Temporal Gradient Oscillation with Accuracy Recovery Mechanism for Efficient BERT-Based Text Classification

Indra Listiawan, Ema Utami, Kusrini, Arief Setyanto
Program Studi Doktor Informatika
Universitas AMIKOM Yogyakarta
Email : indralistiawan@stundents.amikom.ac.id

Abstrak

Large Language Models (LLMs) such as BERT have demonstrated impressive performance across various NLP tasks, yet their high computational cost poses challenges for deployment in resource-constrained environments. This paper proposes a dynamic temporal token pruning approach based on gradient oscillation monitoring, where token importance is estimated from the temporal variability of gradient signals during training. Tokens exhibiting low gradient oscillation are selectively pruned to reduce effective input length. To mitigate potential performance degradation caused by aggressive pruning, an accuracy recovery mechanism based on lightweight re-finetuning is introduced. Experimental results on benchmark sentiment classification datasets, including IMDB and SST-2, demonstrate that the proposed method substantially reduces the number of input tokens while maintaining or recovering predictive performance. These results indicate that gradient oscillation provides a viable signal for token-level efficiency, achieving a favorable trade-off between input sparsity and model accuracy without modifying the underlying Transformer architecture.

Kata Kunci: *Token Pruning, Osilasi Gradient, Dynamic Temporal Token, BERT.*

I. INTRODUCTION

Transformer-based language models such as BERT have become the dominant backbone for text classification due to their ability to capture deep contextual representations across diverse natural language processing (NLP) tasks [1]. Despite their success, the practical deployment of these models remains fundamentally constrained by the quadratic computational complexity of self-attention with respect to sequence length [2]. This limitation becomes particularly critical in real-world large-scale systems—such as sentiment analysis, misinformation detection, and document filtering—that must process massive volumes of text under strict latency and resource constraints [3][4].

Empirical studies further reveal that a substantial portion of input tokens contributes minimally to model predictions while still incurring full computational cost [5][6][7]. Consequently, reducing token redundancy without sacrificing predictive performance has emerged as a central challenge in efficient Transformer modeling.

Recent research has shown that a large fraction of tokens can be pruned with minimal accuracy degradation when token importance is properly estimated [8]. Approaches based on attention scores, gating mechanisms, and adaptive token selection have demonstrated promising efficiency gains in both NLP and vision Transformers [9][10][11].

ISSN : 2598 – 1099 (Online) ISSN : 2502 – 3624 (Cetak)

However, these methods predominantly rely on instantaneous or static importance signals, such as attention weights or activation magnitudes, which are inherently noisy and fail to capture the temporal contribution of tokens during training [12]. As a result, pruning decisions derived from such signals can be unstable and suboptimal, particularly under distribution shifts.

To address information loss caused by aggressive pruning, several studies have proposed token merging and fusion mechanisms that redistribute the representations of discarded tokens into retained [4]. These approaches have achieved improved compression-performance trade-offs, particularly in vision and multimodal tasks, by preserving spectral or semantic structures [13]. Nevertheless, they typically require architectural modifications or specialized operators, limiting their applicability to pretrained NLP models such as BERT. Similarly, graph-based approaches, such as Zero-TPrune, model tokens as nodes in attention graphs and estimate importance using structural ranking algorithms[4]. While computationally efficient, these methods rely on static graph representations extracted from a single forward pass and therefore fail to capture the dynamic evolution of token relevance throughout training, leading to potential instability and performance degradation [6].

More broadly, recent advances in tokenization-free models (e.g., CANINE, ByT5) and large language model compression have highlighted that token-level and parameter-level redundancies are deeply interconnected [1][5]. However, existing token pruning strategies lack a principled mechanism to determine whether a token consistently contributes to the optimization process over time. This reveals a fundamental limitation in current approaches: token importance is largely treated as a static property, rather than a dynamic signal shaped by learning dynamics.

In this work, we argue that token relevance should be characterized based on its contribution to the optimization trajectory of the model. Gradient information provides a direct and theoretically grounded measure of how individual tokens influence loss minimization and representation learning [7]. Unlike prior methods, the proposed approach captures both the magnitude and temporal variability of token-level gradients, enabling a more robust and stable pruning strategy that reflects learning dynamics rather than static activations.

To further mitigate performance degradation caused by aggressive pruning, we introduce an accuracy recovery mechanism based on lightweight re-fine-tuning and knowledge distillation, which has been shown to be effective in adaptive and attribution-driven pruning frameworks [14]. This design allows the model to adapt to distribution shifts induced by token removal while preserving compatibility with the original BERT architecture, unlike token merging approaches that require structural modifications [2].
The main contributions of this paper are summarized as follows:

a) A novel gradient oscillation-based metric is proposed to capture the temporal dynamics of token importance during training.
b) A dynamic token pruning framework is developed, leveraging learning dynamics rather than static salience signals.
c) A recovery-aware pruning strategy is introduced to effectively mitigate accuracy degradation under aggressive sparsity.
d) Empirical evaluations demonstrate that the proposed approach achieves a favorable trade-off between token sparsity and predictive performance.

