A Survey on Token Compression Techniques and Their Impact on Transformer Model Efficiency

www.surveyx.cn

Abstract

This survey explores the critical role of token compression techniques in enhancing the efficiency and scalability of Transformer models across diverse applications. Token compression, encompassing pruning, reduction, and merging, addresses the substantial computational demands of large-scale models, particularly in vision-language models, image and video generation, and large language models. Techniques like ATP-LLaVA demonstrate adaptive token pruning's effectiveness, achieving significant token count reductions while maintaining high performance. In generative tasks, CogVideoX exemplifies state-of-the-art performance in text-to-video generation, underscoring token compression's importance in optimizing complex models. Cropr showcases the versatility of token pruning in accelerating Vision Transformers with minimal performance penalties across multiple tasks. Efficient pruning in text-to-image models significantly reduces model size without compromising image quality, highlighting these techniques' transformative impact on model scalability. The survey also addresses challenges such as balancing efficiency and semantic richness, adapting to diverse modalities, and optimizing pruning strategies. By integrating these strategies, Transformer models achieve enhanced adaptability and computational efficiency, essential for handling complex and resource-intensive tasks. Overall, token compression techniques are pivotal in advancing AI systems' capabilities, enabling the deployment of efficient and scalable solutions for complex computational tasks with enhanced performance and reduced computational demands.

1 Introduction

1.1 Importance of Efficiency in Transformer Models

Efficiency in Transformer models is paramount in large-scale applications due to their significant computational and memory demands. The attention mechanism central to these models outperforms traditional convolutional and recurrent architectures in natural language processing (NLP), highlighting the necessity for efficient processing strategies [1]. The high computational costs of Vision Transformers (ViTs) in object detection and instance segmentation further underscore the urgency for efficiency improvements [2].

The quadratic complexity of Multimodal Large Language Models (MLLMs) poses substantial challenges as sequence lengths increase, necessitating efficient designs for scalability across various domains [3]. The excessive utilization of visual tokens in MLLMs like LLaVA accentuates the critical need for efficiency enhancements [4]. In Vision-Language Models (VLMs), high computational costs hinder practical usability, driving the demand for resource optimization strategies [5].

Moreover, the large size of text-to-image models restricts their deployment on resource-constrained devices, further emphasizing the importance of efficiency [6]. The computational complexity involved in generating long-duration, temporally consistent videos with rich motion semantics in diffusion models illustrates the essential role of efficiency in enabling advanced applications [7]. The chal-

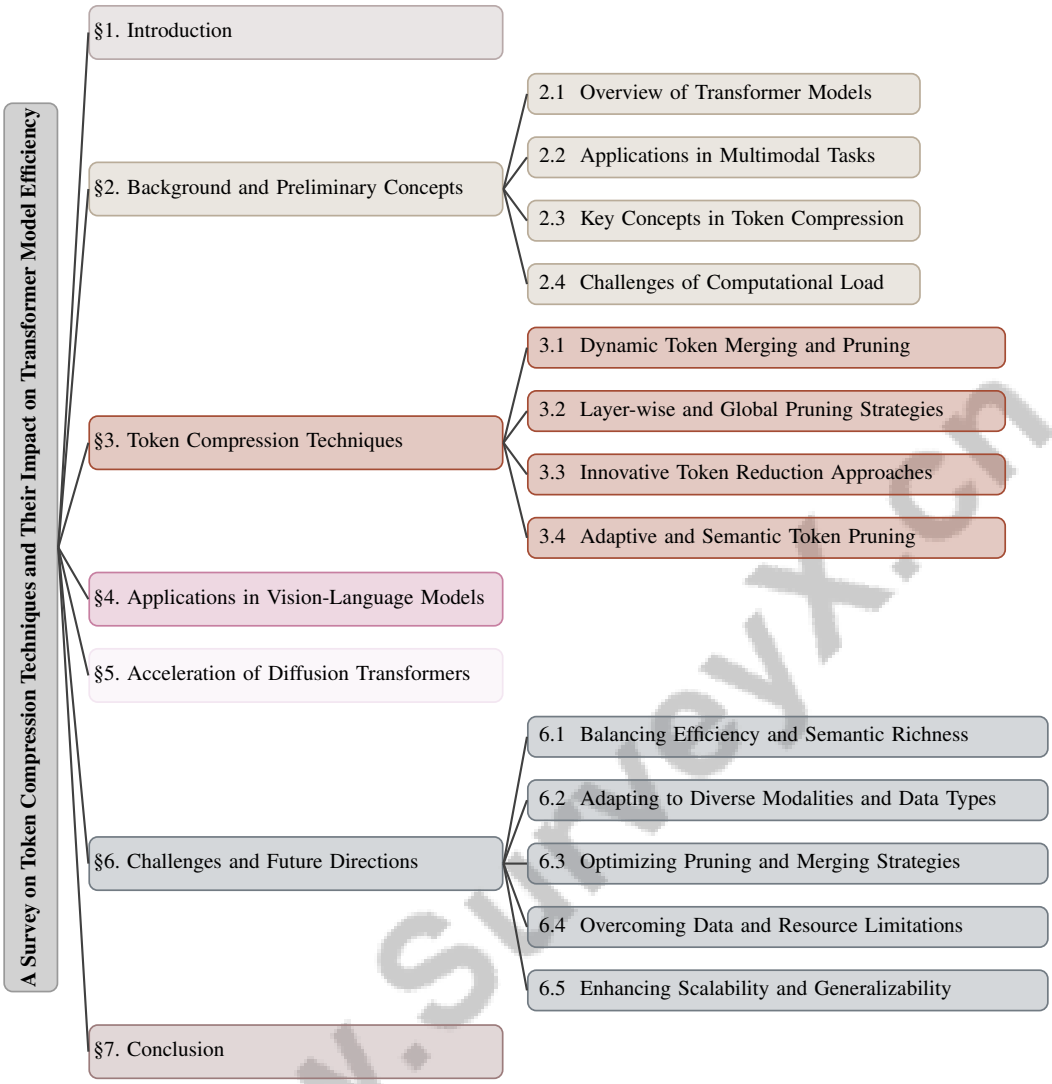

Figure 1: chapter structure

lenges of producing high-resolution images and long videos highlight the necessity of optimizing Transformer models for efficient performance [8].

Addressing these issues is vital for enhancing the scalability and applicability of Transformers across diverse, resource-intensive domains, driven by the inherent computational and memory demands of large-scale applications [9].

1.2 Role of Token Compression Techniques

Token compression techniques are essential for improving the efficiency of Transformer models by alleviating the computational and memory burdens associated with extensive datasets. Techniques such as token pruning, reduction, and merging are designed to optimize model performance while minimizing accuracy loss. For instance, LazyLLM employs a dynamic token pruning method that selectively computes key-value pairs for significant tokens, addressing inefficiencies of static pruning methods [10]. Similarly, SpAtten enhances efficiency by pruning unimportant tokens and heads within the attention mechanism, thereby reducing computational demands [1].

In ViTs, token compression methods like token and channel pruning are crucial for managing high computational costs while preserving accuracy and inference speed [11]. Techniques such as VisToG further improve efficiency by intelligently reducing redundant visual tokens while maintaining

critical semantic information [12]. Focal Pruning (FoPru) exemplifies effective token compression by decreasing redundant visual tokens without requiring retraining, streamlining Transformer models [13].

The integration of token compression strategies, including pruning, combining, and merging, significantly enhances the scalability and efficiency of Transformer models. These approaches reduce computational costs and memory requirements while preserving or even enhancing model accuracy. For example, Learned Token Pruning (LTP) adaptively eliminates less critical tokens based on learned attention scores, achieving up to a 2.1× reduction in FLOPs with minimal accuracy loss. Additionally, techniques like PITOME prioritize the retention of informative tokens during merging, yielding 40-60

1.3 Focus of the Survey

This survey investigates the pivotal role of token compression techniques across various applications, emphasizing their impact on vision-language models, image and video generation, and large language models. The focus is on how these techniques enhance the efficiency and performance of Transformer models by reducing computational load while maintaining or improving model accuracy. In vision-language models, token compression is critical for optimizing resource usage, facilitating deployment in practical scenarios. In image and video generation, the survey addresses challenges and advancements in generative learning within non-textual modalities, noting that video generation has lagged behind language foundation models [14]. Furthermore, the survey explores the integration of token compression strategies in large language models, highlighting the need for scalable solutions to manage increasing dataset complexity and size. Through this comprehensive examination, the survey seeks to provide insights into the current state and future directions of token compression techniques in enhancing AI system capabilities across diverse applications.

