山东师范大学
SHANDONG NORMAL UNIVERSITY

Attribution-Driven Adaptive Token Pruning for Transformers

Yaoyao Yan; Yu Hui; Weizhi Xu

CONTENTS

01 Introduction

Background & Motivation

Transformers (e.g., GPT, BERT) dominate NLP due to strong modeling power.

Issue: Self-attention scales quadratically → heavy computation & latency. → e.g., 512 tokens → >1.5B FLOPs.

Need: Efficient compression without accuracy loss.

→ **Problems:**

Attention ≠ true token contribution
Static pruning ignores sequence complexity

Our Method: AD-TP

Attribution-Driven Token Importance via Integrated Gradients (IG) → more accurate token contribution.

Adaptive Token Retainer (ATR) = Adaptive Retention Ratio Predictor (ARP) → decide how many tokens to keep Token Saliency Predictor (TSP) → decide which tokens to keep

Key Contributions

- Dynamic pruning based on sequence complexity → major FLOPs reduction
- Attribution-based importance estimation (using IG)
- Dual-Normalized Distillation + self-supervision for better token-level reasoning

02 Motivation

Motivation

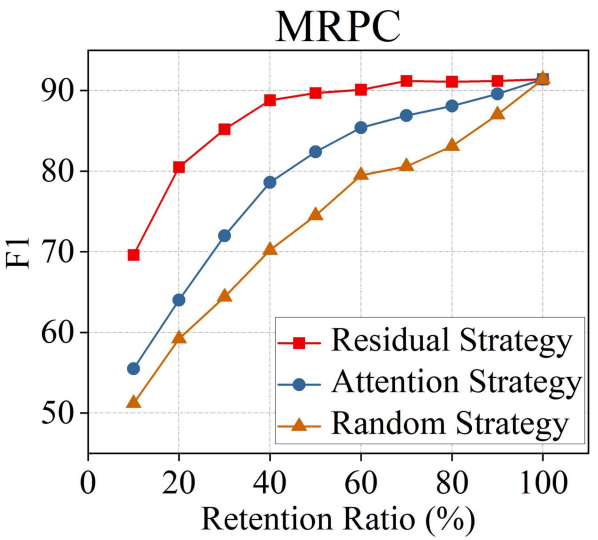
◇ 1. Token Importance Estimation Matters

Most pruning methods use attention scores to estimate token importance.

⚠ Problem: Attention focuses on token-to-token interactions, ignoring individual semantic contribution.

Comparison (SST-2, MRPC):
Random (lower bound)
Attention-based
Residual-based (upper bound)
→ Attention ≠ true contribution → large improvement margin remains.

☑Our solution: Use Integrated Gradients (IG) to compute attribution-based importance for each token → more accurate & interpretable selection.



Motivation

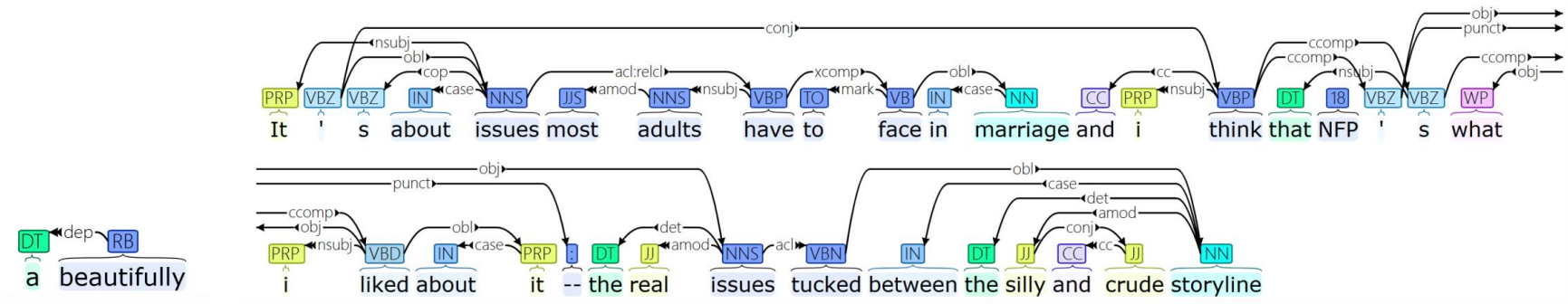
◇ 2. Limitations of Fixed or Learnable Thresholds

Existing frameworks adopt fixed or global thresholds, ignoring sequence-level diversity.

