Jasper-Flash: Elastic Token Compression and Progressive Distillation for Inference-Scalable Text Embedding Models

Anonymous ACL submission

Abstract

Deploying text embedding models under resource constraints is hindered by massive parameters and standard self-attention's quadratic complexity. However, existing sequence reduction strategies remain predominantly static. To address this, inspired by Matryoshka Representation Learning (MRL), we propose an Elastic Token Compression (ETC) framework that enables flexible sequence scaling for inference-time scalability. Furthermore, to stabilize training, we introduce Compression-Adaptive Progressive Distillation (CAPD) utilizing multi-teacher fusion and dynamic sampling to construct a robust, compression-tolerant semantic space. We present Jasper-Token-Compression-600M, which allows on-the-fly adjustment of encoding latency based on resources while maintaining highly competitive performance and demonstrating superior representation capacity across varying compression bounds. Our core framework remains anonymously accessible at https://anonymous.4open.science/r/Jasper-Token-Compression-Training-0DDF.

1 Introduction

While text embedding models have profoundly advanced downstream tasks such as information retrieval and document clustering (Muennighoff et al., 2022), deploying them in resource-constrained scenarios remains highly challenging (Google Research, 2024). Specifically, achieving lightweight yet high-performance representations faces two formidable bottlenecks: the *memory bottleneck* caused by massive parameter scales (Nie et al., 2024), and the *computational bottleneck* stemming from the inherent quadratic complexity of self-attention (Vaswani et al., 2017). These limitations result in prohibitive memory overhead and significant inference latency, particularly as input sequences scale in length (Vaswani et al., 2017).

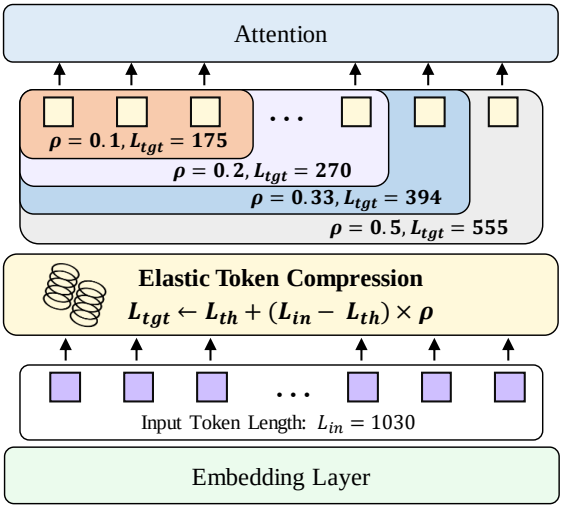

Figure 1: Illustration of the Elastic Token Compression. Positioned between the embedding and attention layers, this module utilizes a target compression ratio ρ and a length threshold (default $L_{th} = 80$) to transform a single input into multi-scale elastic target sequences.

To mitigate the aforementioned bottlenecks, token compression techniques such as merging and pruning have been widely explored (Bolya et al., 2023; EECS Department, UC Berkeley, 2023). However, even amidst recent advancements in efficient sequence modeling (Gu and Dao, 2023; Dao et al., 2022), sequence reduction strategies remain predominantly static. Furthermore, while Matryoshka Representation Learning (MRL) enables elasticity in the *representation dimension* for storage efficiency (Kusupati et al., 2022), the high computational costs during inference remain unaddressed. Therefore, there is an urgent need for embedding models with inference elasticity to balance task performance and computational overhead under varying resource constraints.

Given that the computational bottleneck of self-attention is fundamentally tied to sequence length (Vaswani et al., 2017), introducing elastic-

ity along the *sequence dimension* presents a natural and highly effective solution. To this end, inspired by MRL, we propose **Jasper-Flash**, a resource-efficient embedding framework designed to achieve **inference-time elasticity**, based on an **Elastic Token Compression (ETC)** module and a **Compression-Adaptive Progressive Distillation (CAPD)** paradigm. Specifically, as illustrated in Figure 1, the ETC module utilizes length thresholds and compression ratio sampling to flexibly scale input sequences to target lengths. Furthermore, because sequence compression is inherently a lossy truncation that poses significant optimization challenges and risks representation degradation (Wang et al., 2022), we introduce the CAPD paradigm to stabilize the learning process. By coupling multi-teacher fusion with a four-stage progressive curriculum, CAPD establishes a robust semantic foundation that successfully adapts to varying compression constraints (Formont et al., 2025).