1.4 Structure of the Survey

The survey is meticulously structured to explore token compression techniques and their impact on Transformer model efficiency. It begins with an **Introduction**, discussing the importance of efficiency in Transformer models, especially for large-scale applications, and the role of token compression techniques in enhancing computational efficiency. This is followed by a detailed **Background and Preliminary Concepts** section, providing an overview of Transformer models and their applications in multimodal tasks, along with fundamental concepts such as token compression, pruning, reduction, and merging, and the challenges posed by computational load in large models.

The core of the survey focuses on **Token Compression Techniques**, detailing various strategies including dynamic token merging and pruning, layer-wise and global pruning strategies, innovative token reduction approaches, and adaptive and semantic token pruning. Each subsection discusses methodologies and algorithms, supported by relevant literature.

Subsequently, the survey examines **Applications in Vision-Language Models**, highlighting the application of token compression techniques in enhancing model efficiency and performance in tasks such as video understanding and image generation. The section on **Acceleration of Diffusion Transformers** explores the role of token compression in accelerating diffusion transformers, particularly in image and video generation tasks, discussing training-free inference acceleration methods and caching and reuse strategies.

The survey also addresses , emphasizing the complexities of implementing token compression techniques, including trade-offs between model performance and computational efficiency. It discusses the need for standardized evaluations of various methods, such as extractive compression, token pruning, and token merging, which exhibit varying effectiveness across tasks. Furthermore, potential avenues for future research are identified, highlighting the importance of characterizing reduction patterns and enhancing the robustness of compression methods for broader applicability in natural language processing and computer vision contexts [15, 16, 17, 18, 19]. Key issues such as balancing efficiency and semantic richness, adapting to diverse modalities and data types, optimizing pruning and merging strategies, overcoming data and resource limitations, and enhancing scalability and generalizability are examined.

In the **Conclusion**, we encapsulate the pivotal findings of our survey, emphasizing the critical role of token compression techniques, such as token pruning and merging, in significantly improving

Transformer model efficiency. These methods enhance model performance across various NLP and vision tasks while reducing computational costs and memory usage, as demonstrated by our experiments showing up to a 5

2 Background and Preliminary Concepts

2.1 Overview of Transformer Models

Transformer models have revolutionized artificial intelligence, becoming foundational across diverse domains due to their adaptability and efficacy. Originally developed for natural language processing (NLP), Transformers excel in tasks such as text comprehension and generation, underpinning large language models (LLMs) like GPT and BERT. Their self-attention mechanism effectively captures long-range dependencies, enhancing performance in linguistic tasks [20].

In computer vision, Vision Transformers (ViTs) offer a robust alternative to convolutional neural networks (CNNs), achieving competitive results in image analysis. However, their quadratic computational complexity presents challenges for real-time processing, prompting the development of adaptive token pruning and block-structured pruning to reduce computational costs and latency while maintaining performance [21, 22, 23, 11, 24]. These strategies are crucial for deployment on edge devices and in real-time applications.

In video processing, Transformers' flexibility supports various analysis tasks, yet inefficiencies in handling long-form video data remain problematic, as computational costs rise with sequence length [7]. This is notably evident in diffusion models, where quadratic scaling with input length hinders high-resolution image generation [25].

Transformers are also pivotal in multimodal applications, particularly in Vision-Language Models (VLMs) that integrate visual and textual data. The computational demands necessitate innovations in tokenization and model optimization [26]. Previous image tokenization methods in LLMs highlight the need for improved strategies to enhance multimodal comprehension and generation [27].

In generative tasks, such as text-to-image and text-to-video generation, models like CogVideoX exemplify the potential of diffusion transformers in creative applications [7]. Despite widespread adoption, ongoing research addresses challenges in deploying ViTs and U-Net architectures on-device, focusing on scalable and efficient diffusion models [28].

Transformers are integral to contemporary AI systems, advancing unimodal tasks—like speech and image recognition—and complex multimodal applications that integrate text, images, and audio. Their self-attention mechanisms enhance performance across domains, enabling sophisticated applications such as text-to-image generation, image captioning, and visual question answering. The ability of Transformers to efficiently learn representations and fuse information across modalities underscores their pivotal role in AI evolution [29, 30, 31, 15]. Their significance is evident in superior performance and versatility across a broad spectrum of applications, from text processing to complex vision-language integration.

2.2 Applications in Multimodal Tasks

Transformer models are essential for multimodal tasks, adeptly integrating and processing diverse data types for complex operations such as image captioning, text-to-image generation, and visual question answering. They excel in managing sequential data, crucial for maintaining temporal consistency and efficient memory usage in tasks like video generation [32]. In video generation, particularly when combined with diffusion models, Transformers tackle the challenges of creating high-quality videos from varied inputs, leveraging their capacity to manage long-range dependencies [33].

The integration of audio and visual data in multimodal tasks is enhanced by diffusion models, which facilitate the joint generation of high-quality audio and video content, showcasing Transformers' versatility in complex multimodal applications [34]. Despite their success in language-based generative tasks, diffusion models often outperform LLMs in image and video generation, indicating the necessity for specialized approaches [35].

In text-to-image generation, various generative models, including GANs, autoregressive models, and diffusion models, have been employed to bridge the gap between textual and visual data, emphasizing

the expanding capabilities of Transformers [34]. Additionally, advancements in audio-conditioned diffusion models are addressing the challenge of generating high-quality images from audio inputs, further illustrating the growing scope of Transformer applications [36].

VLMs face inefficiencies due to high computational overhead in processing visual tokens, especially in high-resolution images and videos. Addressing data redundancy is critical for enhancing throughput and latency in real-world applications, highlighting the need for optimized token handling strategies [37]. Frameworks like the Multi-scale Efficient Graph-Transformer (MEGT) exemplify efforts to enhance classification accuracy and efficiency in managing both low-resolution and high-resolution data [38].

Transformer models are at the forefront of multimodal intelligence, continuously evolving to meet the demands of integrating and processing diverse data types across various applications. Their adaptability and efficiency in handling complex tasks underscore their pivotal role in advancing multimodal AI systems, as they unify diverse modalities to facilitate comprehensive multimodal models [39].

2.3 Key Concepts in Token Compression

Token compression is critical for enhancing Transformer model efficiency by systematically reducing the number of tokens processed, thus alleviating the computational burden of the self-attention mechanism. This approach includes strategies such as token pruning, which eliminates less important tokens based on learned importance scores, and token combining, which condenses input sequences to optimize performance while maintaining accuracy. These methods can lead to significant reductions in memory usage and processing time, with improvements in accuracy and F1 scores of up to 5

Pruning selectively eliminates less important tokens to decrease model complexity and computational demands. The CATP method exemplifies this by emphasizing query tokens based on their cross-attention probabilities with image tokens, optimizing the token selection process [40]. Similarly, LazyLLM dynamically prunes tokens by evaluating their importance for subsequent generation steps, ensuring that only the most relevant tokens are retained [10]. SpAtten introduces cascade token pruning, head pruning, and progressive quantization to streamline computational demands while maintaining model accuracy [1]. Focal Pruning (FoPru) enhances efficiency by removing redundant visual tokens based on significance scores derived from attention maps, allowing for a training-free approach to pruning [13].

Reduction strategies minimize token count while preserving essential information. Techniques such as Visual Token Grouping (VisToG) utilize pre-trained vision encoders to cluster similar visual tokens, effectively reducing the number of tokens processed without compromising critical data [12]. Asymmetric token merging (ATME) and consistent dynamic channel pruning (CDCP) are additional methods that reduce computational costs while maintaining spatial integrity, optimizing resource usage [11]. FastV exemplifies dynamic token pruning by selectively removing less impactful visual tokens in deeper layers based on attention scores, highlighting the importance of adaptive token management [26].