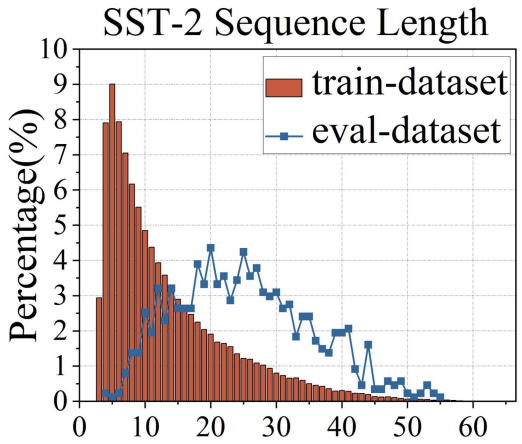
Even LTP learns thresholds per layer but keeps them static during inference.

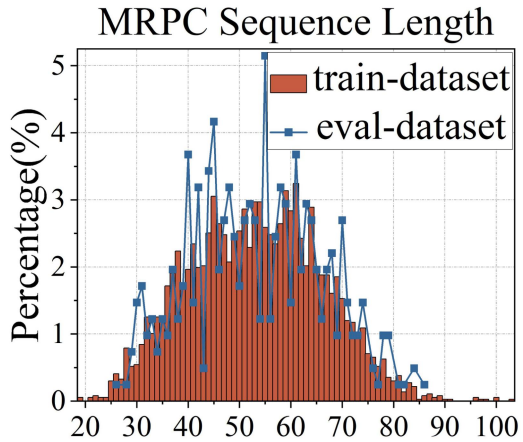
Observation: Sequence length and syntactic complexity vary widely across data.

☑Our approach: Introduce dynamic pruning — adjust retention ratio per sequence based on token saliency & structure complexity.



(a) (b)