Building upon our framework, we introduce the bilingual text embedding model, Jasper-TC-600M. Experiments demonstrate its exceptional inference elasticity: it accelerates 1024-token encoding by over 5x (down to 4.48 ms) while fully preserving semantic robustness. Crucially, even under severe compression, Jasper-TC-600M maintains highly competitive performance. Compared to its 0.6B initialization baseline, our approach yields a substantial performance leap—boosting the mean MTEB score from 70.47 to 74.75—delivering a robust semantic space that remains resilient across varying compression bounds.

In summary, our contributions are as follows:

- **Elastic Token Compression (ETC):** We propose a lightweight architecture that achieves **inference-time elasticity**. It enables on-the-fly sequence scaling under varying computational budgets, thereby flexibly optimizing inference overhead.
- **Compression-Adaptive Progressive Distillation (CAPD):** We introduce a training paradigm synergizing multi-teacher fusion with dynamic sampling to construct a robust, compression-adaptive semantic space.
- **Jasper-Token-Compression-600M:** We train and present an inference-scalable text embedding model that exhibits unprecedented inference elasticity.

2 Compression-Adaptive Progressive Distillation

2.1 Overall Methodology

Combining knowledge distillation with token compression is a highly promising approach for efficient text embedding models. However, this introduces a critical challenge: smoothly integrating token compression during distillation to balance the student model's representational capacity and compression adaptability. Since token compression is inherently a lossy truncation of features, forcing a model to simultaneously learn semantic alignment and sequence compression before establishing a robust feature foundation often leads to training instability and sub-optimal convergence.

The core design of our CAPD framework is the decoupling of learning objectives. Specifically, we first train the student model to align with the teachers' representational space under full-context conditions, and subsequently introduce flexible sequence compression to establish a robust adaptation foundation. Finally, following the standard representation learning paradigm, we perform fine-grained task adaptation to elicit specialized downstream performance.

2.2 Heterogeneous Multi-Teacher Distillation

In the first stage of CAPD, we aim to construct a comprehensive semantic foundation resilient to the information loss of token compression. Because sequence reduction is inherently lossy, any capability biases inherited from a single-teacher model risk disproportionate performance degradation on weaker tasks (Hooker et al., 2019). Consequently, relying solely on a single teacher leaves the student highly vulnerable to these compression-induced bottlenecks.

Mathematically, given independently ℓ_2-normalized sub-vectors u_i and v_i from n teacher models, their globally normalized concatenated representations, defined as $U = \mathrm{Norm}(u_1 \parallel \cdots \parallel u_n)$ and $V = \mathrm{Norm}(v_1 \parallel \cdots \parallel v_n)$, strictly satisfy:

$$U \cdot V = \frac{1}{n} \sum_{i=1}^{n} (u_i \cdot v_i) \tag{1}$$

where $\mathrm{Norm}(\cdot)$ denotes the ℓ_2 normalization, and $\parallel$ represents concatenation along the feature dimension. The proof is provided in the Appendix A.

Leveraging this capability-averaging property and extending the dual-teacher fusion pipeline, we

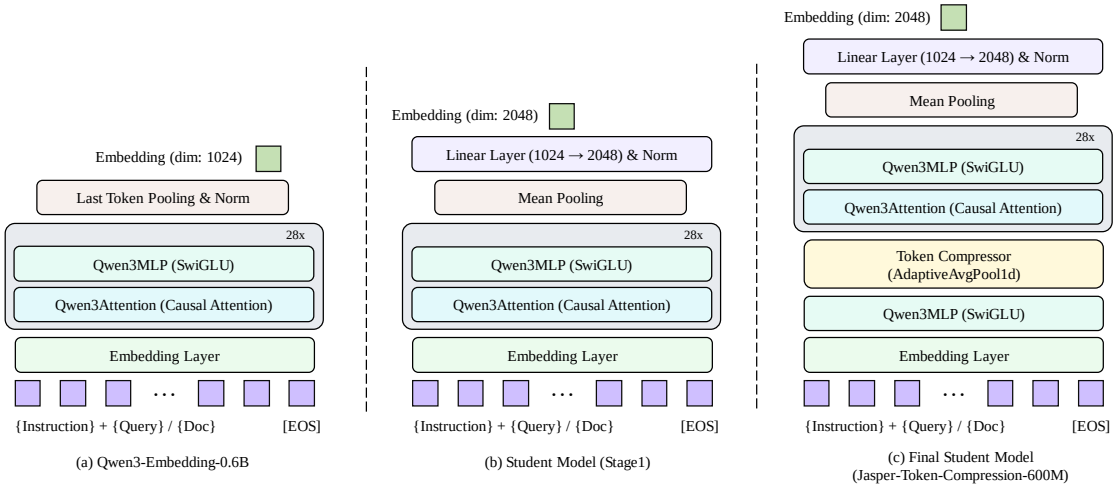

Figure 2: **Model architecture evolution.** (a) Qwen3-Embedding-0.6B, the original base model. (b) Student Model (Stage 1), adapted for unified teacher distillation by replacing last-token pooling with mean pooling and introducing a linear projection layer. (c) Final Student Model (Jasper-Token-Compression-600M), equipped with the elastic token compression module for flexibly sequence reduction.