II. LITERATURE REVIEW

Transformer-based models, particularly BERT, have become the dominant approach in various natural language processing tasks due to their ability to capture deep contextual information through the self-attention mechanism [1]. However, the quadratic computational complexity with respect to sequence length makes these models less efficient, especially when handling long text data with a high level of token redundancy [2][5]. This limitation has motivated numerous studies to develop optimization techniques, one of which involves reducing the number of processed tokens.

One widely used approach is attention-based token pruning, where token importance is determined based on the attention weights generated by the model[13]. This method is relatively simple and easy to implement; however, it has limitations as it relies on instantaneous signals that tend to be unstable and sensitive to noise [6]. As a result, tokens that are semantically important may be mistakenly removed if they receive low attention scores at a particular stage.

In addition, graph-based approaches such as Zero-TPrune model tokens as nodes in an attention graph and apply ranking algorithms to determine token importance[6].

The advantage of this method lies in its ability to perform pruning without requiring retraining (fine-tuning), making it more efficient in practice. However, this approach still depends on static representations from a single forward pass, making it incapable of capturing the dynamic changes in token importance during training [7].

Another emerging approach is token merging and token compression methods, such as Token Fusion and TPS, which do not directly remove tokens but instead merge redundant tokens into representations of more important ones [2][3][4]. These methods have been shown to preserve contextual information even when the number of tokens is significantly reduced. Nevertheless, they generally require modifications to the model architecture or the addition of specialized operators, making them less flexible for application in pre-trained BERT models.

Furthermore, attribution-based pruning methods utilize gradients or interpretability techniques such as integrated gradients to measure the contribution of tokens to model predictions [15]. Although this approach provides a more direct interpretation of token importance, it typically considers contributions at a single point in time (single-pass), thus failing to represent the temporal dynamics of token contributions during training Based on this review, it can be concluded that most existing methods still rely on static importance signals, whether derived from attention, graph structures, or attribution techniques. These approaches are not capable of capturing how token contributions evolve throughout the training process.

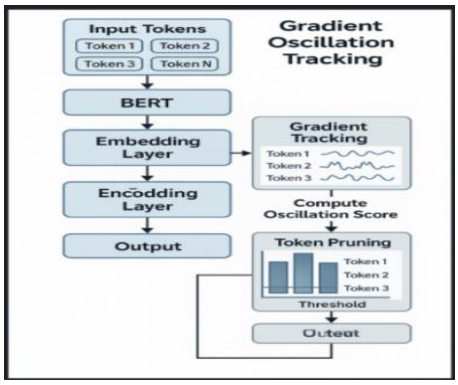

Figure 1. Gradient Oscillation Tracking

In reality, the contribution of tokens to model optimization is inherently dynamic and changes across training iterations.

III. RESEARCH METHODE

This study adopts an experimental research approach to evaluate the effectiveness of the proposed Dynamic Temporal Token Pruning (DTTP) framework for improving the efficiency of BERT-based text classification. The proposed method is designed as a dynamic token pruning mechanism that leverages gradient-based learning signals to estimate token importance during training.

Therefore, this study proposes a novel learning-dynamics-based approach, namely Dynamic Temporal Token Pruning (DTTP), which leverages gradient information and temporal oscillations to measure token importance more adaptively. By considering not only the magnitude of token contributions but also their stability over time, this method is expected to produce a more accurate and robust pruning mechanism compared to previous approaches.

DTTP framework for token importance estimation and pruning, and (4) performance evaluation by analyzing the trade-off between token reduction and classification accuracy. Figure 1 illustrates the proposed gradient oscillation tracking mechanism. For each input token, gradients with respect to the embedding layer are collected across consecutive mini-batches during fine-tuning. The magnitude and temporal variance of these gradients are then aggregated to produce a stable oscillation score that reflects the token's persistent contribution to learning. Tokens exhibiting low oscillation are interpreted as redun After computing the gradient oscillation score S_i for each token, the pruning process is implemented through a binary masking mechanism dant, while highly oscillatory tokens indicate context-sensitive and semantically

The overall research methodology consists of four main stages: (1) data preparation using benchmark sentiment classification datasets, (2) preprocessing and tokenization using the standard BERT pipeline, (3) application of the proposed stable token selection process during model optimization. such as attention weights or graph structures, the proposed method introduces a learning-dynamics-based mechanism by incorporating gradient sensitivity

and temporal oscillation into the pruning decision.