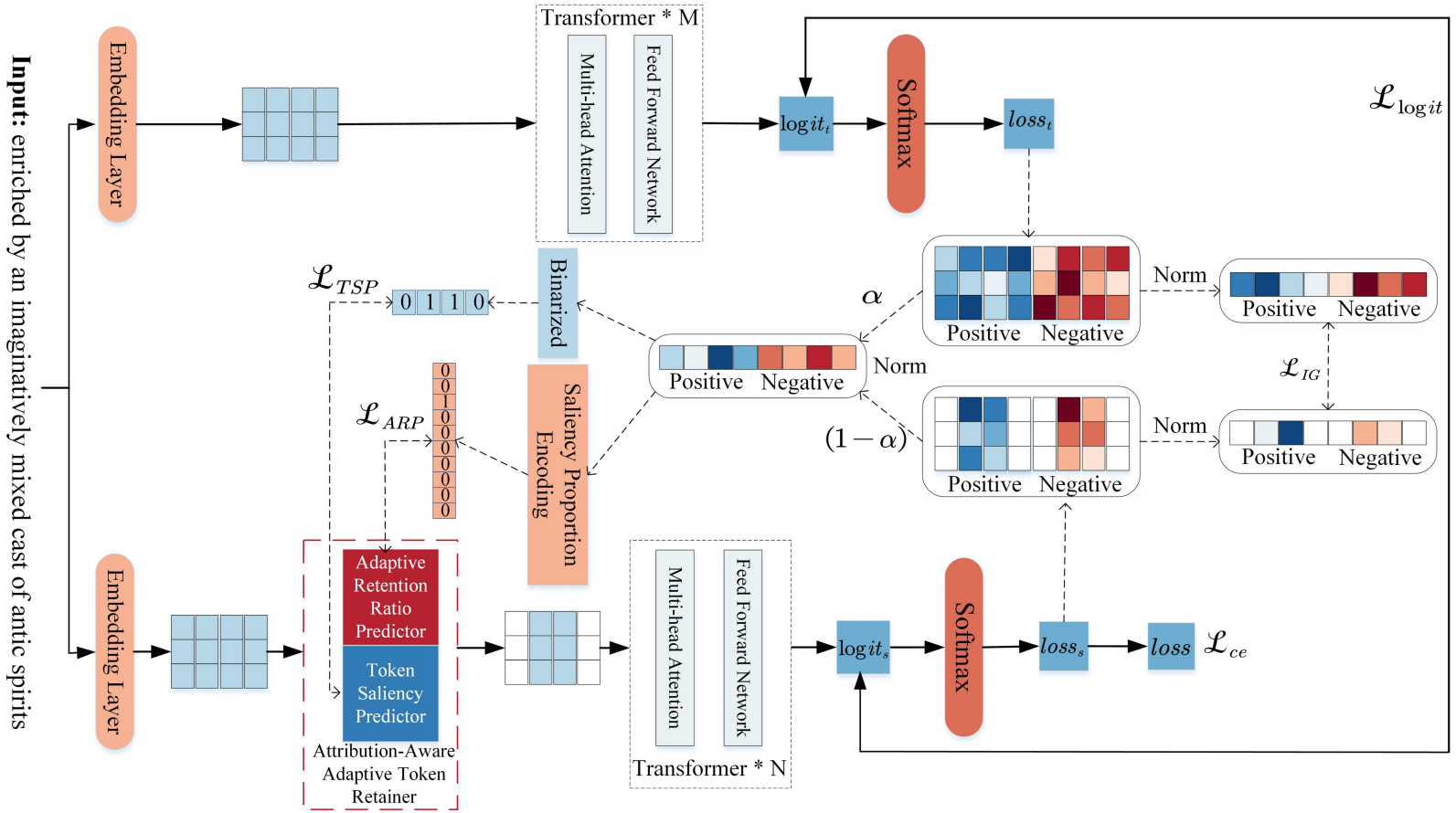




03 Method

Overview

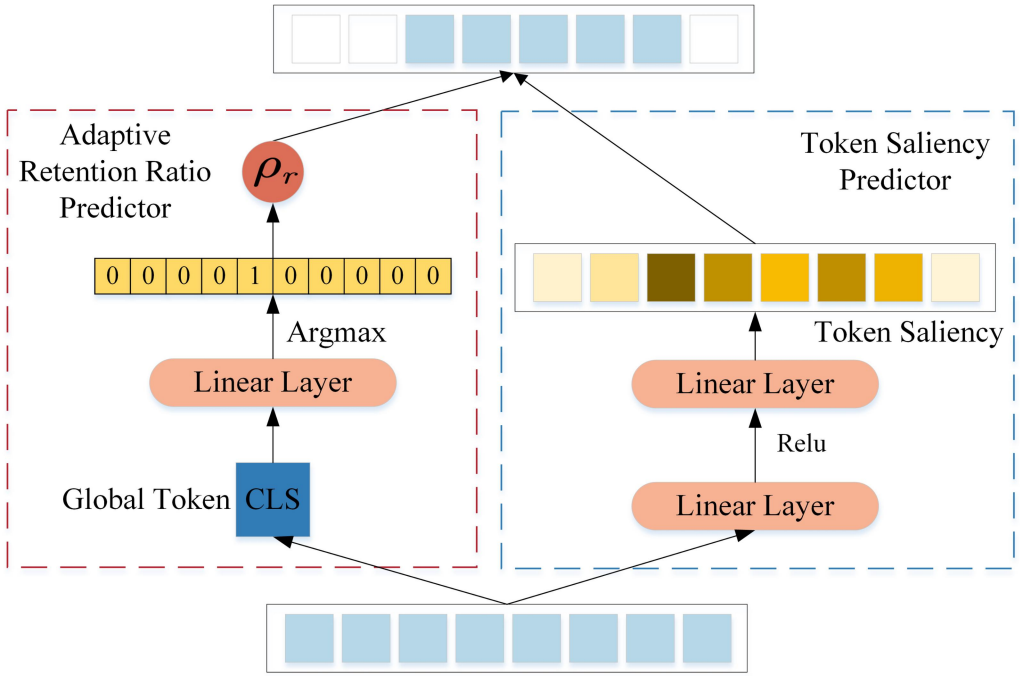
- AD-TP introduces an Attribution-Aware Adaptive Token Retainer (ATR)→ composed of ARP (Adaptive Retention Ratio Predictor) & TSP (Token Saliency Predictor).
- Goal: Efficiently prune tokens while maintaining model accuracy.
- Uses Integrated Gradients (IG) for token attribution + teacher & self-supervision for training.



Attribution-Aware Adaptive Token Retainer

ARP:
Uses [CLS] embedding → predicts sequence complexity → outputs retention ratio ρ_r.
Selects ratio from candidate list via linear + argmax function.

TSP:
Assigns saliency score S to each token through linear + ReLU layers.
Keeps top ρ_r% tokens → adaptive sparse sequence.



Integrated Gradients for Token Saliency

IG computes token attribution: $$IG_{ij} = (e_{ij} - e'_{ij}) \times \frac{1}{m} \sum_{k=1}^{m} \frac{\partial F^c(E' + \frac{k}{m}(E - E'))}{\partial e_{ij}}$$

Teacher keeps top-K IG dimensions (reduce noise), student uses all.