Token merging involves combining similar or non-crucial tokens with more significant ones, thereby reducing the overall token count without sacrificing important information. This approach is exemplified by methods employing importance and similarity scores to guide the merging process, enhancing feature representation while minimizing computational overhead [11]. By leveraging these strategies, token compression techniques effectively manage computational demands, facilitating the deployment of Transformer models across diverse applications while ensuring optimal performance and preserving the integrity of processed information.

2.4 Challenges of Computational Load

The computational load inherent in large Transformer models presents significant challenges, particularly when processing extensive and complex multimodal data. The quadratic complexity of self-attention mechanisms complicates the efficient application of Transformers across various computer vision tasks [20]. This complexity is exacerbated in video understanding models, where substantial computational costs necessitate innovative strategies to enhance efficiency. The overhead

introduced by tokenization and variable-sized blocks in U-Net architectures further complicates real-time deployment, underscoring the need for scalable solutions [28].

In diffusion transformers, challenges are intensified by existing CNN-based models' limitations, which struggle to effectively scale and adapt to diverse text inputs while maintaining high visual quality. This is particularly evident in image and video generation, where traditional CNN architectures, such as U-Net, dominate despite the advantages of Transformer-based diffusion models. Recent advancements, such as GenTron, highlight the potential for enhancing visual quality by significantly scaling model parameters from approximately 900 million to over 3 billion. Additionally, integrating large language models (LLMs) into the text-to-image diffusion framework has revealed issues related to prompt encoding, where misalignment between LLM training paradigms and diffusion model requirements can degrade performance. Innovative strategies like the LLM-Infused Diffusion Transformer (LI-DiT) demonstrate superior prompt understanding and flexibility by effectively harnessing LLM strengths while overcoming positional biases inherent in decoder-only architectures [41, 42]. The computational demands of managing multimodal inputs, alongside the complexity of maintaining coherent generation across modalities, highlight the critical need for advancements in computational efficiency. Furthermore, inefficiencies in current deep neural network architectures, such as 3D CNNs and Transformers, lead to computational redundancy and increased latency during online inference, hindering real-time application.

Quantization of activations and weights in transformer-only structures often requires extensive optimization or retraining, complicating the deployment of efficient models. Existing benchmarks face limitations in computational efficiency, particularly when scaling to high-resolution outputs, necessitating innovative approaches to overcome these constraints. Additionally, inconsistencies in calculation patterns and token selection strategies between fully trained models and dynamic vision transformers present further obstacles, necessitating adaptive approaches to enhance computational efficiency. The inability of current pruning techniques to adaptively allocate model capacities based on varying prompt complexities exacerbates resource utilization inefficiencies. The challenges associated with Transformer models underscore the critical need for efficiency enhancements, essential for increasing both scalability and practical applicability in various resource-intensive domains. Recent advancements, such as Learned Token Pruning (LTP) and PITOME, demonstrate innovative approaches to reduce computational costs and memory usage while maintaining accuracy. LTP adaptively removes less important tokens during processing, achieving significant reductions in floating-point operations (FLOPs) and enhancing throughput, while PITOME merges token representations to optimize performance across tasks like image classification and retrieval. Additionally, methods focused on edge deployment, such as non-linear latency-workload strategies, further illustrate the potential for tailored efficiency improvements that address specific operational constraints. Collectively, these strategies highlight a pathway toward more effective and versatile Transformer architectures capable of meeting the demands of diverse applications [43, 30, 15, 44, 45].

In recent years, the optimization of Transformer models has garnered significant attention within the field of natural language processing. A variety of token compression techniques have been developed to enhance the efficiency of these models. Figure **??** illustrates these techniques, categorizing them into dynamic token merging and pruning, layer-wise and global pruning strategies, as well as innovative token reduction approaches, and adaptive and semantic token pruning. Each category not only highlights specific methods but also elucidates their applications, demonstrating notable improvements in efficiency, reductions in computational costs, and the maintenance of performance across diverse tasks and models. This comprehensive overview underscores the importance of selecting appropriate token compression strategies to achieve optimal model performance while minimizing resource expenditure.

Figure 2: This figure illustrates various token compression techniques for optimizing Transformer models, categorized into dynamic token merging and pruning, layer-wise and global pruning strategies, innovative token reduction approaches, and adaptive and semantic token pruning. Each category highlights specific methods and their applications, demonstrating improvements in efficiency, computational cost reductions, and maintenance of performance across different tasks and models.

3 Token Compression Techniques

3.1 Dynamic Token Merging and Pruning

Method Name	Optimization Techniques	Application Domains	Performance Metrics
LLM[10]	Dynamic Token Pruning	Large Language Models	Ttft Speedup
SpAtten[1]	Cascade Pruning	Nlp Models	Speedup
CAIT[11]	Token Pruning	Multimodal Tasks	Inference Speed
FV[26]	Dynamic Pruning	Vision-language Tasks	Flops Reduction
FoPru[13]	Rank Pruning	Multimodal Tasks	Accuracy, Speedup

Table 1: Overview of various dynamic token pruning and optimization methods, detailing their specific techniques, application domains, and performance metrics. This table highlights the diverse strategies employed to enhance computational efficiency in large language models, NLP models, multimodal tasks, and vision-language tasks.

Dynamic token merging and pruning are essential for optimizing Transformer models by selectively processing tokens based on their relevance, thereby enhancing computational efficiency while preserving critical information. The ColBERT model exemplifies this by reducing index size by up to 30% with minimal performance loss through token pruning. A combined approach with fuzzy logic yields a 5% increase in classification accuracy and a 5.6% improvement in the F1 score, alongside a memory cost reduction to 0.61x and a speedup of 1.64x [18, 17].

As illustrated in Figure 3, the categorization of dynamic token merging and pruning techniques encompasses token pruning approaches, multimodal token merging, and visual token pruning. This figure highlights key methods such as ColBERT, LazyLLM, and FastV, providing a visual representation of their relationships and functionalities. The LazyLLM framework prioritizes token importance during both prefilling and decoding, significantly boosting inference speed [10]. SpAtten demonstrates the efficacy of on-the-fly pruning for tokens and heads in reducing computational demands [1]. VisToG dynamically reduces redundant visual tokens, while Speech2Video employs cross-modal distillation for generating talking face videos, showcasing dynamic token merging in multimodal tasks [12, 46]. The CAIT method combines asymmetric token merging with consistent dynamic channel pruning to compress Vision Transformers (ViTs), achieving high accuracy and rapid inference [11]. FastV enhances efficiency through dynamic image token pruning by learning adaptive attention patterns and selectively pruning visual tokens [26].

Focal Pruning (FoPru) exploits attention distribution to retain critical tokens, significantly boosting inference efficiency [13]. These approaches illustrate the transformative potential of dynamic token merging and pruning in optimizing Transformer performance and scalability, facilitating enhanced adaptability and efficiency in handling complex tasks. Additionally, Table 1 provides a comprehensive summary of dynamic token pruning methods, illustrating their optimization techniques, application domains, and associated performance metrics.

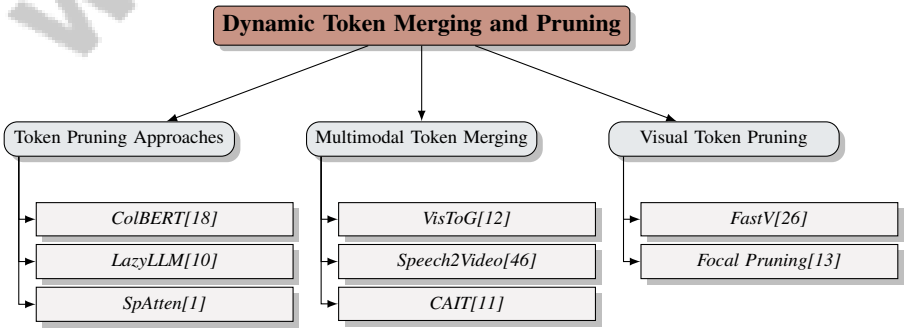

Figure 3: This figure illustrates the categorization of dynamic token merging and pruning techniques into token pruning approaches, multimodal token merging, and visual token pruning, highlighting key methods such as ColBERT, LazyLLM, and FastV.