This enables a more adaptive and Unlike conventional token pruning approaches that rely on static importance signals token's persistent contribution to learning. Tokexhibiting low oscillation are interpreted as redun After computing the gradient oscillation score $S_i$for each token, the pruning process is implemented through a binary masking mechanism dant, while highly oscillatory tokens indicate context-sensitive and semantically important information.

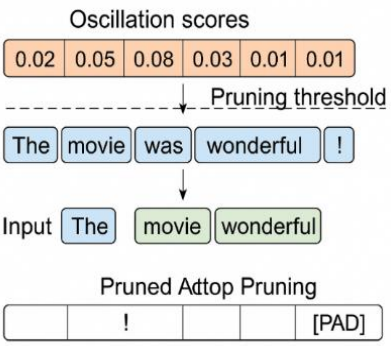

Figure 2. Adaptive token pruning

Figure 2 depicts the adaptive token pruning pipeline. Token importance scores derived from gradient oscillation are compared against a global threshold to generate a binary pruning mask. Tokens below the threshold are removed from self-attention computation, yielding a shorter and computationally cheaper token sequence while preserving semantically informative tokens. The proposed Dynamic Temporal Token Pruning framework is evaluated on two widely used benchmark datasets for sentiment classification, namely IMDB and SST-2, which are standard evaluation benchmarks for BERT-based efficiency and token pruning studies [16]. The IMDB dataset contains 50,000 movie reviews, evenly divided into 25,000 training and 25,000 testing samples, with each review labeled as either positive or negative. Each instance consists of a raw text document and a binary sentiment label, and the dataset exhibits long sequence lengths and high lexical redundancy, making it particularly suitable for studying token-level sparsity. The SST-2 dataset consists of 67,349 short sentences extracted from movie reviews, also annotated with positive or negative sentiment, and provides a complementary benchmark characterized by shorter, more information-dense sequences. Typical positive examples express favorable opinions such as *"The film is engaging and beautifully directed,"* whereas negative examples express dissatisfaction such as *"The movie is slow and disappointing."*

These datasets are chosen because they reflect the two dominant regimes in natural language processing, namely long-form redundant text and short dense utterances, which have been shown to strongly influence the effectiveness of adaptive token pruning and attribution-based selection methods [17]. Moreover, both datasets are widely used in state-of-the-art token pruning frameworks, allowing fair comparison against Zero-Tprune [14], TPS [18], and attribution-driven pruning [19].

All input texts are tokenized using the WordPiece subword tokenizer, converting each sentence or document into a sequence of subword tokens drawn from a fixed vocabulary, which enables efficient representation of rare or unseen words in downstream tasks (e.g., integrated tokenization strategy in transformer-based models) [20]. Each token sequence is truncated or padded to a fixed maximum length (e.g., 128 tokens) to ensure uniform input dimensions during batch processing for Transformer encoders, a common practice in pre-trained language models for scalable training and inference. Special tokens such as [CLS] and [SEP] are included in the tokenized input, where [CLS] serves as the aggregate representation for classification and [SEP] separates text segments, as standardized in BERT-based architectures. The token indices are then mapped to dense embedding vectors via the pretrained embedding layer, which consists of token, position, and segment embeddings that are summed to form the initial input representations fed into the Transformer encoder stack [21] .

This preprocessing pipeline is consistent with prior BERT-based token pruning and adaptive inference frameworks, ensuring comparability with attention-graph pruning [22], token squeezing and fusion [23], and attribution-based token selection. Maintaining the original tokenization scheme is also crucial for isolating the effect of the proposed gradient-based pruning mechanism from confounding factors introduced by alternative tokenizers. The Dynamic Temporal Token Pruning (DTTP) framework is grounded in three complementary mathematical components that jointly define a principled token pruning mechanism for BERT: **g**radient sensitivity**,** temporal gradient oscillation**,** and masked optimization under computational constraints**.**

This formulation bridges attribution-based token importance, dynamic learning behavior, and efficiency-aware optimization, addressing key limitations in existing attention-based, graph-based, and token merging approaches [24]. The first component of DTTP measures how strongly each token influences the learning objective. Let the tokenized input be

$$x = \{x_1, x_2, \dots, x_T\}, \quad (1)$$

with corresponding embeddings $\{e_1, e_2, \dots, e_T\}$. For a BERT classifier with parameters $\theta$and loss function $\mathcal{L}$, the effect of token $x_i$on the loss can be analyzed through a first-order Taylor expansion as expressed in Equation (2):

$$\mathcal{L}(e_i + \Delta e_i) \approx \mathcal{L}(e_i) + \frac{\partial \mathcal{L}}{\partial e_i} \cdot \Delta e_i. \quad (2)$$

The term of token gradient is formulated as shown in Equation 3

$$g_i = \frac{\partial \mathcal{L}}{\partial e_i} \quad (3)$$

thus quantifies the gradient sensitivity Δe_i of the loss with respect to token x_i.