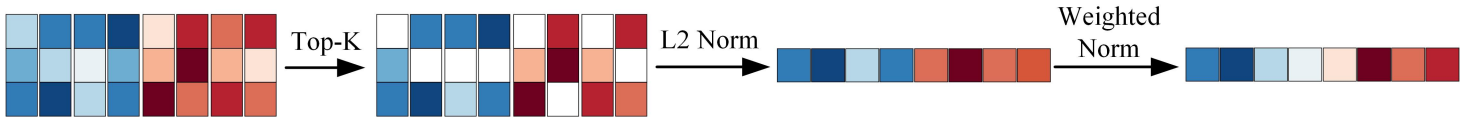
Computes token-level attribution magnitudes via L2 norm.

Joint Supervision (Teacher + Self)

Weighted fusion of teacher & student attribution: $$a_c^i = (1-k)a_{t,c}^i + ka_{s,c}^i$$

Gradually shifts from teacher-guided → self-supervised learning.

Dual normalization aligns scales (L2 + sequence-level).



Overall Objective:

$$\mathcal{L} = (1-\alpha)\mathcal{L}_{CE} + \alpha\mathcal{L}_{logit} + \beta\mathcal{L}_{IG} + \gamma(\mathcal{L}_{ARP} + \mathcal{L}_{TSP})$$

04 Experiments

Table 1: Comparison of AD-TP and mainstream pruning methods on GLUE in terms of accuracy and FLOPs. T and S denote the teacher and student models, respectively. All baseline results are from ToP [23], except BERT-base (T) and BERT6 (S). Bold and underlined values indicate the best and second-best results.

Model	Method	CoLA		RTE		QQP		MRPC		SST-2		MNLI		QNLI		STS-B	
		Matthews	FLOPs	Acc.	FLOPs	Acc.	FLOPs	F1	FLOPs	Acc.	FLOPs	Acc.	FLOPs	Acc.	FLOPs	Pearson	FLOPs
BERT-base (T)	-	60.3±0.7	1.00×	69.7±0.5	1.00×	91.5±0.9	1.00×	91.4±0.3	1.00×	93.7±0.5	1.00×	84.9±0.4	1.00×	91.7±0.5	1.00×	89.0±0.2	1.00×
BERT6 (S)	-	51.2±1.1	1.00×	66.1±0.4	1.00×	90.4±0.3	1.00×	89.2±0.8	1.00×	91.0±0.8	1.00×	81.7±0.9	1.00×	89.3±0.8	1.00×	87.8±0.5	1.00×
PoWER-BERT	Atten-value	52.3	4.50×	67.4	3.40×	90.2	4.50×	88.1	2.70×	92.1	2.40×	83.8	2.60×	90.1	2.00×	85.1	2.00×
LAT	Atten-value	-	-	-	-	-	-	-	-	92.8	2.90×	84.4	2.80×	-	-	-	-
LTP	Atten-value	52.3	8.66×	63.2	6.84×	90.4	7.44×	87.1	6.02×	92.3	3.59×	83.9	3.74×	89.3	3.91×	87.5	5.25×
ToP	Atten-value	**60.5**	9.62×	70.0	**7.10×**	91.2	**8.04×**	89.2	6.32×	93.5	3.82×	84.7	4.27×	**90.6**	4.35×	87.6	5.32×
AD-TP6	IG	55.3±0.4	13.78×±0.2	70.1±0.7	4.25×±0.4	90.4±0.5	7.05×±0.3	89.8±0.9	6.83×±0.3	92.0±0.7	7.98x±0.4	84.9±1.0	5.45x±0.7	89.5±0.5	4.17×±0.3	88.2±0.6	6.63×±0.7
AD-TP12	IG	56.9±0.6	**13.89×±0.4**	**71.2±0.7**	4.89×±0.7	**91.4±0.6**	7.85×±0.5	**90.8±0.6**	**6.97×±0.4**	**94.7±0.3**	**8.45×±0.6**	**85.4±0.4**	**5.69×±0.2**	90.2±0.6	**4.52×±0.4**	**88.7±0.4**	**6.66×±0.3**

Table 2: Comparison of Token Pruning Methods on Long-Sequence Tasks.