introduce a heterogeneous multi-teacher distillation paradigm. Fusing complementary teachers neutralizes individual biases, yielding a balanced semantic space that serves as a robust prerequisite for subsequent sequence scaling. Assuming the output vector of the i-th teacher model is E_{t_i}, the target teacher representation E_t is formulated as:

$$E_t = \mathrm{Norm}\Big(\mathrm{Norm}(E_{t_1}) \parallel \cdots \parallel \mathrm{Norm}(E_{t_n})\Big)$$

To maintain computational efficiency while achieving a balanced semantic space, we specifically select two highly complementary SOTA models from the MTEB leaderboard: Qwen3-Embedding-8B (Zhang et al., 2025), which excels in asymmetric retrieval tasks (69.44), and QZhou-Embedding (7B) (Yu et al., 2025), which dominates in STS tasks (91.65). To prevent direct concatenation from yielding an excessively high-dimensional target vector, we apply customized dimensionality reduction. Leveraging its native Matryoshka Representation Learning (MRL), we directly extract the first 1024 dimensions of Qwen3, denoted as E_{qwen}. For QZhou, we compress its first 3072 dimensions into a 1024-dimensional vector via block-wise summation, denoted as E_{qzhou}. As proven in Appendix B, this operation unbiasedly preserves the semantic topology. Ultimately, the dual-teacher fusion vector used to guide the student model is formulated as:

$$E_t = \mathrm{Norm}\Big(\mathrm{Norm}(E_{qwen}) \parallel \mathrm{Norm}(E_{qzhou})\Big)$$

For the student model, we adopt Qwen3-Embedding-0.6B (Figure 2(a)) as the initialization backbone, which produces a 1024-dimensional output. To fully align the student representations with the 2048-dimensional fused teacher features, as shown in Figure 2(b), we replace the default last-token pooling with mean pooling and introduce a randomly initialized linear projection layer to upsample the features to 2048 dimensions. Following ℓ_2 normalization to obtain the student vector E_s, the learning objective is defined as minimizing the cosine similarity loss between E_s and E_t:

$$\mathcal{L}_{cosine} = 1 - E_s \cdot E_t \tag{2}$$

2.3 Elastic Token Compression

Our framework introduces a lightweight compression module immediately following the embedding layer, designed to provide inference-time elasticity (Figure 2(c)). Specifically, the text sequence first undergoes a feature transformation via a randomly initialized SwiGLU network (Shazeer, 2020) (i.e., Qwen3MLP), a non-linear projection that maximizes semantic retention before spatial reduction. The module then leverages 1D adaptive average pooling (`AdaptiveAvgPool1d`) to deterministically scale the sequence to a target length L_{tgt}. Furthermore, we employ a threshold-based scaling strategy (Algorithm 1) to prevent the destructive truncation of short queries. Given a length threshold L_{th} and a target compression ratio ρ, proportional compression is applied exclusively

Algorithm 1 Target Sequence Length Calculation

Input: Input length L_{in}, length threshold L_{th}, target compression ratio ρ
Output: Target sequence length L_{tgt}

if $L_{in} \leq L_{th}$ **then**
  $L_{tgt} \leftarrow L_{in}$ ▷ No compression below the threshold
else
  $L_{tgt} \leftarrow L_{th} + (L_{in} - L_{th}) \times \rho$
end if
return L_{tgt}

Algorithm 2 Compression Ratio Sampling

Input: Baseline compression ratio ρ
Output: Sampled compression ratio r

$p \leftarrow \text{Uniform}(0, 1)$
if $p < 0.2$ **then**
  $r \leftarrow \text{Uniform}(0.1, \rho)$
else if $p < 0.6$ **then**
  $r \leftarrow \rho$
else if $p < 0.8$ **then**
  $r \leftarrow \text{Uniform}(\rho, 2\rho)$
else
  $r \leftarrow \text{Uniform}(2\rho, 1.0)$
end if
return r

to sequences exceeding this threshold ($L_{in} > L_{th}$), thereby balancing computational efficiency with fine-grained context preservation.