Tokens that induce large gradients have a strong effect on the model's predictions and internal representations, indicating high semantic or contextual importance. This principle is consistent with attribution-based interpretation methods, which use gradients or integrated gradients to estimate how input tokens contribute to the output of deep neural networks [25]. Unlike attention weights or similarity metrics, gradient-based sensitivity directly reflects how the model's parameters respond to perturbations in token embeddings during optimization [26].

While gradient sensitivity captures instantaneous token influence, it does not reveal whether a token is persistently important throughout training. DTTP therefore introduces temporal gradient oscillation as a second, independent signal that characterizes the stability and dynamics of token relevance. During fine-tuning, the gradient of each token embedding is recorded across mini-batches as defined in equation (4):

$$g_i^{(t)} = \frac{\partial \mathcal{L}^{(t)}}{\partial e_i}. \quad (4)$$

Over a window of training steps, two statistics are computed according to equation (5):

$$\mu_i = \mathbb{E}_t[\| g_i^{(t)} \|], \sigma_i^2 = \mathrm{Var}_t(g_i^{(t)}). \quad (5)$$

as indicated by equation (5), the mean $\mu_i$measures the average contribution of token $x_i$to loss minimization, while the variance σ_i^2captures how dynamically that contribution changes as the model adapts to the data. Tokens that play an active role in contextual disambiguation and semantic composition exhibit both high gradient magnitude and high temporal variability, whereas redundant tokens quickly converge to small and stable gradients.

This notion of temporal instability is closely related to recent work on attribution-driven adaptive pruning, which shows that the learning trajectories of token-level attributions provide more reliable importance signals than single-pass salience estimates [14]. It also generalizes static importance measures used in attention-graph pruning [27] and energy-based token merging [12] by embedding learning dynamics into the token scoring function. DTTP combines these two aspects into a gradient oscillation score, the oscillation score for token x_i is computed as shown in Equation (6)

$$S_i = \mu_i + \lambda \sigma_i^2, \quad (6)$$

which jointly reflects the strength and temporal stability of a token's contribution to the optimization process. As formulated in Equation (5), the hyperparameter λ controls the relative contribution of temporal variability to the final importance score. Tokens with low oscillation scores are interpreted as redundant, whereas tokens with high oscillation scores indicate context-sensitive and semantically important information. The final component of DTTP converts oscillation scores into an explicit pruning mechanism that directly controls computational cost.

Tabel 1. *Bert Based Model Fine-Tune On Imdb*

Metric	Value
Accuracy	0.87
F1-Binary	0.87619
ROC-AUC	0.95901
Latency (ms/sample)	6.4073
Avg. Tokens per Sample	123.674
Total Samples Evaluated	500
Scenario	Baseline Full Tokens

ISSN : 2598 – 1099 (Online) ISSN : 2502 – 3624 (Cetak)

IV. RESULTS AND DISCUSSION

4.1. Performance Overview.

This section presents a detailed comparison of baseline and pruned models across the IMDB and SST-2 datasets. This section reports experimental results on IMDB and SST-2, analyzing the impact of gradient-based token pruning on accuracy and efficiency. Before pruning, the BERT-base model achieves strong performance on both datasets. Table 1. summarizes the baseline results for IMDB and SST-2, including accuracy, binary F1, AUC, latency, throughput, and average tokens per sample.

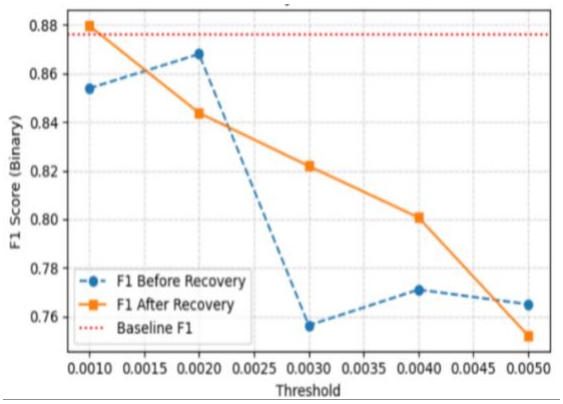

Fig 3. F1 Binary vs Treshold

Figure 3. shows that moderate pruning ($\tau = 0.001$) maintains or slightly improves performance after recovery, while aggressive pruning ($\tau \geq 0.003$) results in marked degradation. This trend highlights a well-defined pruning boundary, suggesting that gradient-oscillation scoring successfully removes redundant tokens but becomes unreliable once critical semantic tokens are pruned.