Model	Method	SQuADv2.0		20News	
		F1	FLOPs	Acc.	FLOPs
BERT-base	-	77.1±0.3	1.00x	86.7±0.7	1.00x
BERT6	-	73.8±0.8	1.00x	85.0±1.2	1.00x
Transkimmer	Prediction	75.7	4.67x	86.1	8.11x
PoWER-BERT	Attention	-	-	86.5	2.91x
LTP	Attention	75.6	3.10x	85.2	4.66x
ToP	Attention	75.9	4.12x	87.0	8.26x
AD-TP6	IG	**76.2±0.6**	**5.07x±0.2**	**87.3±0.5**	**8.74x±0.5**

Table 3: Comparison with Distillation and Structured Pruning.

Model	CoLA		MRPC		20News	
	Matt.	FLOPs	F1	FLOPs	Acc.	FLOPs
BERT6	51.2±1.1	1.00x	89.2±0.8	1.00x	85.0±1.2	1.00x
DistilBERT6	49.0	2.00x	86.9	2.00x	85.8	2.00x
CoFi6	38.0	9.10x	86.3	**7.70x**	85.9	5.90x
AD-TP6	**55.3±0.4**	**13.78x±0.2**	**89.8±0.9**	6.83x±0.3	**87.3±0.5**	**8.74x±0.5**
BERT-base12	60.3±0.7	1.00x	91.4±0.3	1.00x	86.7±0.7	1.00x
CoFi12	39.8	9.10x	90.0	4.00x	86.4	7.70x
AD-TP12	**56.9±0.6**	**13.89x±0.4**	**90.8±0.6**	**6.97x±0.4**	**88.6±0.3**	**8.82x±0.2**

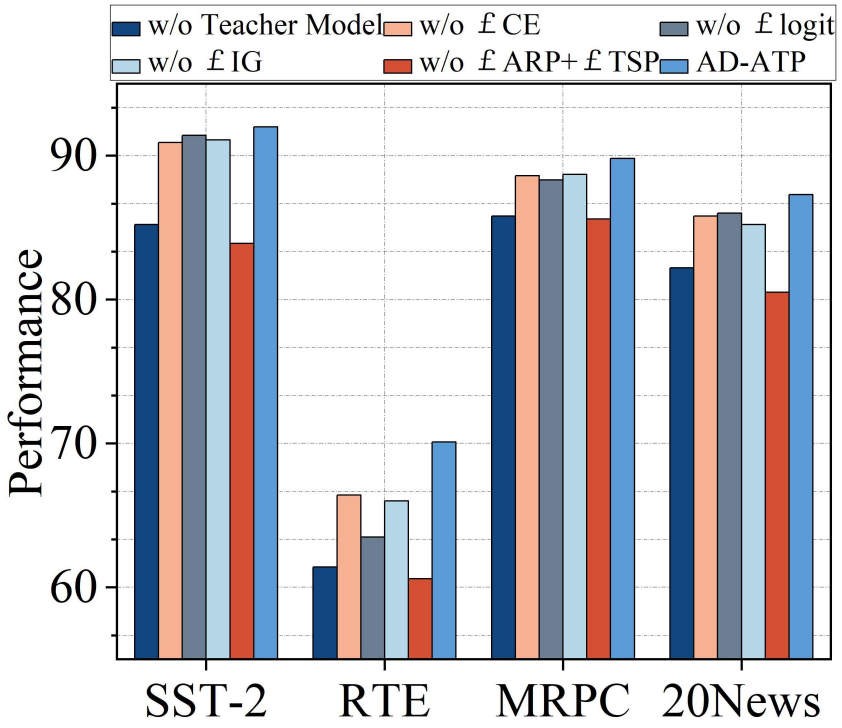
Ablation Study — Effect of Teacher Model & Adaptive Retainer

1. Impact of the Teacher Model and Loss Components

Five AD-TP variants are tested by removing:
(1) Teacher model (2) Cross-entropy loss (3) Logits-based loss (4) Attribution loss (5) Retainer loss

Results: Removing any component degrades performance. The absence of the teacher or retainer loss causes the largest drop (7–8 %).

Conclusion: Both the teacher model and the Attribution-Aware Adaptive Token Retainer are critical for effective pruning and accuracy preservation.



Ablation Study — Effect of Teacher Model & Adaptive Retainer

2. Impact of the Adaptive Retention Predictor (ARP)

- All methods share the same architecture; only retention-ratio selection differs.
- Adaptive strategy consistently achieves higher accuracy under similar FLOPs.
- Adjusting pruning strength based on sequence complexity yields better trade-offs.

Table 4: Ablation study of the adaptive token retainer.

