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, Weizhu Chen, et al. Not all tokens are what you need for pretraining. *NeurIPS*, 37:29029–29063, 2024.

[59] Zhihang Lin, Mingbao Lin, Luxi Lin, and Rongrong Ji. Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In *AAAI*, volume 39, pages 5334–5342, 2025.

[60] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023.

[61] Xiangrui Liu, Yan Shu, Zheng Liu, Ao Li, Yang Tian, and Bo Zhao. Video-xl-pro: Reconstructive token compression for extremely long video understanding. *arXiv preprint arXiv:2503.18478*, 2025.

[62] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. Ts2-net: Token shift and selection transformer for text-video retrieval. In *ECCV*, pages 319–335. Springer, 2022.

[63] Zhihang Liu, Chen-Wei Xie, Pandeng Li, Liming Zhao, Longxiang Tang, Yun Zheng, Chuanbin Liu, and Hongtao Xie. Hybrid-level instruction injection for video token compression in multi-modal large language models. *CVPR*, 2025.

[64] Luca Masserano, Abdul Fatir Ansari, Boran Han, Xiyuan Zhang, Christos Faloutsos, Michael W Mahoney, Andrew Gordon Wilson, Youngsuk Park, Syama Rangapuram, Danielle C Maddix, et al. Enhancing foundation models for time series forecasting via wavelet-based tokenization. *arXiv preprint arXiv:2412.05244*, 2024.

[65] Jesse Mu, Xiang Li, and Noah Goodman. Learning to compress prompts with gist tokens. *NeurIPS*, 36:19327–19352, 2023.

[66] Piotr Nawrot, Jan Chorowski, Adrian Łańcucki, and Edoardo M Ponti. Efficient transformers with dynamic token pooling. *ACL*, 2023.

[67] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2024.

[68] Yue Pan, Minxuan Zhou, Chonghan Lee, Zheyu Li, Rishika Kushwah, Vijaykrishnan Narayanan, and Tajana Rosing. Primate: Processing in memory acceleration for dynamic token-pruning transformers. In *Proceedings of the 29th Asia and South Pacific Design Automation Conference*, ASPDAC '24, page 557âĂŞ563. IEEE Press, 2024.

[69] Dhruv Parikh, Shouyi Li, Bingyi Zhang, Rajgopal Kannan, Carl Busart, and Viktor Prasanna. Accelerating vit inference on fpga through static and dynamic pruning. In *2024 IEEE 32nd Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*, pages 78–89. IEEE, 2024.

[70] William Peebles and Saining Xie. Scalable diffusion models with transformers. *ICCV*, 2023.

[71] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *NeurIPS*, 34:13937–13949, 2021.

[72] Shuhuai Ren, Sishuo Chen, Shicheng Li, Xu Sun, and Lu Hou. Testa: Temporal-spatial token aggregation for long-form video-language understanding. In *EMNLP*, 2023.

[73] Michael Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. *NeurIPS*, 34:12786–12797, 2021.

[74] Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*, 2024.

[75] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[76] Leqi Shen, Tianxiang Hao, Tao He, Sicheng Zhao, Yifeng Zhang, Pengzhang Liu, Yongjun Bao, and Guiguang Ding. Tempme: Video temporal token merging for efficient text-video retrieval. *ICLR*, 2025.

[77] Xuan Shen, Peiyan Dong, Lei Lu, Zhenglun Kong, Zhengang Li, Ming Lin, Chao Wu, and Yanzhi Wang. Agile-quant: Activation-guided quantization for faster inference of llms on the edge. In *AAAI*, volume 38, pages 18944–18951, 2024.

[78] Dingjie Song, Wenjun Wang, Shunian Chen, Xidong Wang, Michael Guan, and Benyou Wang. Less is more: A simple yet effective token reduction method for efficient multi-modal llms. *COLING*, 2025.

[79] Dimitris Spathis and Fahim Kawsar. The first step is the hardest: Pitfalls of representing and tokenizing temporal data for large language models. *Journal of the American Medical Informatics Association*, 31(9):2151–2158, 2024.

[80] Xiaorui Su, Shvat Messica, Yepeng Huang, Ruth Johnson, Lukas Fesser, Shanghua Gao, Faryad Sahneh, and Marinka Zitnik. Multimodal medical code tokenizer. *arXiv preprint arXiv:2502.04397*, 2025.