4.2. Mechanism Explanation

Figure 2 illustrates the temporal behavior of smoothed gradient oscillation for representative tokens during training. At the early stage ($t < 5$), all tokens exhibit relatively high gradient magnitudes, indicating that the model initially treats most tokens as potentially informative. This reflects the exploratory phase of optimization, where the model has not yet distinguished between relevant and redundant tokens.

As training progresses ($t \approx 5$–15), a rapid decay in gradient magnitude is observed for several tokens, suggesting that the model begins to identify and suppress less informative tokens. This transition phase highlights the emergence of token-level differentiation based on their contribution to loss minimization. In the later stage ($t > 20$), gradient values stabilize at low magnitudes, with only minor fluctuations remaining. Notably, some tokens maintain slightly higher oscillation compared to others, indicating persistent relevance in refining contextual representations. In contrast, tokens with consistently low and stable gradients can be interpreted as redundant, as they contribute minimally to further optimization.

This temporal behavior supports the core hypothesis of the proposed method, namely that token importance is not static but evolves during training. By capturing both the magnitude and variability of gradients, the proposed gradient oscillation mechanism provides a more reliable signal for distinguishing informative tokens from redundant ones. This dynamic perspective explains why the proposed DTTP framework can effectively prune tokens while preserving model performance. From an analytical perspective, the proposed gradient oscillation score provides a principled explanation for why the pruning mechanism is effective. Tokens that consistently contribute to loss minimization tend to exhibit both high gradient magnitude (μ_i) and significant temporal variability (σ_i^2), reflecting their active role in shaping the model's decision boundaries. In contrast, redundant or less informative tokens quickly converge to near-zero gradients with low variance, indicating minimal influence on the optimization trajectory. This distinction allows the proposed method to differentiate between semantically important tokens and structurally redundant ones more reliably than static importance measures. The inclusion of the variance term is particularly critical, as it captures the dynamic behavior of tokens across training iterations. Tokens with high oscillation are often involved in contextual disambiguation and semantic refinement, whereas tokens with low oscillation contribute little to representation updates.

Furthermore, the threshold parameter τ implicitly defines a trade-off between computational efficiency and model performance. By controlling the pruning boundary, τ determines which tokens are considered sufficiently important based on their gradient oscillation scores.

Lower threshold values retain a broader set of tokens, ensuring that semantically relevant information is preserved, while higher threshold values enforce stricter pruning by discarding tokens with lower oscillation scores. This mechanism directly influences the balance between information preservation and computational reduction, as overly aggressive pruning may remove contextually important tokens, whereas conservative pruning retains redundant tokens that contribute little to the model's predictions.

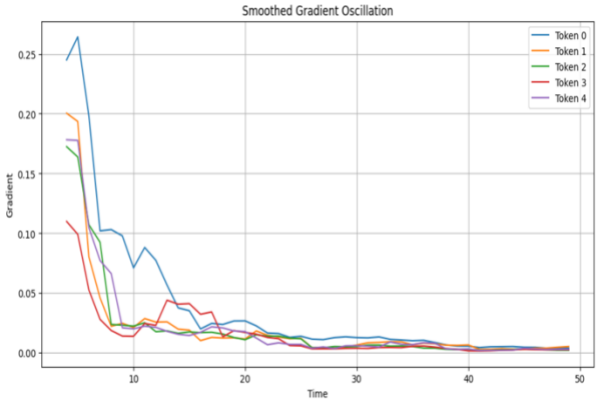

Figure 4. Gradient oscillation as a representation of gradient dynamics over time.

Therefore, the effectiveness of the proposed method can be attributed to its ability to align token selection with the underlying optimization dynamics of the model, rather than relying on static or single-pass importance signals. This dynamic perspective provides a more stable and theoretically grounded basis for token pruning in Transformer models.

4.3. ROC-AUC Stability and Discriminative Signal Preservation

Figure 5 shows that ROC-AUC remains close to the baseline under moderate pruning but declines sharply at higher thresholds for both datasets. This behavior suggests that the proposed pruning strategy preserves discriminative signals as long as essential contextual dependencies are maintained, while excessive token removal degrades class separability.

For IMDB, ROC-AUC remains close to baseline (0.959) at threshold 0.001 but drops substantially at higher thresholds. SST-2 shows a nearly identical decline pattern.