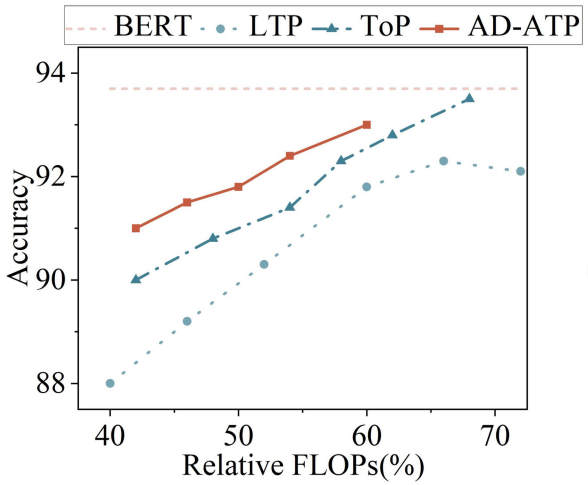
Model	SST-2		20News	
	Acc.	ρ	Acc.	ρ
Static	91.2	0.30	85.7	0.30
Random	90.5	0.30	82.5	0.30
Adaptive	**92.0**	0.29	**87.3**	0.27

Effectiveness Under Different FLOPs Budgets

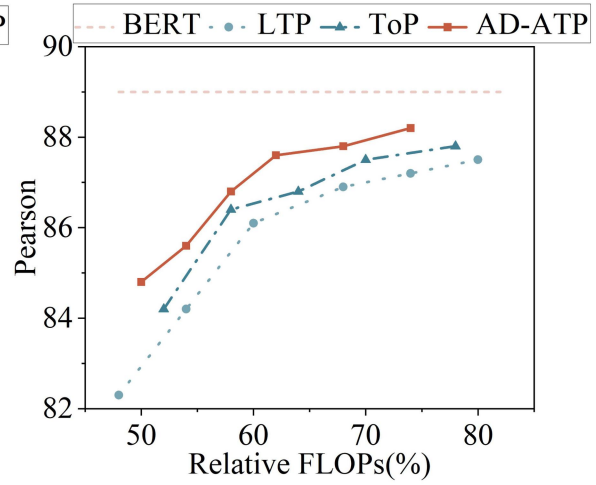
Evaluate AD-TP under various FLOPs budgets.

Compare with two pruning baselines: LTP and ToP.

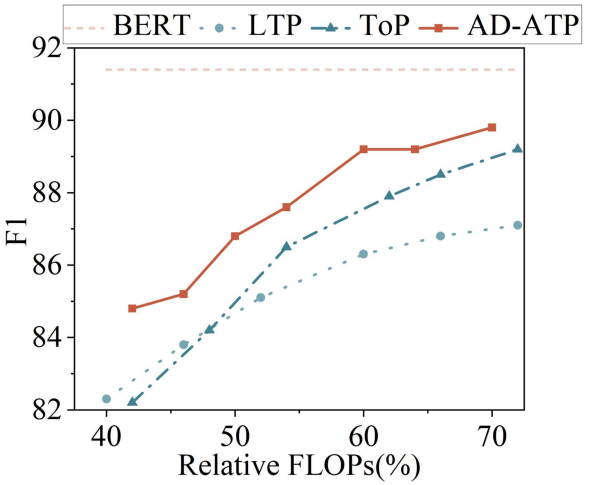
Used official implementations and grid search to ensure fair comparison (each method tuned for best trade-off at each FLOPs level).