During the training phase, our objective shifts to constructing a continuous and robust elastic representational space. We implement a two-stage progressive curriculum to achieve this goal. In the initial stage, the model adapts to a fixed baseline compression ratio ($r = \rho$), guided solely by the cosine similarity loss $\mathcal{L}_{cosine}$. Once this foundational alignment is established, the training advances to a dynamic generalization stage. Here, the compression ratio r is no longer static but is dynamically sampled from predefined intervals within each batch (Algorithm 2). This stochastic sampling forces the model to learn resilient feature mappings under constantly varying compression bounds. To guarantee that the compressed student representation faithfully maintains the original relative semantic topology of the teacher models, we introduce a Pairwise Similarity Loss:

$$\mathcal{L}_{similarity} = \text{MSE}(E_s E_s^\top, E_t E_t^\top) \quad (3)$$

The final joint optimization objective for this dynamic stage integrates both constraints using balancing weights λ_1 and λ_2:

$$\mathcal{L}_{s3} = \lambda_1 \mathcal{L}_{cosine} + \lambda_2 \mathcal{L}_{similarity} \quad (4)$$

2.4 Regularized Adaptation for Retrieval

The final stage shifts focus to fine-grained downstream task adaptation, specifically targeting asymmetric dense retrieval. Because precise matching is highly sensitive to the token-level signal dilution inherent in feature pooling, we continue to employ the dynamic compression sampling strategy during this phase. Exposing the model to varying compression ratios within each batch forces the retrieval-oriented representations to remain robust across the entire elastic spectrum. To prevent catastrophic forgetting of the foundational knowledge while specializing this space, we adopt a regularized fine-tuning approach.

Alongside the task-specific loss $\mathcal{L}_{task}$, we retain the cosine similarity loss $\mathcal{L}_{cosine}$ as a "semantic anchor" to stabilize the elastic structure. Additionally, a soft KL divergence loss $\mathcal{L}_{soft}$ is introduced to align the student's output distribution with the teachers' discriminative nuances. The joint optimization objective is formulated as:

$$\mathcal{L}_{s4} = \lambda_1 \mathcal{L}_{cosine} + \lambda_3 \mathcal{L}_{soft} + \mathcal{L}_{task} \quad (5)$$

We focus on task adaption for retrieval via contrastive learning, instantiate $\mathcal{L}_{task}$ using the InfoNCE loss (Oord et al., 2018) ($\mathcal{L}_{cl}$), augmented by in-batch and mined hard negatives. This multi-objective mechanism significantly enhances retrieval discriminability under flexible sequence scaling while effectively preserving the broad semantic knowledge established in previous stages.

3 Experiments

3.1 Experimental Setup

Our training pipeline comprises four progressive stages, as illustrated in Figure 3. During the first three distillation stages, we utilize a 12-million bilingual corpus (strictly balanced 1:1 for English and Chinese), followed by 2.5 million fine-tuning instances in the final stage to elicit specialized retrieval capabilities. Detailed dataset compositions and comprehensive hyperparameter settings are deferred to Appendix C and D. For elastic token compression, we empirically set the length threshold

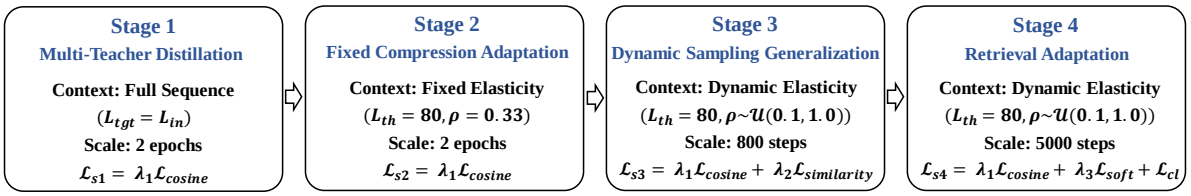

Figure 3: **Overview of the Compression-Adaptive Progressive Distillation (CAPD) Training Pipeline.** The four-stage curriculum smoothly transitions from full-sequence distillation to flexible sequence compression, concluding with retrieval-specific fine-tuning.