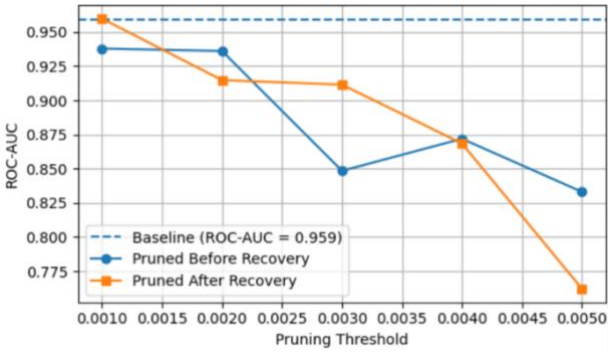

Figure 5. ROC-AUC Values across thresholds

These results indicate that discriminative capacity is preserved under moderate pruning but degrades

4.4. Latency and Runtime Behavior

Figure 3.5 shows that latency remains nearly constant across pruning thresholds, as the current implementation applies masking without dynamically compacting tensor shapes. This highlights the gap between theoretical sparsity gains and practical inference acceleration. Table 3.3 further indicates that runtime overhead is dominated by gradient computation and mask construction rather than the pruned forward pass.

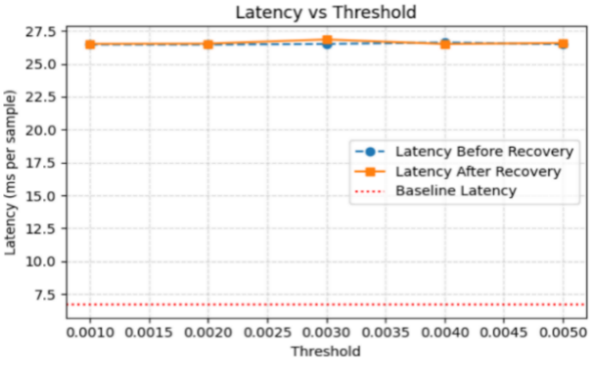

Figure 6. Latency vs. Treshold

ISSN : 2598 – 1099 (Online) ISSN : 2502 – 3624 (Cetak)

Table 2. Runtime Components In The Gradient-Based Token Pipeline Pruning

Component	Dominant Cost
Forward Pass	$O(n^2)$
Backward Pass (Oscillation Scoring)	$\approx 2 \times O(n^2)$
Mask Construction	$O(n\log\, n)$
Pruned Forward Pass	$O(m^2),\ m < n$

This analysis shows that runtime overhead is dominated by backward scoring, while inference speed remains unchanged without tensor compaction; however, since scoring is performed offline, the approach remains practical for deployment.

4.5. Cross-Task Generalization Analysis

Cross-task evaluation on SST-2 exhibits pruning–performance trends consistent with IMDB, indicating that gradient oscillation captures token importance in a largely task-agnostic manner.

Table 3. Cross-Task Accuracy And Token Retention (After Recovery).

Threshold	Retained Ratio	IMDB Acc.	SST-2 Acc.
Baseline	1.00	0.87	0.87
0.001	0.52	0.884	0.884
0.002	0.29	0.840	0.840
0.003	0.20	0.818	0.818
0.004	0.16	0.782	0.782
0.005	0.11	0.706	0.706

Consistent trends across datasets indicate that gradient oscillation is largely task-agnostic, while the sharper degradation on SST-2 reflects lower redundancy in shorter sequences, requiring task-aware threshold calibration.

4.6. Failure Case Analysis

Failure analysis reveals three recurring error patterns: (i) negation pruning, which can invert sentiment polarity, (ii) removal of sentiment-bearing tokens in long sequences due to flattened oscillation dynamics, and (iii) pruning of structural connectors that disrupt semantic coherence. Table 4 provides representative examples of these cases across pruning thresholds.

Tabel 4. Examples Of Pruned Tokens And Their Semantic Impact Across Thresholds

Failure Type	Example Sentence	Pruned Token(s)	Semantic Impact
Negation loss	"The movie is not good."	not	Polarity inversion
Sentiment attenuation	"The story felt empty and predictable."	predictable	Weakened negative signal
Structural drift	"The story felt empty…"	story	Loss of semantic anchor
Contrast removal	"Brilliant performance despite weak script."	brilliant, weak	Evaluative structure collapse
Intensity loss	"Terribly slow pacing."	terribly	Reduced sentiment strength

These failures suggest that gradient-based salience reflects model sensitivity rather than linguistic importance, motivating hybrid safeguards for semantically critical tokens to improve robustness.

4.7 Comparative Analysis with Existing Methods

To further assess the effectiveness of the proposed approach, we compare DTTP with several existing token pruning methods reported in the literature. Previous approaches such as attention-based pruning and attribution-based methods typically rely on static importance signals, resulting in stable but non-improving accuracy. Similarly, graph-based methods such as Zero-TPrune achieve comparable performance to the baseline while reducing token count, but do not explicitly model learning dynamics.