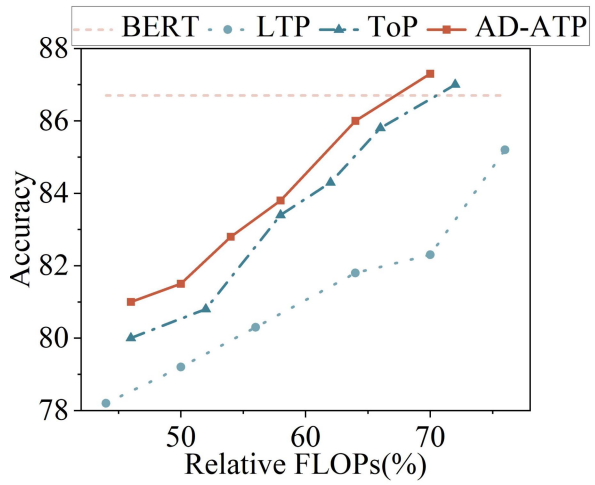
SST-2



STS-B



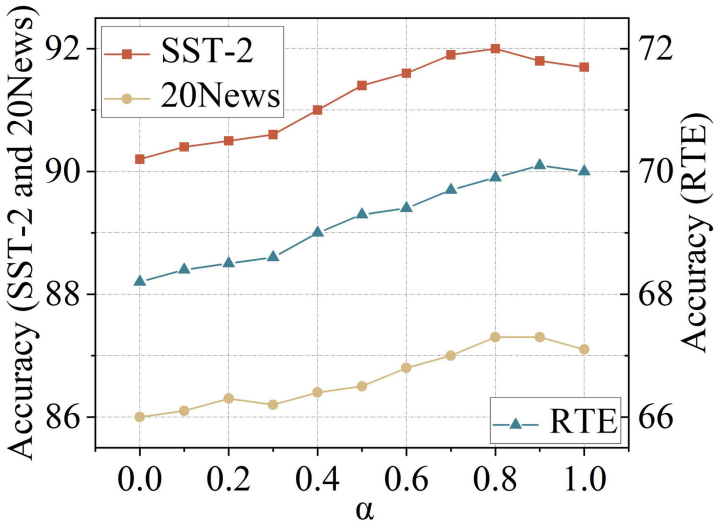
MRPC

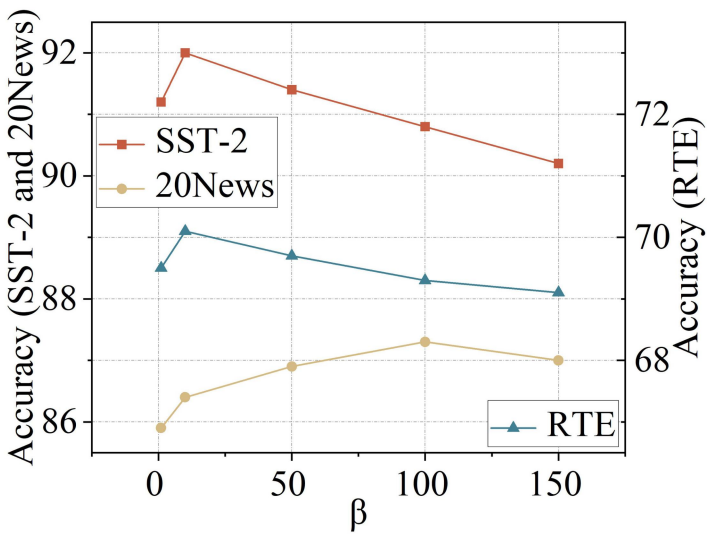


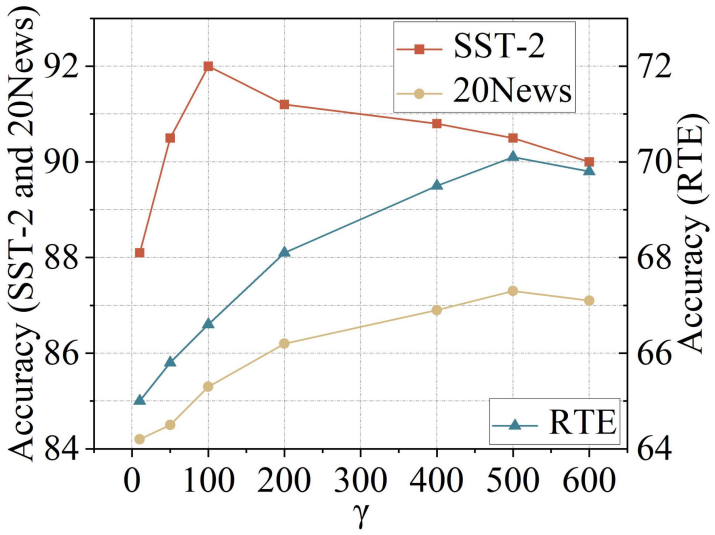
20NEWs

Impact of α, β, and γ

- Smaller α → weaker teacher guidance → performance drops.
- Larger β & γ → better attribution and retention learning.But too large → overfitting to sub-objectives → lower accuracy.
- Optimal α is consistent across tasks; best β, γ vary → task-specific tuning required for attribution & retention.