Model	Params	Dim.	M.Task	M.Type	Clas.	Clus.	Pair.	Rank.	Retr.	STS	Sum.
QZhou-Embedding	7B	3584	**75.97**	**69.52**	88.97	**61.65**	**92.43**	**51.77**	67.12	**91.65**	33.05
Qwen3-Embedding-8B	8B	4096	75.23	68.71	**90.43**	58.57	87.52	51.56	**69.44**	88.58	**34.83**
Qwen3-Embedding-4B	4B	2560	74.61	68.10	89.84	57.51	87.01	50.76	68.46	88.72	34.39
Qwen3-Embedding-0.6B	595M	1024	70.47	64.72	84.58	54.05	84.37	48.18	61.83	86.57	33.43
Jasper-TC-600M ($\rho = 0.50$)	595M	2048	74.75	68.46	90.35	59.44	90.15	50.60	66.19	88.79	33.66
Jasper-TC-600M ($\rho = 0.33$)	595M	2048	74.71	68.43	90.37	59.33	90.15	50.60	66.17	88.71	33.66
Jasper-TC-600M ($\rho = 0.20$)	595M	2048	74.58	68.33	90.32	59.22	90.15	50.59	65.77	88.75	33.53
Jasper-TC-600M ($\rho = 0.10$)	595M	2048	74.22	67.93	90.27	58.80	90.15	50.59	64.76	88.77	32.16

Table 1: Main results on MTEB (English). Abbreviations: Params (Parameters), Dim. (Dimension), Clas. (Classification), Clus. (Clustering), Pair. (Pair Classification), Rank. (Reranking), Retr. (Retrieval), Sum. (Summarization). M.Task and M.Type denote Mean(Task) and Mean(TaskType), respectively. The row highlighted in gray represents our best-performing configuration ($\rho = 0.50$). Best scores across all models are **bolded**.

L_{th} to 80 and the baseline compression ratio ρ_{base} to 0.33. This strategy safeguards short sentence-level queries against destructive truncation, while ensuring that long documents retain at least one-third of their core features to filter redundancy effectively. Applying these configurations, we train our final model, Jasper-Token-Compression-600M.

3.2 Main Results

To evaluate our model, we conduct comprehensive evaluations on MTEB (English) and C-MTEB (Chinese) across varying compression ratios $\rho \in \{0.5, 0.33, 0.2, 0.1\}$. We primarily report MTEB results here (Table 1); full bilingual results are detailed in Appendix E.

Despite its compact 0.6B parameter scale, Jasper-TC-600M ($\rho = 0.50$) exhibits highly competitive performance. Compared to the initialized baseline model, Qwen3-Embedding-0.6B, our approach significantly boosts the Mean (Task) score on the English MTEB from 70.47 to 74.75 (+4.28). Notably, this performance even surpasses that of Qwen3-Embedding-4B (74.61), a much larger model. This remarkable performance leap underscores the efficacy of our CAPD framework in constructing a high-quality foundational space via multi-teacher fusion. More importantly, the proposed ETC module successfully anchors this high performance profile across the entire elasticity spectrum.

Furthermore, our model demonstrates exceptional inference elasticity. As shown in Table 1, decreasing the compression ratio ρ from 0.50 to an extreme of 0.10 during inference results in only a marginal drop in the Mean (Task) score, from 74.75 to 74.22. These results further confirm that the model can provide highly reliable, elastic text representations tailored to varying computational budgets and latency constraints.

3.3 Inference Performance & Efficiency

To quantify the efficiency gains of ETC, we systematically evaluated the encoding latency across various input lengths on RTX 4090 GPUs (Table 2). Experiments demonstrate that the ETC mechanism significantly optimizes inference efficiency for extended sequences: at input lengths of 1024 and 2048 tokens, model latency is drastically reduced from the baseline's 24.24 ms and 49.99 ms to a minimum of 4.48 ms and 6.95 ms, respectively. This remarkable acceleration is attributed to our ETC module being constructed entirely on dense matrix multiplications, a design that maximizes hardware parallelization and minimizes memory access latency, ensuring highly efficient deployment across various resource-constrained scenarios.

Model	Performance	Encoding Latency (ms) vs. Input Length				
	Mean (Task)	128	256	512	1024	2048
Qwen3-Embedding-0.6B	70.47	3.22	6.47	12.20	24.24	49.99
Jasper-TC-600M ($\rho = 0.50$)	74.75	2.62	3.96	7.39	13.11	25.07
Jasper-TC-600M ($\rho = 0.33$)	74.71	2.52	3.52	5.41	9.38	17.52
Jasper-TC-600M ($\rho = 0.20$)	74.58	2.38	2.91	4.00	6.56	11.48
Jasper-TC-600M ($\rho = 0.10$)	74.22	2.09	2.56	3.18	4.48	6.95

Table 2: MTEB Mean(Task) performance and encoding latency across various compression ratios ρ. The results illustrate that Jasper-Token-Compression-600M maintains high performance while significantly reducing latency, especially for long sequences.