3.2 Layer-wise and Global Pruning Strategies

Layer-wise and global pruning strategies optimize Transformer models by selectively reducing parameters while maintaining performance. Layer-wise methods target individual layers, while global strategies assess the entire architecture. Efficient Coarse-to-Fine Layer-Wise Pruning (ECoFLaP) leverages both local and global insights to enhance performance and minimize computational costs in large Vision-Language Models (LVLMs) [29, 47].

Layer-wise pruning, such as Token Pruning (ToP), uses improved attention value-based scoring to dynamically remove unimportant tokens during inference [48]. Cropr retains the top K tokens based on importance scores, effectively reducing complexity while preserving essential information [49]. Conversely, global pruning strategies, like FasterVLM, use training-free visual token pruning based on [CLS] token attention scores, allowing effective pruning without retraining [50]. The Selective Vision Transformer (SViT) preserves pruned tokens in feature maps for potential reactivation, maintaining accuracy while reducing computational load [2].

Evo-ViT enhances efficiency by preserving spatial structure during token management [51], while ATP-LLaVA applies layer-specific pruning using self-attention maps for targeted retention and improved efficiency [52]. In text-to-image models, post-training pruning techniques enhance efficiency by targeting both the text encoder and image generator [6].

Layer-wise and global pruning strategies are essential for optimizing Transformer models. Layer-wise pruning, like Learned Token Pruning (LTP), adaptively removes less important tokens based on learned attention scores, significantly enhancing efficiency and throughput. Global methods like Zero-TPrune leverage the attention graph for token pruning without fine-tuning, facilitating flexible deployment across tasks with minimal overhead. These strategies improve adaptability and efficiency in deploying Transformers, particularly in resource-constrained environments [6, 30, 53, 44, 45]. They ensure robust performance across applications, advancing model efficiency and scalability.

3.3 Innovative Token Reduction Approaches

Innovative token reduction techniques enhance Transformer model efficiency by minimizing token redundancy while retaining crucial information. Learned Token Pruning (LTP) and Zero-TPrune exemplify this by dynamically eliminating less important tokens based on learned attention scores, achieving significant computational cost reductions and improved throughput without substantial accuracy loss. PITOME focuses on merging token representations based on energy scores, allowing a reduction of 40-60

The TOken REduction (TORE) method significantly decreases the number of tokens required in Transformer-based Human Mesh Recovery, enhancing recovery accuracy [9]. FiCoCo employs a three-stage process of filtering, correlating, and compressing, achieving substantial computational load reductions without compromising performance [3]. CogVideoX improves video generation quality and coherence through effective token management using a 3D causal VAE and expert transformer [7]. SpAtten enhances attention computation efficiency by leveraging token and head sparsity, demonstrating the efficacy of sparse token management [1].

These innovative token reduction techniques highlight the critical role of strategic token management in optimizing Transformer model efficiency. Methods like Learned Token Pruning, Spectrum-Preserving Token Merging, and Zero-TPrune enhance performance by significantly reducing computational costs, memory usage, and latency. For instance, Learned Token Pruning achieves up to a 2.1× reduction in FLOPs while maintaining accuracy, and Zero-TPrune enables efficient pruning without fine-tuning, making it suitable for deployment on resource-constrained edge devices. These advancements facilitate broader applications of Transformer models in complex scenarios, emphasizing the importance of effective token management strategies [15, 18, 53, 19, 45]. By integrating these approaches, Transformer models achieve improved adaptability and computational efficiency, essential for modern AI systems.

3.4 Adaptive and Semantic Token Pruning

Adaptive and semantic token pruning techniques enhance Transformer model efficiency by dynamically adjusting token processing based on semantic relevance. These methods optimize computational resources while ensuring high model performance across various applications, including advanced

text-to-image models and late-interaction retrieval systems. Effective pruning can significantly reduce model sizes like Stable Diffusion 2 without substantial quality loss, highlighting insights into information encoding. Analysis of late-interaction models emphasizes the importance of co-occurrence signals in improving retrieval effectiveness, while evaluations of prompt compression methods reveal that extractive techniques often outperform others in maintaining accuracy [54, 6, 16, 36].

The LazyLLM method exemplifies adaptive token pruning by deferring computation of less relevant tokens, enhancing efficiency without sacrificing accuracy [10]. GTP employs a graph-based propagation mechanism to reduce computational complexity while retaining crucial information from eliminated tokens, demonstrating effective adaptive strategies [55].

Semantic-based pruning techniques merge similar tokens based on similarity scores, effectively managing high-resolution images and reducing redundancy while preserving critical information. These methods showcase the efficacy of semantic token pruning in enhancing model efficiency by selectively retaining relevant tokens, thereby reducing storage overhead and computational demands while maintaining performance, as evidenced by studies on ColBERT and other late-interaction models [54, 18, 17].

Adaptive and semantic token pruning techniques are essential for enhancing Transformer model efficiency, intelligently reducing the number of tokens processed during inference. Methods like Learned Token Pruning (LTP) dynamically eliminate less important tokens based on learned attention scores, achieving significant reductions in computational load—up to 2.1× fewer floating-point operations (FLOPs) with minimal accuracy loss. Similarly, approaches like Zero-TPrune utilize the attention graph of pre-trained models to facilitate zero-shot token pruning, enabling a 34.7

4 Applications in Vision-Language Models

Token compression techniques are pivotal in enhancing the efficiency and performance of vision-language models (VLMs), particularly as the demand for advanced models increases. This section explores methodologies that significantly improve video understanding through strategic token compression.

4.1 Token Compression in Video Understanding

Token compression significantly enhances video understanding by improving computational efficiency while retaining essential spatial-temporal information. The CogVideoX model exemplifies this by generating coherent narratives and preserving motion semantics [7]. Similarly, the Speech2Video model demonstrates the potential of cross-modal token management in video generation from audio inputs [46].

The VisToG framework, although primarily tested on static images, suggests that intelligent visual token grouping can enhance efficiency in dynamic video contexts [12]. Cropr, effective in tasks like semantic segmentation and object detection, highlights the benefits of token pruning in Vision Transformers (ViTs) [49]. Additionally, intermediate fusion mechanisms align visual concepts with high-level semantics, enhancing generation quality and computational efficiency [34].

FastV employs dynamic token pruning by learning adaptive attention patterns, selectively pruning visual tokens to optimize video understanding [26]. The STOIC model showcases efficient resource usage in image synthesis, applicable to video synthesis tasks [28].

Recent advancements in token compression, such as Motion Guided Token Compression (MGTC) and Progressive Visual Token Compression (PVC), reduce computational burdens while maintaining performance and perceptual quality. These methods strategically select and compress tokens based on informational value, addressing redundancy in video data and optimizing processing capabilities. This enhances accuracy and responsiveness across applications, from real-time video analysis to complex multimodal tasks [56, 33, 18, 57, 58].

Figure 4 illustrates key models, techniques, and advanced methods in token compression for video understanding, highlighting their contributions to computational efficiency and video representation quality.

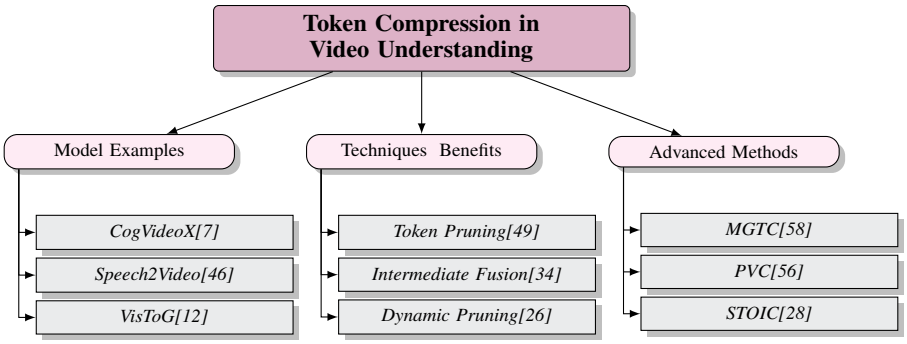

Figure 4: This figure illustrates key models, techniques, and advanced methods in token compression for video understanding, highlighting their contributions to computational efficiency and video representation quality.