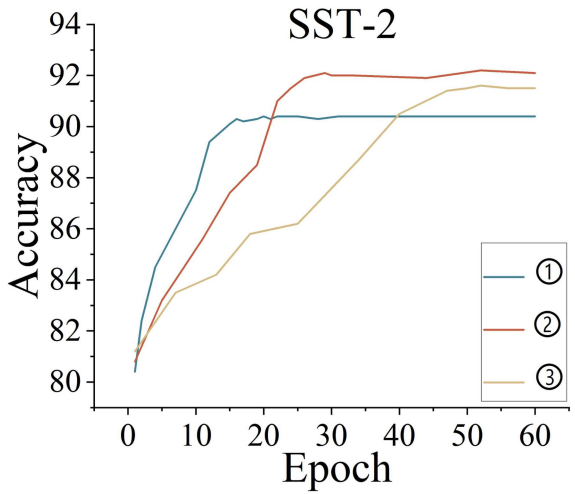


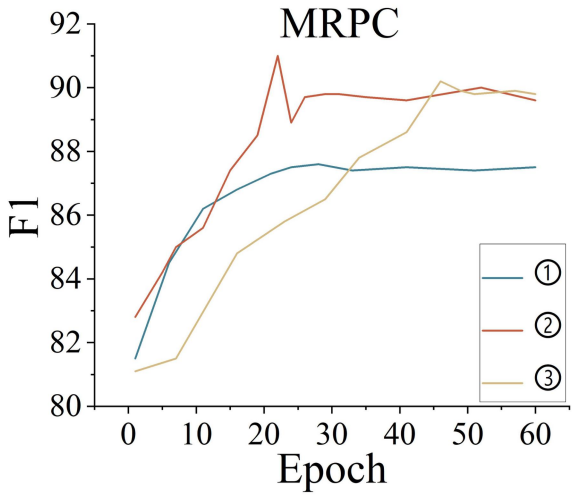
Impact of Candidate List Length L_p

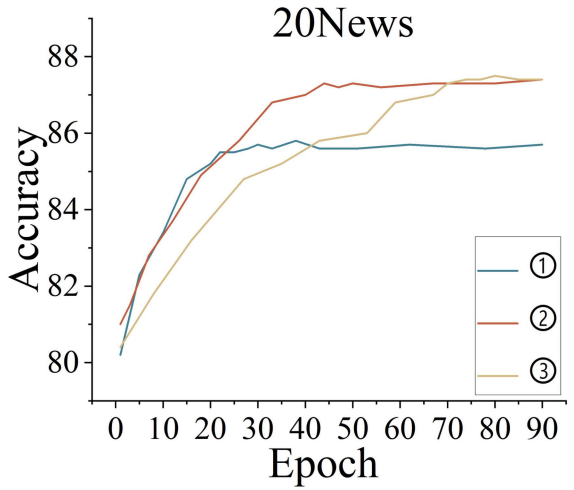
Three candidate lists with different granularity levels:

- List ①: 5 coarse-grained ratios
- List ②: 10 medium-grained ratios
- List ③: 20 fine-grained ratios

Evaluate on SST-2 and 20News datasets.







Evaluation of Token Saliency Predictor (TSP)

Purpose

1. Evaluate how well TSP approximates Integrated Gradients (IG)–based token importance.
2. Assess both ranking consistency and numerical similarity between TSP and IG scores.

Experimental Setup

Randomly sampled 1,000 validation examples from SST-2 and MRPC datasets.

Metrics used:

Spearman ρ — measures rank correlation between TSP and IG token importance.

MSE — measures numerical distance between TSP and IG saliency scores.

Key Insights

- High ranking consistency ($\rho > 0.75$) → TSP captures IG-derived importance ordering.
- Low MSE → numerical predictions closely match IG attributions.
- TSP effectively learns and generalizes the saliency distribution from IG supervision.

Table 5: Consistency Between TSP and IG Scores.

Dataset	Spearman ρ	MSE
SST-2	0.812	0.0119
MRPC	0.778	0.0143

Sensitivity to the Number of IG Steps

Background

Integrated Gradients (IG) step count m controls approximation granularity.
Default: m=1 to minimize computation.
Goal: analyze how varying $m \in \{1,3,5,7\}$ affects model performance.

Observations

Model performance shows low sensitivity to m.
Increasing steps yields slight & unstable gains (<1%).
Computational cost grows linearly with m.
Example: m=10 → ~10× longer training time.
Conclusion: m=1 achieves best balance between accuracy and efficiency.

Table 6: Performance of AD-TP with different IG step counts.

Dataset	m = 1	m = 3	m = 5	m = 7
SST-2	92.0	92.1	92.5	92.3
MRPC	89.8	89.6	90.0	90.5

山东师范大学
SHANDONG NORMAL UNIVERSITY

THANKS

弘德明志 博学笃行