Model / Variant	Dim.	Clas.	Clus.	Pair.	Rank.	Retr.	STS	Sum.	M.Task	M.Type
Ablation on Heterogeneous Multi-Teacher Distillation (MTEB-Subset)										
Single Teacher (Qwen)	2048	90.12	53.03	84.17	**40.50**	57.87	84.70	32.74	68.91	63.30
Jasper-TC-600M (Stage 1)	2048	**90.35**	**55.80**	**86.20**	40.04	**58.56**	**86.15**	**32.80**	**70.11**	**64.28**
Ablation on Token Compression Module (MTEB-Subset)										
Hard Truncation	2048	76.53	41.66	74.31	26.40	22.54	78.57	11.45	51.62	47.35
Token Pruning	2048	80.19	43.98	74.83	26.59	29.78	78.53	**29.39**	55.38	51.90
Token Merging	2048	80.23	43.71	75.06	27.18	26.84	78.40	28.51	54.55	51.42
Learnable Conv1D	2048	72.70	40.36	71.74	22.80	10.21	76.11	18.32	46.90	44.60
Jasper-TC-600M (Stage 1-2)	2048	**89.94**	**51.39**	**83.97**	**37.58**	**49.58**	**85.03**	29.27	**66.28**	**61.01**
Ablation of Retrieval Adaption (Full MTEB)										
w/o Retrieval Adaption	2048	**90.49**	**59.71**	90.08	**50.84**	65.53	88.73	33.28	74.65	68.38
Jasper-TC-600M (Stage 1-4)	2048	90.35	59.44	**90.15**	50.60	**66.19**	**88.79**	**33.66**	**74.75**	**68.46**

Table 3: Comprehensive ablation study on the MTEB benchmark. We independently ablate the dual-teacher distillation strategy, the token compression architecture (evaluated at $\rho = 0.33$), and the progressive training pipeline. The *Jasper-TC-600M (Stage1-4)* denotes our final model equipped with heterogeneous dual-teachers, the MLP+AvgPool elastic module, and the complete 4-stage pipeline.

4 Ablation Study & Analysis

To manage computational costs, Sections 4.1-4.4 employ a scaled-down setting (1 epoch) on a streamlined MTEB subset (23 tasks), which efficiently reveals consistent relative performance trends. Conversely, Section 4.5 utilizes the full training configuration on the complete benchmark to ensure direct comparability with main results.

4.1 Ablation on Heterogeneous Multi-Teacher Distillation

To validate our dual-teacher distillation, we construct a capacity-controlled baseline supervised by Qwen3-Embedding-8B. Since global concatenation and normalization make homogeneous dual teachers mathematically equivalent to a single-teacher space (see Equation 1), this baseline efficiently tests both "dual vs. single" and "heterogeneous vs. homogeneous" paradigms. Table 3 shows the heterogeneous configuration outperforms the baseline (68.91 vs. 70.11), confirming that fusing complementary experts mitigates bias and builds a more resilient foundation for elastic scaling.

4.2 Ablation on Token Compression Module

To validate our ETC module, we compare it against four baselines under a controlled budget (trained up to Stage 2): **Hard Truncation**, **Token Pruning**, **Token Merging**, and **Learnable 1D Convolution**.

As shown in Table 3, our MLP-AvgPool configuration (66.28) significantly outperforms all baselines. Notably, dynamic routing mechanisms like Pruning and Merging suffer massive degradation, dropping over 20 points in asymmetric Retrieval. Their discrete operations fail to converge within a short 1-epoch budget, often destructively discarding or diluting high-entropy entity tokens critical for exact matching. Similarly, the Learnable 1D Conv performs poorest (46.90). Injecting randomly initialized parameters causes severe initialization shock, disrupting the pre-aligned semantic topology from Stage 1, while its local receptive field

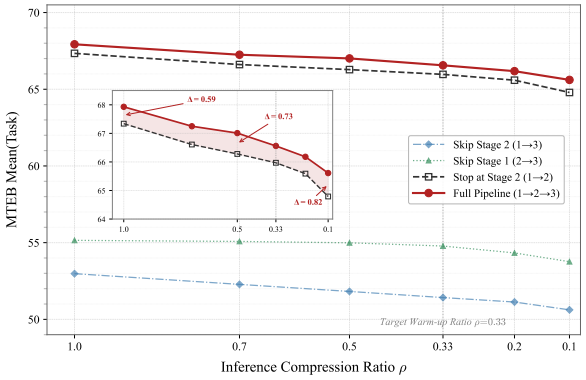

Figure 4: Ablation study of the progressive distillation pipeline across varying compression ratios.

misses long-range dependencies. In contrast, our configuration combines non-linear projection with parameter-free, deterministic spatial pooling. This guarantees a smooth, fully differentiable gradient path and a global receptive field, perfectly aligning with short-budget distillation dynamics without destroying the semantic foundation.