4.2 Efficiency in Image Generation

Token compression is crucial for optimizing image generation by reducing computational demands while maintaining high image quality. Wavelet-based approaches generate high-resolution images with fewer tokens, preserving fidelity [59]. Token Merging techniques enhance processing speed by consolidating similar tokens, maintaining image integrity with minimal computational overhead [60].

The FQGAN model achieves state-of-the-art performance in discrete image reconstruction, demonstrating significant efficiency gains [61]. PixArt- leverages token compression to produce high-fidelity images at 4K resolution, showcasing scalability in handling high-resolution outputs with reduced computational requirements [62].

Token Fusion techniques improve accuracy and speed in image generation tasks [63]. The DiffiT model sets a new state-of-the-art FID score on the ImageNet-256 dataset with fewer parameters, highlighting the efficiency of token compression strategies [64].

The Swin Transformer architecture enhances semantic understanding and image fidelity through advanced token management [65]. The Diffusion Mamba (DiM) architecture evaluates performance and efficiency in generating high-quality images and videos, illustrating the impact of efficient token management [25].

Moreover, the APTP approach adapts model resources to varying prompt complexities, improving performance metrics while reducing computational costs compared to static methods [66]. The LIT architecture shows improved performance on image recognition tasks over existing hierarchical vision transformers, significantly reducing computational complexity and memory usage [20].

Token compression techniques transform image generation efficiency, enabling the production of coherent and contextually relevant images from textual input while optimizing computational resources. Innovations such as text information-guided dynamic visual token recovery and token pruning strategies demonstrate that intelligently reducing visual tokens can maintain model performance while achieving substantial reductions in computational load. This is crucial as models like Stable Diffusion grow in size and complexity, impacting accessibility on resource-constrained devices. Effective token management enhances processing speed and facilitates the deployment of advanced image generation systems across platforms [6, 16, 67, 18, 57].

4.3 Vision-Language Model Optimization

Token compression is critical for optimizing vision-language models (VLMs) by reducing computational complexity while maintaining high performance across diverse tasks. These techniques enhance inference speed and memory efficiency, essential for deploying VLMs in real-world applications. The BLIP-2 model utilizes Cross-Attention Token Pruning (CATP) to optimize tasks such as image captioning and visual question answering (VQA), demonstrating enhanced efficiency and performance [40].

The CAIT method maintains favorable transferability for downstream tasks like semantic segmentation, highlighting its optimization capabilities for VLMs [11]. Focal Pruning (FoPru) effectively reduces the number of visual tokens while preserving accuracy, leading to significant improvements in inference efficiency [13].

Recent advancements in token compression, including token pruning, merging, and recovery mechanisms, underscore their role in enhancing VLMs. These methods optimize computational efficiency and memory usage while maintaining or improving performance across various applications. For example, integrating token pruning with fuzzy logic can reduce memory costs by 39

5 Acceleration of Diffusion Transformers

Enhancing the efficiency of diffusion transformers requires exploring methods that improve performance without necessitating additional training. This section explores training-free inference acceleration techniques, which are essential for optimizing the operational capabilities of these models. By utilizing existing model architectures and resources, significant improvements in inference speed and computational efficiency can be realized. Techniques such as SpAtten and the -DiT method illustrate the potential of training-free approaches in diffusion transformers.

5.1 Training-Free Inference Acceleration Methods

Training-free inference acceleration methods are crucial for improving the efficiency of diffusion transformers by minimizing computational overhead without further training. SpAtten enhances inference speed by streamlining the attention mechanism, reducing DRAM access and computation [1]. These techniques optimize existing model components for faster and more efficient inference.

The -DiT method utilizes a caching mechanism to improve the efficiency of DiT blocks, significantly enhancing generation speed and quality without retraining. This approach underscores the advantages of cache-based strategies in managing large model parameters and high computational demands, as evidenced by recent advances in prompt compression and adaptive pruning techniques that optimize memory efficiency and reduce latency [68, 16, 69]. Similarly, FORA accelerates DiT models by caching and reusing intermediate outputs from attention and MLP layers across denoising steps, effectively lowering computational demands.

Innovations like the Skip-DiT framework enhance training-free acceleration by incorporating skip branches into conventional DiT models. This modification enables a caching mechanism, Skip-Cache, facilitating efficient feature reuse across timesteps during inference. The implementation of skip branches smooths feature transitions between DiT blocks, achieving a 1.5× speedup with minimal impact on output quality. This advancement addresses computational complexities in the sequential denoising process, making it more viable for real-time applications in image and video generation [70, 1, 71, 72, 73].

DiTFastAttn enhances attention computation efficiency by identifying and leveraging redundancies: spatial, temporal, and conditional. Techniques like Window Attention with Residual Sharing and Attention Sharing across Timesteps significantly reduce computational demands, achieving up to a 76% reduction in attention FLOPs and a 1.8× increase in processing speed for high-resolution image generation tasks [1, 74, 18, 17]. FlexDiT complements this by dynamically adapting token density, maintaining high-quality generation while ensuring computational efficiency.

The summarization compression method effectively reduces token size while preserving accuracy, showcasing its efficacy in accelerating inference without additional training. The Pruning All-Rounder (PAR) method offers a training-free approach that optimizes the pruning of tokens and layers in Large Vision-Language Models (LVLMs) through a meta-router for dynamic pruning flow management, enhancing inference efficiency without extensive training. Additionally, the FitPrune method exemplifies rapid visual token pruning, achieving substantial reductions in computational complexity with minimal accuracy loss [67, 75, 68, 50].

The VATP approach consistently outperforms attention-score-only methods by emphasizing value vector norms in KV cache reduction, thereby enhancing inference speed. The ToP method reduces computational costs through innovative training-free strategies, such as the -Cache mechanism, which optimizes DiT performance during image generation by caching and reusing intermediate outputs

across denoising steps, enhancing real-time application viability while maintaining quality metrics like IS Score and FID. Experimental results indicate that ToP can achieve up to a 1.6× speedup in generation time, underscoring its effectiveness in generative modeling applications [70, 73].

As illustrated in Figure 5, the key methods for training-free inference acceleration in diffusion transformers are categorized into attention mechanisms, cache mechanisms, and pruning strategies, highlighting specific techniques like SpAtten, -DiT, and PAR. This figure succinctly encapsulates the diverse strategies employed to enhance the efficiency of these models.

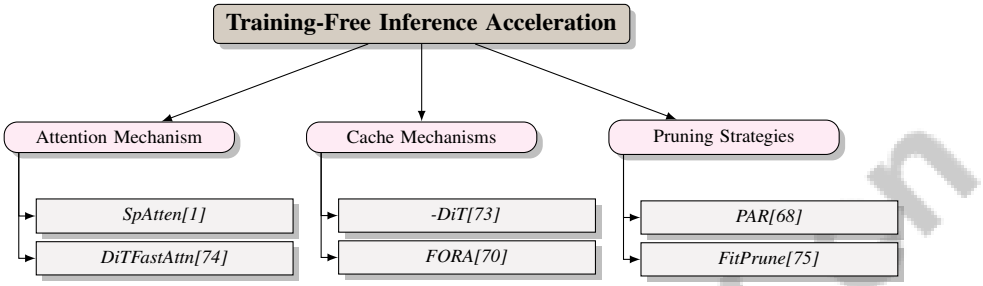

Figure 5: This figure illustrates the key methods for training-free inference acceleration in diffusion transformers, categorized into attention mechanisms, cache mechanisms, and pruning strategies, highlighting specific techniques like SpAtten, -DiT, and PAR.

5.2 Caching and Reuse Strategies

Caching and reuse strategies are pivotal for enhancing the efficiency of diffusion transformers by significantly reducing redundant computations and optimizing resource utilization. The Fast-FORward CAching (FORA) method accelerates diffusion transformers by caching and reusing intermediate outputs from attention and MLP layers throughout the denoising process. This approach minimizes computational overhead and integrates seamlessly with existing models without retraining. Techniques such as pruning and post-training quantization further optimize these models, allowing substantial reductions in size and computational requirements while maintaining output quality. These advancements make diffusion transformers more viable for real-time applications, addressing challenges posed by increasing complexity and resource demands [76, 6, 70].