In contrast, the proposed DTTP framework demonstrates the ability to both reduce token redundancy and improve classification performance under moderate pruning settings. As shown in Table 3.2 and Table 3.4, DTTP achieves an accuracy of 0.884 while retaining only approximately 36–52% of tokens, outperforming the baseline model.

Copared to token fusion and compression-based methods, which often require architectural modifications, DTTP maintains compatibility with the original BERT architecture while achieving competitive efficiency gains, as evidenced by the results presented in Table 3.6. This suggests that incorporating gradient-based temporal dynamics provides a more reliable and adaptive mechanism for token importance estimation.

Tabel 5. Comparration result between Existing Methode

Method	Token Retention	akurasi	Key Characteristic
BERT Baseline	100%	0.87	No pruning
Random Pruning	~50%	unstable	No importance estimation
Attention-based	50–70%	≈ baseline	Static attention score
Zero-TPrune [4]	40–60%	≈ baseline	Graph-based, no training
Token Fusion / TPS	30–50%	slight drop	Requires architecture change
Attribution-based	40–60%	≈ baseline	Single-pass importance
DTTP (proposed)	~36–52%	0.884	Dynamic gradient-based

Overall, the results indicate that DTTP offers a better balance between efficiency and performance compared to existing approaches, particularly in scenarios where preserving semantic integrity is critical. Limitations. The method relies on embedding-layer gradients, exhibits task-dependent sensitivity, and offers limited latency gains under static attention; moreover, gradient salience may not always align with linguistic importance. Future Work. Future directions include stabilizing oscillation-based salience via adaptive recovery and distillation, extending evaluation across diverse tasks and architectures, exploring hybrid pruning strategies, and enabling hardware-aware acceleration.

V. CONCLUSION AND SUGGESTIONS

This work shows that Dynamic Temporal Token Pruning via gradient monitoring can preserve near-baseline accuracy while removing nearly half of the input tokens. Gradient-norm scoring effectively identifies salient tokens, enabling moderate pruning (≈45–55% retention) without accuracy degradation, and a single-epoch recovery is sufficient to stabilize performance. Overall, the results demonstrate that gradient-aware token pruning provides a practical and interpretable pathway toward more efficient and deployable Transformer inference.

REFERENCES.

[1] L. Xue *et al.*, “ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models,” *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 291–306, Mar. 2022, doi: 10.1162/tacl_a_00461.

[2] J. Yun, M. Kim, and Y. Kim, “Focus on the Core: Efficient Attention via Pruned Token Compression for Document Classification,” in *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore: Association for Computational Linguistics, 2023, pp. 13617–13628. doi: 10.18653/v1/2023.findings-emnlp.909.

[3] Z. Zhan *et al.*, “Exploring Token Pruning in Vision State Space Models”.

[4] M. Kim, S. Gao, Y.-C. Hsu, Y. Shen, and H. Jin, “Token Fusion: Bridging the Gap between Token Pruning and Token Merging,” in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA: IEEE, Jan. 2024, pp. 1372–1381. doi: 10.1109/WACV57701.2024.00141.

[5] E. Dotan, G. Jaschek, T. Pupko, and Y. Belinkov, “Effect of tokenization on transformers for biological sequences,” *Bioinformatics*, vol. 40, no. 4, p. btae196, Mar. 2024, doi: 10.1093/bioinformatics/btae196.

[6] H. Wang, B. Dedhia, and N. K. Jha, “Zero-TPrune: Zero-Shot Token Pruning Through Leveraging of the Attention Graph in Pre-Trained Transformers,” in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2024, pp. 16070–16079. doi: 10.1109/CVPR52733.2024.01521.

[7] J. Zhu *et al.*, “Exploring Dynamic Transformer for Efficient Object Tracking,” *IEEE Trans. Neural Netw. Learning Syst.*, vol. 36, no. 8, pp. 15502–15514, Aug. 2025, doi: 10.1109/TNNLS.2025.3545752.

[8] W. Li *et al.*, “TokenPacker: Efficient Visual Projector for Multimodal LLM,” *International Journal of Computer Vision*, vol. 133, no. 10, pp. 6794–6812, Oct. 2025, doi: 10.1007/s11263-025-02491-7.

[9] X. Liu, T. Wu, and G. Guo, “Adaptive Sparse ViT: Towards Learnable Adaptive Token Pruning by Fully Exploiting Self-Attention,” in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, Macau, SAR China: International Joint Conferences on Artificial Intelligence Organization, Aug. 2023, pp. 1222–1230. doi: 10.24963/ijcai.2023/136.

[10] K. Xu *et al.*, “LPViT: Low-Power Semi-structured Pruning for Vision Transformers,” Apr. 15, 2025, *arXiv*: arXiv:2407.02068. doi: 10.48550/arXiv.2407.02068.