4.3 Ablation on Progressive Distillation Pipeline

Results in Figure 4 validate the CAPD pipeline: bypassing multi-teacher distillation (**Skip Stage 1: 2→3**) or skipping the fixed-compression adaption (**Skip Stage 2: 1→3**) triggers catastrophic representation collapse under extreme compression. This proves the necessity of decoupling learning objectives as outlined in Section 2.1.

Furthermore, comparing early halting (**Stop at Stage 2: 1→2**) with the complete configuration (**Full Pipeline: 1→2→3**) reveals that the primary benefit of dynamic compression sampling (Stage 3) is not merely flattening the degradation curve, but inducing a significant global regularization effect. It forces the model to maintain semantic topology across randomly fluctuating lengths, enabling the extraction of more robust, position-agnostic core features, ultimately achieving a global upward shift of the entire elasticity curve.

4.4 Parameter Sensitivity Analysis

We evaluate the sensitivity of L_{th} and ρ under a Stage 2 controlled setting (Figure 5). As shown in the top panel ($L_{th} = 80$), reducing ρ from 1.0 to 0.33 halves encoding latency with only graceful MTEB degradation, though extreme compression ($\rho = 0.1$) triggers a sharp performance drop. Conversely, fixing $\rho = 0.33$ (bottom panel) reveals

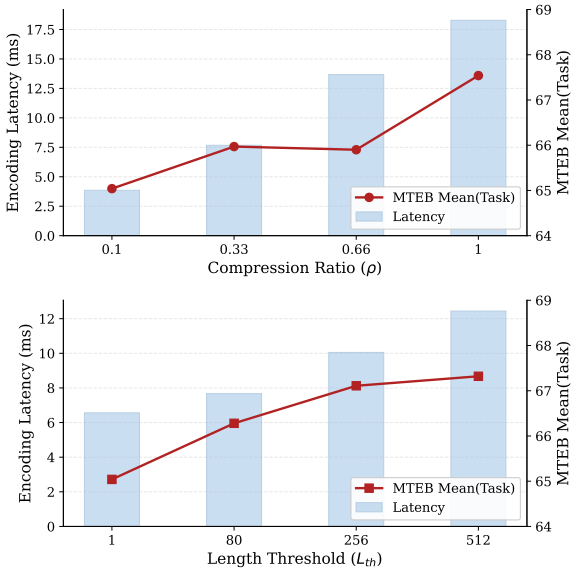

Figure 5: Parameter sensitivity analysis. (a) Impact of the compression ratio ρ ($L_{th} = 80$). (b) Impact of the length threshold L_{th} ($\rho = 0.33$).

that $L_{th} = 1$ destructively dilutes short queries, whereas $L_{th} = 80$ sharply recovers performance with a negligible latency penalty; further relaxation (e.g., 256) inflates latency for marginal gains. This firmly justifies our configuration ($\rho = 0.33$, $L_{th} = 80$) as the optimal sweet spot that minimizes long-document computational overhead while safeguarding short-query semantics.

4.5 Ablation of Retrieval Adaption

To validate the contrastive fine-tuning phase (Stage 4) for asymmetric dense retrieval, we ablate this stage as reported in Table 3. Following the targeted task adaptation in Stage 4, Retrieval performance improves notably (from 65.53 to 66.19). Crucially, while performance on other symmetric benchmarks exhibits only marginal, zero-sum empirical fluctuations, the substantial and isolated gain in Retrieval confirms a successful targeted enhancement rather than random variance. This demonstrates that Stage 4 effectively elicits specialized fine-grained retrieval capabilities without inducing catastrophic forgetting across the foundational semantic space.

5 Qualitative Analysis

To intuitively demonstrate our elastic compression, we analyze a retrieval sample where the core evidence resides at the extreme tail. As Figure 6 illustrates, while standard Hard Truncation" physically discards this segment causing complete retrieval

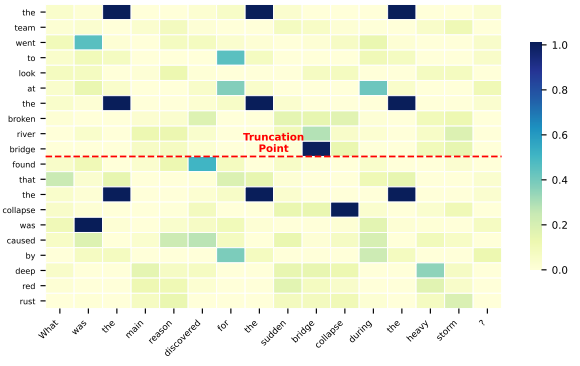

Figure 6: Token-level similarity heatmap (vertical axis: compressed document tokens; horizontal axis: query tokens). It illustrates how ETC's semantic distillation preserves core tail evidence beyond the hard truncation point (red dashed line).