FORA exemplifies the utility of caching strategies by accelerating the sampling process in diffusion transformers through the caching of features from computationally intensive layers for reuse across multiple time steps [70]. Similarly, Skip-DiT introduces skip connections that facilitate feature flow across layers, enhancing information preservation and allowing efficient caching across timesteps [72].

DiTFastAttn further demonstrates the effectiveness of caching and reuse strategies by caching and reusing outputs from previous computations, significantly reducing the need for redundant calculations and optimizing diffusion transformer efficiency [74]. By leveraging these strategies, diffusion transformers achieve enhanced computational efficiency, enabling faster and more resource-efficient processing across various applications.

6 Challenges and Future Directions

The advancement of AI, particularly through Transformer models, presents challenges that necessitate a nuanced understanding of their operational intricacies and performance constraints. This section delves into the critical balance between efficiency and semantic richness, especially in multimodal applications, and examines strategies that navigate these challenges to advance robust AI systems.

6.1 Balancing Efficiency and Semantic Richness

Balancing efficiency and semantic richness in Transformer models is crucial, particularly in multimodal contexts where preserving high-quality semantic representation is vital. Techniques like Focal Pruning (FoPru) effectively reduce token processing while maintaining critical information, emphasizing selective token management to optimize computational efficiency without compromising semantic

integrity [13]. Maintaining spatial integrity during token pruning is essential for tasks reliant on complete spatial structures, such as image and video processing [11]. Moreover, methods like FastV highlight the limitations of relying solely on attention scores for token importance, underscoring the need for adaptive strategies that consider broader factors to prevent information loss [26].

In natural language processing, dynamic token management strategies, including token pruning and combining, enhance computational efficiency while improving model performance, evidenced by a 5

6.2 Adapting to Diverse Modalities and Data Types

Adapting token compression techniques to diverse modalities and data types is critical for developing efficient AI systems amid the increasing complexity of modern applications. The computational demands of multimodal generation models necessitate effective handling of varied data types for scalability and performance [6]. The Speech2Video method exemplifies the need for token compression strategies that disentangle and process features from multiple modalities, such as audio and video, to enhance cross-modal understanding [46]. Token pruning methods like CATP demonstrate varying effectiveness based on specific layers, highlighting the importance of adaptable techniques catering to different modalities [40].

Future research could expand models like Cropr to additional vision tasks and modalities, enhancing applicability [49]. Frameworks such as Turbo, optimized for various Vision-Language Model (VLM) architectures, are crucial for improving AI system performance across diverse multimodal environments [5]. Recent studies highlight the potential of flexible token compression techniques to significantly reduce computational costs while improving classification accuracy and F1 scores. For instance, extractive prompt compression can achieve up to a 10× reduction in context length with minimal accuracy loss, outperforming other techniques like token pruning [18, 16]. This adaptability is vital for managing the complexities of multimodal data, ensuring robust performance in increasingly sophisticated computational environments.

6.3 Optimizing Pruning and Merging Strategies

Optimizing token pruning and merging strategies is essential for enhancing the efficiency and scalability of Transformer models, particularly in resource-intensive applications. Future research could explore architectural optimizations, including neural architecture search (NAS) and enhancements to MLP blocks for better local pattern encoding, as seen in vision-based models [20]. Refining dynamic token pruning parameters, as demonstrated by LazyLLM, could significantly enhance performance across model architectures [10]. Further research on Compressed Inference Networks (CINs) could focus on improving accuracy while maintaining efficiency, alongside exploring various applications [77]. Optimizing pruning rates and extending token pruning strategies to different architectures may improve adaptability and effectiveness [2].

Promising avenues for future research include exploring IdleViT applications in other domains and optimizing token selection processes, which could yield substantial efficiency improvements [78]. Addressing the limitations of current sparse attention mechanisms and refining pruning strategies, as highlighted in SpAtten, could lead to more robust and efficient models [1]. Recent advancements underscore the necessity for ongoing enhancement of token pruning and merging techniques to meet the increasing complexity and performance demands of Transformer models. Strategies like Learned Token Pruning (LTP) and PITOME demonstrate significant efficiency improvements by adaptively removing less important tokens and merging similar representations, respectively. Moreover, Zero-TPrune introduces a zero-shot approach that leverages attention graphs of pre-trained models for effective token pruning without the computational costs of fine-tuning, ensuring that Transformer architectures remain competitive in resource-constrained environments [53, 15, 45].

6.4 Overcoming Data and Resource Limitations

Addressing data and resource limitations is critical for deploying Transformer models in large-scale applications with substantial computational demands. The complexity of visual token redundancy in intricate visual inputs necessitates innovative strategies to streamline processing without compromising performance [50]. While token sparsification techniques improve efficiency, they may lead to sub-optimal boundary predictions in segmentation tasks, highlighting the need for refined

pruning strategies that uphold accuracy [79]. Efficient prompt design strategies are essential for reducing computational overhead and optimizing resource utilization, leveraging advanced techniques to maintain high performance across diverse tasks [36].

Exploring training-free methods and dynamic token management strategies can alleviate the computational burden of large Transformer models. By dynamically optimizing token density through pruning and combining, along with caching and reuse techniques, models can enhance inference speed and reduce resource consumption. For example, methods like DyCoke achieve an 18x speedup with minimal accuracy degradation, underscoring the effectiveness of these optimizations for real-time applications [80, 18, 81]. Enhancing the accessibility of large foundation models through democratized AI technologies is imperative to mitigate the high inference costs and latency associated with deploying models with hundreds of billions of parameters. This approach fosters innovative applications across sectors like customer service and content generation, ensuring broader access to these powerful tools [6, 82, 36, 69]. Developing open-source solutions and refining model architectures are crucial for ensuring efficiency and accessibility, optimizing Transformer models for deployment in diverse environments.

6.5 Enhancing Scalability and Generalizability

Developing scalable and generalizable token compression techniques is vital for optimizing Transformer models across diverse applications, particularly in multi-modal generative AI and complex vision tasks. The CAIT method exemplifies scalability by optimizing compression techniques across various transformer architectures, enhancing generalizability in practical applications [11]. The VisToG framework minimizes image token redundancies, benefiting larger Multi-modal Large Language Models (MLLMs) [12]. The TORE method highlights the potential for application in complex tasks, emphasizing the need for scalable techniques [9]. Future research could refine filtering and correlation stages of the FiCoCo method to enhance scalability and applicability across domains [3]. Integrating scalable token compression methods, as demonstrated by FastV, into various multimodal models could yield significant performance improvements [26].

The Speech2Video framework illustrates potential enhancements in video generation fidelity for unobserved speakers, underscoring the necessity for scalable token compression techniques in cross-modal applications [46]. The adaptability of proposed methods to various MLLMs and benchmarks further emphasizes the importance of developing techniques that integrate seamlessly with existing architectures [83]. Moreover, the compatibility of scalable token compression techniques with existing architectures, as shown by the multi-scale token merging method, underscores the importance of generalizability in enhancing performance across multiple datasets [21]. The Focal Pruning (FoPru) approach reinforces the need for scalable techniques applicable to various Large Vision-Language Models (LVLMs) [13].

These advancements highlight the critical role of developing scalable and generalizable token compression techniques in optimizing Transformer models for a wide range of applications. By ensuring robust performance and adaptability, these techniques facilitate the deployment of efficient AI systems in complex computational environments, making them suitable for on-device applications [28].

7 Conclusion

Token compression techniques play a pivotal role in enhancing the efficiency and scalability of Transformer models across various domains. These techniques, such as adaptive token pruning and efficient token merging, significantly reduce computational demands while maintaining high performance levels. In applications like text-to-video generation, these methods enable state-of-the-art results, underscoring their importance in optimizing complex generative tasks. The acceleration of Vision Transformers through minimal performance trade-offs further illustrates the broad applicability and effectiveness of token compression strategies. These advancements not only streamline model size but also preserve output quality, particularly in text-to-image models, facilitating their deployment on resource-constrained platforms. By improving the computational efficiency of Transformer models, token compression techniques are instrumental in advancing AI capabilities, supporting the implementation of scalable and high-performing solutions across diverse applications.