[81] Yi Su, Yuechi Zhou, Quantong Qiu, Juntao Li, Qingrong Xia, Ping Li, Xinyu Duan, Zhefeng Wang, and Min Zhang. Accurate kv cache quantization with outlier tokens tracing. *ACL*, 2025.

[82] Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Hanjie Chen, et al. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:2503.16419*, 2025.

[83] Burak Suyunu, Özdeniz Dolu, and Arzucan Özgür. evobpe: Evolutionary protein sequence tokenization. *arXiv preprint arXiv:2503.08838*, 2025.

[84] Sabera Talukder, Yisong Yue, and Georgia Gkioxari. Totem: Tokenized time series embeddings for general time series analysis. *arXiv preprint arXiv:2402.16412*, 2024.

[85] Yehui Tang, Kai Han, Yunhe Wang, Chang Xu, Jianyuan Guo, Chao Xu, and Dacheng Tao. Patch slimming for efficient vision transformers. In *CVPR*, pages 12165–12174, 2022.

[86] Yao Tao, Yehui Tang, Yun Wang, Mingjian Zhu, Hailin Hu, and Yunhe Wang. Saliency-driven dynamic token pruning for large language models. *arXiv preprint arXiv:2504.04514*, 2025.

[87] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

[88] Hanrui Wang, Zhekai Zhang, and Song Han. Spatten: Efficient sparse attention architecture with cascade token and head pruning. In *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 97–110. IEEE, 2021.

[89] Hongjie Wang, Difan Liu, Yan Kang, Yijun Li, Zhe Lin, Niraj K Jha, and Yuchen Liu. Attention-driven training-free efficiency enhancement of diffusion models. In *CVPR*, pages 16080–16089, 2024.

[90] Junke Wang, Xitong Yang, Hengduo Li, Li Liu, Zuxuan Wu, and Yu-Gang Jiang. Efficient video transformers with spatial-temporal token selection. In *ECCV*, pages 69–86. Springer, 2022.

[91] Rui Wang, Hongru Wang, Boyang Xue, Jianhui Pang, Shudong Liu, Yi Chen, Jiahao Qiu, Derek Fai Wong, Heng Ji, and Kam-Fai Wong. Harnessing the reasoning economy: A survey of efficient reasoning for large language models. *arXiv preprint arXiv:2503.24377*, 2025.

[92] Yixiao Wang, Yifei Zhang, Mingxiao Huo, Ran Tian, Xiang Zhang, Yichen Xie, Chenfeng Xu, Pengliang Ji, Wei Zhan, Mingyu Ding, et al. Sparse diffusion policy: A sparse, reusable, and flexible policy for robot learning. *arXiv preprint arXiv:2407.01531*, 2024.

[93] Siyuan Wei, Tianzhu Ye, Shen Zhang, Yao Tang, and Jiajun Liang. Joint token pruning and squeezing towards more aggressive compression of vision transformers. In *CVPR*, pages 2092–2101, 2023.

[94] Zichen Wen, Yifeng Gao, Weijia Li, Conghui He, and Linfeng Zhang. Token pruning in multimodal large language models: Are we solving the right problem? *arXiv preprint arXiv:2502.11501*, 2025.

[95] David Wingate, Mohammad Shoeybi, and Taylor Sorensen. Prompt compression and contrastive conditioning for controllability and toxicity reduction in language models. *EMNLP Findings*, 2022.

[96] Qiong Wu, Wenhao Lin, Weihao Ye, Yiyi Zhou, Xiaoshuai Sun, and Rongrong Ji. Accelerating multimodal large language models via dynamic visual-token exit and the empirical findings. *arXiv preprint arXiv:2411.19628*, 2024.

[97] Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Towards semantic equivalence of tokenization in multimodal llm. *ICLR*, 2025.

[98] Tong Wu, Junzhe Shen, Zixia Jia, Yuxuan Wang, and Zilong Zheng. From hours to minutes: Lossless acceleration of ultra long sequence generation up to 100k tokens. *arXiv preprint arXiv:2502.18890*, 2025.

[99] Xinjian Wu, Fanhu Zeng, Xiudong Wang, and Xinghao Chen. Ppt: Token pruning and pooling for efficient vision transformers. *arXiv preprint arXiv:2310.01812*, 2023.