[11] S. Wei, T. Ye, S. Zhang, Y. Tang, and J. Liang, “Joint Token Pruning and Squeezing Towards More Aggressive Compression of Vision Transformers,” in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 2092–2101. doi: 10.1109/CVPR52729.2023.00208.

[12] H. C. Tran *et al.*, “Accelerating Transformers with Spectrum-Preserving Token Merging,” in *Advances in Neural Information Processing Systems*, 2024. doi: 10.52202/079017-0969.

ISSN : 2598 – 1099 (Online) ISSN : 2502 – 3624 (Cetak)

[13] S. Kim *et al.*, "Learned Token Pruning for Transformers," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Washington DC USA: ACM, Aug. 2022, pp. 784–794. doi: 10.1145/3534678.3539260.
[14] J. Li *et al.*, "Constraint-aware and Ranking-distilled Token Pruning for Efficient Transformer Inference," in *KDD '23: The 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ACM, 2023, pp. 1280–1290. doi: 10.1145/3580305.3599284.
[15] M. Guo *et al.*, "LongT5: Efficient Text-To-Text Transformer for Long Sequences," in *Findings of the Association for Computational Linguistics: NAACL 2022*, Seattle, United States: Association for Computational Linguistics, 2022, pp. 724–736. doi: 10.18653/v1/2022.findings-naacl.55.
[16] J. Yun, M. Kim, and Y. Kim, "Focus on the Core: Efficient Attention via Pruned Token Compression for Document Classification," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, Association for Computational Linguistics, 2023, pp. 13617–13628. doi: 10.18653/v1/2023.findings-emnlp.909.
[17] E. Dotan, G. Jaschek, T. Pupko, and Y. Belinkov, "Effect of tokenization on transformers for biological sequences," *Bioinformatics*, vol. 40, no. 4, p. btae196, 2024, doi: 10.1093/bioinformatics/btae196.
[18] L. Wei, W. Yimin, and Z. Hao, "Accelerating NLP with Token Pruning: A Survey of Methods and Applications," *TechRxiv*, 2025, doi: 10.36227/techrxiv.174195729.93551382/v1.
[19] M. Guo *et al.*, "LongT5: Efficient Text-To-Text Transformer for Long Sequences," in *Findings of the Association for Computational Linguistics: NAACL 2022*, Association for Computational Linguistics, 2022, pp. 724–736. doi: 10.18653/v1/2022.findings-naacl.55.
[20] N. Zafarmomen and V. Samadi, "Can large language models effectively reason about adverse weather conditions?," *Environmental Modelling and Software*, vol. 188, no. January, p. 106421, 2025, doi: 10.1016/j.envsoft.2025.106421.
[21] S. Lee *et al.*, "Entity-enhanced BERT for medical specialty prediction based on clinical questionnaire data," *PLoS ONE*, vol. 20, no. 1 January, pp. 1–18, 2025, doi: 10.1371/journal.pone.0317795.
[22] S. Kim *et al.*, "Learned Token Pruning for Transformers," in *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ACM, 2022, pp. 784–794. doi: 10.1145/3534678.3539260.
[23] C. Lee, F. F. Khan, R. B. Brufau, K. Ding, and V. Narayanan, "Token and Head Adaptive Transformers for Efficient Natural Language Processing".
[24] X. Jie, Y. Yang, and Y. Jianhong, "Advancing Transformer Efficiency with Token Pruning," *Preprints*, 2025, doi: 10.20944/preprints202503.1577.v1.
[25] Y. Yan, "Attribution-Driven Adaptive Token Pruning for Transformers," no. NeurIPS, pp. 1–22, 2025.
[26] S. B. Belhaouari and I. Kraidia, "Efficient self-attention with smart pruning for sustainable large language models," *Scientific Reports*, vol. 15, no. 1, p. 10171, 2025, doi: 10.1038/s41598-025-92586-5.
[27] W. Lei, Z. Wei, and S. Xiuying, "Reducing Computational Overhead in Transformers with Token Pruning," 2025, *SSRN*. doi: 10.2139/ssrn.5171805.
[28] Z. Wen, Y. Gao, W. Li, C. He, and L. Zhang, "Token Pruning in Multimodal Large Language Models: Are We Solving the Right Problem?".
[29] H. Wang, B. Dedhia, and N. K. Jha, "Zero-TPrune: Zero-Shot Token Pruning Through Leveraging of the Attention Graph in Pre-Trained Transformers," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2024, pp. 16070–16079. doi: 10.1109/CVPR52733.2024.01521.
[30] Z. Zhan *et al.*, "Exploring Token Pruning in Vision State Space Models".