References

[1] Hanrui Wang, Zhekai Zhang, and Song Han. Spatten: Efficient sparse attention architecture with cascade token and head pruning, 2024.

[2] Yifei Liu, Mathias Gehrig, Nico Messikommer, Marco Cannici, and Davide Scaramuzza. Revisiting token pruning for object detection and instance segmentation, 2023.

[3] Yuhang Han, Xuyang Liu, Pengxiang Ding, Donglin Wang, Honggang Chen, Qingsen Yan, and Siteng Huang. Rethinking token reduction in mllms: Towards a unified paradigm for training-free acceleration, 2024.

[4] Ke Wang and Hong Xuan. Llava-zip: Adaptive visual token compression with intrinsic image information, 2024.

[5] 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 *European Conference on Computer Vision*, pages 436–455. Springer, 2024.

[6] Samarth N Ramesh and Zhixue Zhao. Efficient pruning of text-to-image models: Insights from pruning stable diffusion, 2024.

[7] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihan Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer, 2024.

[8] Jing Wang, Ao Ma, Jiasong Feng, Dawei Leng, Yuhui Yin, and Xiaodan Liang. Qihoo-t2x: An efficient proxy-tokenized diffusion transformer for text-to-any-task, 2024.

[9] Zhiyang Dou, Qingxuan Wu, Cheng Lin, Zeyu Cao, Qiangqiang Wu, Weilin Wan, Taku Komura, and Wenping Wang. Tore: Token reduction for efficient human mesh recovery with transformer, 2023.

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

[11] Ao Wang, Hui Chen, Zijia Lin, Sicheng Zhao, Jungong Han, and Guiguang Ding. Cait: Triple-win compression towards high accuracy, fast inference, and favorable transferability for vits, 2023.

[12] Minbin Huang, Runhui Huang, Han Shi, Yimeng Chen, Chuanyang Zheng, Xiangguo Sun, Xin Jiang, Zhenguo Li, and Hong Cheng. Efficient multi-modal large language models via visual token grouping, 2024.

[13] Lei Jiang, Weizhe Huang, Tongxuan Liu, Yuting Zeng, Jing Li, Lechao Cheng, and Xiaohua Xu. Fopru: Focal pruning for efficient large vision-language models, 2024.

[14] Lijun Yu. Towards multi-task multi-modal models: A video generative perspective, 2024.

[15] Hoai-Chau Tran, Duy M. H. Nguyen, Duy M. Nguyen, Trung-Tin Nguyen, Ngan Le, Pengtao Xie, Daniel Sonntag, James Y. Zou, Binh T. Nguyen, and Mathias Niepert. Accelerating transformers with spectrum-preserving token merging, 2024.

[16] Siddharth Jha, Lutfi Eren Erdogan, Sehoon Kim, Kurt Keutzer, and Amir Gholami. Characterizing prompt compression methods for long context inference, 2024.

[17] Carlos Lassance, Maroua Maachou, Joohee Park, and Stéphane Clinchant. A study on token pruning for colbert, 2021.

[18] Jungmin Yun, Mihyeon Kim, and Youngbin Kim. Focus on the core: Efficient attention via pruned token compression for document classification, 2024.

[19] Joakim Bruslund Haurum, Sergio Escalera, Graham W. Taylor, and Thomas B. Moeslund. Which tokens to use? investigating token reduction in vision transformers, 2023.

[20] Zizheng Pan, Bohan Zhuang, Haoyu He, Jing Liu, and Jianfei Cai. Less is more: Pay less attention in vision transformers, 2021.

[21] Zhe Bian, Zhe Wang, Wenqiang Han, and Kangping Wang. Multi-scale and token mergence: Make your vit more efficient, 2023.

[22] Zhenglun Kong, Peiyan Dong, Xiaolong Ma, Xin Meng, Mengshu Sun, Wei Niu, Xuan Shen, Geng Yuan, Bin Ren, Minghai Qin, Hao Tang, and Yanzhi Wang. Spvit: Enabling faster vision transformers via soft token pruning, 2022.

[23] Kaixin Xu, Zhe Wang, Chunyun Chen, Xue Geng, Jie Lin, Mohamed M. Sabry Aly, Xulei Yang, Min Wu, Xiaoli Li, and Weisi Lin. Lpvit: Low-power semi-structured pruning for vision transformers, 2024.

[24] Xiangcheng Liu, Tianyi Wu, and Guodong Guo. Adaptive sparse vit: Towards learnable adaptive token pruning by fully exploiting self-attention, 2023.

[25] Shentong Mo and Yapeng Tian. Scaling diffusion mamba with bidirectional ssms for efficient image and video generation, 2024.

[26] 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, 2024.

[27] Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. Planting a seed of vision in large language model, 2023.

[28] Sanchar Palit, Sathya Veera Reddy Dendi, Mallikarjuna Talluri, and Raj Narayana Gadde. Scalable, tokenization-free diffusion model architectures with efficient initial convolution and fixed-size reusable structures for on-device image generation, 2024.

[29] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications, 2020.

[30] Sihao Lin, Pumeng Lyu, Dongrui Liu, Tao Tang, Xiaodan Liang, Andy Song, and Xiaojun Chang. Mlp can be a good transformer learner, 2024.

[31] Kilian Carolan, Laura Fennelly, and Alan F. Smeaton. A review of multi-modal large language and vision models, 2024.

[32] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation, 2024.

[33] Xinyue Hao, Gen Li, Shreyank N Gowda, Robert B Fisher, Jonathan Huang, Anurag Arnab, and Laura Sevilla-Lara. Principles of visual tokens for efficient video understanding, 2024.

[34] Zizhao Hu, Shaochong Jia, and Mohammad Rostami. An intermediate fusion vit enables efficient text-image alignment in diffusion models, 2024.

[35] Hong Chen, Xin Wang, Yuwei Zhou, Bin Huang, Yipeng Zhang, Wei Feng, Houlun Chen, Zeyang Zhang, Siao Tang, and Wenwu Zhu. Multi-modal generative ai: Multi-modal llm, diffusion and beyond, 2024.

[36] Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. Unleashing the potential of prompt engineering in large language models: a comprehensive review, 2024.

[37] Chen Ju, Haicheng Wang, Zeqian Li, Xu Chen, Zhonghua Zhai, Weilin Huang, and Shuai Xiao. Turbo: informativity-driven acceleration plug-in for vision-language models. *arXiv preprint arXiv:2312.07408*, 2023.

[38] Narges Norouzi, Svetlana Orlova, Daan de Geus, and Gijs Dubbelman. Algm: Adaptive local-then-global token merging for efficient semantic segmentation with plain vision transformers, 2024.

[39] Jinyi Hu, Shengding Hu, Yuxuan Song, Yufei Huang, Mingxuan Wang, Hao Zhou, Zhiyuan Liu, Wei-Ying Ma, and Maosong Sun. Acdit: Interpolating autoregressive conditional modeling and diffusion transformer, 2024.

[40] Ruqi Liao, Chuqing Zhao, Jin Li, and Weiqi Feng. Catp: Cross-attention token pruning for accuracy preserved multimodal model inference, 2024.

[41] Bingqi Ma, Zhuofan Zong, Guanglu Song, Hongsheng Li, and Yu Liu. Exploring the role of large language models in prompt encoding for diffusion models, 2024.

[42] Shoufa Chen, Mengmeng Xu, Jiawei Ren, Yuren Cong, Sen He, Yanping Xie, Animesh Sinha, Ping Luo, Tao Xiang, and Juan-Manuel Perez-Rua. Gentron: Diffusion transformers for image and video generation, 2024.

[43] Yizhe Xiong, Hui Chen, Tianxiang Hao, Zijia Lin, Jungong Han, Yuesong Zhang, Guoxin Wang, Yongjun Bao, and Guiguang Ding. Pyra: Parallel yielding re-activation for training-inference efficient task adaptation. In *European Conference on Computer Vision*, pages 455–473. Springer, 2024.

[44] Nick John Eliopoulos, Purvish Jajal, James C. Davis, Gaowen Liu, George K. Thiravathukal, and Yung-Hsiang Lu. Pruning one more token is enough: Leveraging latency-workload non-linearities for vision transformers on the edge, 2024.