[100] Heming Xia, Yongqi Li, Chak Tou Leong, Wenjie Wang, and Wenjie Li. Tokenskip: Controllable chain-of-thought compression in llms. *arXiv preprint arXiv:2502.12067*, 2025.

[101] Wenyi Xiao, Leilei Gan, Weilong Dai, Wanggui He, Ziwei Huang, Haoyuan Li, Fangxun Shu, Zhelun Yu, Peng Zhang, Hao Jiang, et al. Fast-slow thinking for large vision-language model reasoning. *arXiv preprint arXiv:2504.18458*, 2025.

[102] Keqiang Yan, Xiner Li, Hongyi Ling, Kenna Ashen, Carl Edwards, Raymundo Arróyave, Marinka Zitnik, Heng Ji, Xiaofeng Qian, Xiaoning Qian, et al. Invariant tokenization of crystalline materials for language model enabled generation. *Advances in Neural Information Processing Systems*, 37:125050–125072, 2024.

[103] Wilson Yan, Volodymyr Mnih, Aleksandra Faust, Matei Zaharia, Pieter Abbeel, and Hao Liu. Elastictok: Adaptive tokenization for image and video. *ICLR*, 2025.

[104] Cheng Yang, Yang Sui, Jinqi Xiao, Lingyi Huang, Yu Gong, Chendi Li, Jinghua Yan, Yu Bai, Ponnuswamy Sadayappan, Xia Hu, and Bo Yuan. Topv: Compatible token pruning with inference time optimization for fast and low-memory multimodal vision language model. *CVPR*, 2025.

[105] Deming Ye, Yankai Lin, Yufei Huang, and Maosong Sun. Tr-bert: Dynamic token reduction for accelerating bert inference. *NAACL*, 2021.

[106] Xubing Ye, Yukang Gan, Xiaoke Huang, Yixiao Ge, and Yansong Tang. Voco-llama: Towards vision compression with large language models. *CVPR*, 2025.

[107] Xinyu Yuan, Zichen Wang, Marcus Collins, and Huzefa Rangwala. Protein structure tokenization: Benchmarking and new recipe. *arXiv preprint arXiv:2503.00089*, 2025.

[108] Fanhu Zeng and Deli Yu. M2m-tag: Training-free many-to-many token aggregation for vision transformer acceleration. In *Workshop on Machine Learning and Compression, NeurIPS*, 2024.

[109] Yuting Zeng, Weizhe Huang, Lei Jiang, Tongxuan Liu, Xitai Jin, Chen Tianying Tiana, Jing Li, and Xiaohua Xu. S^2-mad: Breaking the token barrier to enhance multi-agent debate efficiency. *NAACL*, 2025.

[110] Zheng Zhan, Zhenglun Kong, Yifan Gong, et al. Exploring token pruning in vision state space models. In *NeurIPS*, 2024.

[111] Zheng Zhan, Yushu Wu, Zhenglun Kong, et al. Rethinking token reduction for state space models. In *EMNLP*, pages 1686–1697, Miami, Florida, USA, nov 2024. ACL.

[112] Diankun Zhang, Guoan Wang, Runwen Zhu, Jianbo Zhao, Xiwu Chen, Siyu Zhang, Jiahao Gong, Qibin Zhou, Wenyuan Zhang, Ningzi Wang, et al. Sparsead: Sparse query-centric paradigm for efficient end-to-end autonomous driving. *arXiv preprint arXiv:2404.06892*, 2024.

[113] Qizhe Zhang, Aosong Cheng, Ming Lu, Renrui Zhang, Zhiyong Zhuo, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. *arXiv preprint arXiv:2412.01818*, 2025.

[114] Shuheng Zhang, Yuqi Liu, Hongbo Zhou, Jun Peng, Yiyi Zhou, Xiaoshuai Sun, and Rongrong Ji. Adaflow: Efficient long video editing via adaptive attention slimming and keyframe selection. *arXiv preprint arXiv:2502.05433*, 2025.

[115] Wangbo Zhao, Yizeng Han, Jiasheng Tang, Kai Wang, Yibing Song, Gao Huang, Fan Wang, and Yang You. Dynamic diffusion transformer. *ICLR*, 2025.

[116] Chang Zou, Xuyang Liu, Ting Liu, Siteng Huang, and Linfeng Zhang. Accelerating diffusion transformers with token-wise feature caching. In *ICLR*, 2025.