[45] Sehoon Kim, Sheng Shen, David Thorsley, Amir Gholami, Woosuk Kwon, Joseph Hassoun, and Kurt Keutzer. Learned token pruning for transformers, 2022.

[46] Shijing Si, Jianzong Wang, Xiaoyang Qu, Ning Cheng, Wenqi Wei, Xinghua Zhu, and Jing Xiao. Speech2video: Cross-modal distillation for speech to video generation, 2021.

[47] Yi-Lin Sung, Jaehong Yoon, and Mohit Bansal. Ecoflap: Efficient coarse-to-fine layer-wise pruning for vision-language models, 2024.

[48] Junyan Li, Li Lyna Zhang, Jiahang Xu, Yujing Wang, Shaoguang Yan, Yunqing Xia, Yuqing Yang, Ting Cao, Hao Sun, Weiwei Deng, Qi Zhang, and Mao Yang. Constraint-aware and ranking-distilled token pruning for efficient transformer inference, 2023.

[49] Benjamin Bergner, Christoph Lippert, and Aravindh Mahendran. Token cropr: Faster vits for quite a few tasks, 2024.

[50] Qizhe Zhang, Aosong Cheng, Ming Lu, Zhiyong Zhuo, Minqi Wang, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. [cls] attention is all you need for training-free visual token pruning: Make vlm inference faster, 2024.

[51] Yifan Xu, Zhijie Zhang, Mengdan Zhang, Kekai Sheng, Ke Li, Weiming Dong, Liqing Zhang, Changsheng Xu, and Xing Sun. Evo-vit: Slow-fast token evolution for dynamic vision transformer, 2021.

[52] Xubing Ye, Yukang Gan, Yixiao Ge, Xiao-Ping Zhang, and Yansong Tang. Atp-llava: Adaptive token pruning for large vision language models, 2024.

[53] Hongjie Wang, Bhishma Dedhia, and Niraj K. Jha. Zero-tprune: Zero-shot token pruning through leveraging of the attention graph in pre-trained transformers, 2024.

[54] Qi Liu, Gang Guo, Jiaxin Mao, Zhicheng Dou, Ji-Rong Wen, Hao Jiang, Xinyu Zhang, and Zhao Cao. An analysis on matching mechanisms and token pruning for late-interaction models, 2024.

[55] Xuwei Xu, Sen Wang, Yudong Chen, Yanping Zheng, Zhewei Wei, and Jiajun Liu. Gtp-vit: Efficient vision transformers via graph-based token propagation, 2024.

[56] Chenyu Yang, Xuan Dong, Xizhou Zhu, Weijie Su, Jiahao Wang, Hao Tian, Zhe Chen, Wenhai Wang, Lewei Lu, and Jifeng Dai. Pvc: Progressive visual token compression for unified image and video processing in large vision-language models, 2024.

[57] Yi Chen, Jian Xu, Xu-Yao Zhang, Wen-Zhuo Liu, Yang-Yang Liu, and Cheng-Lin Liu. Recoverable compression: A multimodal vision token recovery mechanism guided by text information, 2024.

[58] Yukun Feng, Yangming Shi, Fengze Liu, and Tan Yan. Motion guided token compression for efficient masked video modeling, 2024.

[59] Wael Mattar, Idan Levy, Nir Sharon, and Shai Dekel. Wavelets are all you need for autoregressive image generation, 2024.

[60] Daniel Bolya and Judy Hoffman. Token merging for fast stable diffusion, 2023.

[61] Zechen Bai, Jianxiong Gao, Ziteng Gao, Pichao Wang, Zheng Zhang, Tong He, and Mike Zheng Shou. Factorized visual tokenization and generation, 2024.

[62] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ : $Weak-to-strongtrainingofdiffusiontransformerfor4ktext-to-imagegeneration$, 2024.

[63] Minchul Kim, Shangqian Gao, Yen-Chang Hsu, Yilin Shen, and Hongxia Jin. Token fusion: Bridging the gap between token pruning and token merging, 2023.

[64] Ali Hatamizadeh, Jiaming Song, Guilin Liu, Jan Kautz, and Arash Vahdat. Diffit: Diffusion vision transformers for image generation, 2024.

[65] Ruijun Li, Weihua Li, Yi Yang, Hanyu Wei, Jianhua Jiang, and Quan Bai. Swinv2-imagen: Hierarchical vision transformer diffusion models for text-to-image generation, 2022.

[66] Alireza Ganjdanesh, Reza Shirkavand, Shangqian Gao, and Heng Huang. Not all prompts are made equal: Prompt-based pruning of text-to-image diffusion models, 2025.

[67] Mingxing Rao, Bohan Jiang, and Daniel Moyer. Training noise token pruning, 2024.

[68] Wei Suo, Ji Ma, Mengyang Sun, Lin Yuanbo Wu, Peng Wang, and Yanning Zhang. Pruning all-rounder: Rethinking and improving inference efficiency for large vision language models, 2024.

[69] Youngsuk Park, Kailash Budhathoki, Liangfu Chen, Jonas Kübler, Jiaji Huang, Matthäus Kleindessner, Jun Huan, Volkan Cevher, Yida Wang, and George Karypis. Inference optimization of foundation models on ai accelerators, 2024.

[70] Pratheba Selvaraju, Tianyu Ding, Tianyi Chen, Ilya Zharkov, and Luming Liang. Fora: Fast-forward caching in diffusion transformer acceleration, 2024.

[71] Zheng Zhan, Yushu Wu, Yifan Gong, Zichong Meng, Zhenglun Kong, Changdi Yang, Geng Yuan, Pu Zhao, Wei Niu, and Yanzhi Wang. Fast and memory-efficient video diffusion using streamlined inference, 2024.

[72] Guanjie Chen, Xinyu Zhao, Yucheng Zhou, Tianlong Chen, and Cheng Yu. Accelerating vision diffusion transformers with skip branches, 2024.

[73] Pengtao Chen, Mingzhu Shen, Peng Ye, Jianjian Cao, Chongjun Tu, Christos-Savvas Bouganis, Yiren Zhao, and Tao Chen. δ-dit: A training-free acceleration method tailored for diffusion transformers, 2024.

[74] Zhihang Yuan, Hanling Zhang, Pu Lu, Xuefei Ning, Linfeng Zhang, Tianchen Zhao, Shengen Yan, Guohao Dai, and Yu Wang. Ditfastattn: Attention compression for diffusion transformer models, 2024.

[75] Weihao Ye, Qiong Wu, Wenhao Lin, and Yiyi Zhou. Fit and prune: Fast and training-free visual token pruning for multi-modal large language models, 2024.

[76] Yuewei Yang, Jialiang Wang, Xiaoliang Dai, Peizhao Zhang, and Hongbo Zhang. An analysis on quantizing diffusion transformers, 2024.

[77] Lukas Hedegaard. Efficient online processing with deep neural networks. *arXiv preprint arXiv:2306.13474*, 2023.

[78] Xuwei Xu, Changlin Li, Yudong Chen, Xiaojun Chang, Jiajun Liu, and Sen Wang. No token left behind: Efficient vision transformer via dynamic token idling, 2023.

[79] Lei Zhou, Huidong Liu, Joseph Bae, Junjun He, Dimitris Samaras, and Prateek Prasanna. Token sparsification for faster medical image segmentation, 2023.

[80] Wenxi Tan. Infor-coef: Information bottleneck-based dynamic token downsampling for compact and efficient language model, 2023.

[81] Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Dycoke: Dynamic compression of tokens for fast video large language models, 2024.

[82] Mengwei Xu, Wangsong Yin, Dongqi Cai, Rongjie Yi, Daliang Xu, Qipeng Wang, Bingyang Wu, Yihao Zhao, Chen Yang, Shihe Wang, et al. A survey of resource-efficient llm and multimodal foundation models. *arXiv preprint arXiv:2401.08092*, 2024.

[83] Shiyu Zhao, Zhenting Wang, Felix Juefei-Xu, Xide Xia, Miao Liu, Xiaofang Wang, Mingfu Liang, Ning Zhang, Dimitris N. Metaxas, and Licheng Yu. Accelerating multimodal large language models by searching optimal vision token reduction, 2024.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.