failure, our **ETC** module maintains robust query alignment. The token-level heatmap reveals that ETC implicitly preserves high activation signals (dark blue clusters) for core anchor tokens beyond the truncation point. This confirms that our architecture avoids mechanical sequence loss, performing a global semantic distillation" to successfully retain indispensable context. For an extended qualitative analysis across other challenging scenarios, please refer to Appendix H.

6 Related Work

6.1 Text Embeddings and Knowledge Distillation

Text embedding models have evolved from context-unaware vectors and Siamese architectures like SBERT (Reimers and Gurevych, 2019) to weak-supervised pre-training such as E5 (Wang et al., 2022), and further to dominant LLM-based systems including Qwen3 (Zhang et al., 2025), QZhou (Yu et al., 2025), and KaLM (Zhao et al., 2025). To deploy these massive models in resource-constrained environments, knowledge distillation (KD) serves as a critical technique to bridge the performance-efficiency gap. KD frameworks have advanced from strict layer-wise feature matching like TinyBERT (Jiao et al., 2019), MobileBERT (Sun et al., 2020), and MiniLM (Wang et al., 2020) to generative capability extraction via LLM-driven data synthesis or pseudo-labeling in Gecko (Lee et al., 2024). However, while multi-teacher distillation has been explored for LLMs (Jin et al., 2026), these existing text embedding pipelines predominantly rely on a single teacher, inherently inheriting specific domain biases. To address this, we propose a heterogeneous multi-teacher distillation mechanism that fuses complementary expert models to construct an unbiased, balanced representation space.

6.2 Efficient Inference and Token Compression

The quadratic complexity ($\mathcal{O}(L^2)$) of Transformer self-attention creates a severe inference bottleneck. While exact hardware-aware optimizations like FlashAttention (Dao et al., 2022) or linear architectures like Mamba (Gu and Dao, 2023) alleviate this, they either retain asymptotic complexity or require prohibitive retraining costs. Alternatively, sequence reduction within existing Transformers is widely explored. Token pruning methods, such as PoWER-BERT (Goyal et al., 2020), Funnel-Transformer (Dai et al., 2020), and learned pruning policies (EECS Department, UC Berkeley, 2023; Yun et al., 2023), discard less important tokens but risk losing critical entities. Conversely, token merging techniques like ToMe (Bolya et al., 2023; Bolya and Hoffman, 2023) fuse redundant features non-destructively but rely on dynamic operations that impair GPU parallelization. To achieve runtime flexibility, we introduce a dense-matrix-driven Elastic Token Compression (ETC) architecture that elastically scales sequences on-the-fly.

7 Conclusion

We present **Jasper-Flash**, a resource-efficient framework that addresses computational redundancy in text embedding models via **Elastic Token Compression**. Our lightweight **MLP-AvgPool** module enables elastic sequence scaling while maintaining semantic integrity. We propose a **Compression-Adaptive Progressive Distillation (CAPD)** paradigm that couples multi-stage sampling with heterogeneous teacher fusion to construct a robust, compression-adaptive semantic space. Results demonstrate that our 0.6B-parameter model, **Jasper-Token-Compression-600M**, delivers highly competitive bilingual performance, substantially outperforming its initialization baseline on the MTEB and C-MTEB benchmarks. The framework significantly reduces memory overhead and inference latency, particularly when processing variable-length sequences, offering a scalable solution for high-performance representation learning under strict resource constraints.

Limitations

Retrieval Performance Trade-offs. A performance gap in dense retrieval persists compared to the 8B-parameter teacher models. We attribute this to three primary factors: (1) as derived from **Lemma 1**, our multi-teacher fusion anchors the target space to a mathematical average. While this prevents over-fitting to a single teacher's biases and improves generalizability, it inherently restricts the student (66.19) from perfectly matching the specialized retrieval peak of a single proficient teacher; (2) the 0.6B parameter scale faces a physical capacity bottleneck when attempting to internalize the extensive world knowledge of 8B-level models; and (3) while our compression module is highly efficient, the average pooling operation prioritizes global semantic context, which can occasionally smooth over fine-grained, token-level entity signals required for exact-match asymmetric retrieval.

Granularity of Token Compression. While our MLP-AvgPool compression module achieves remarkable macro-elasticity through dynamic sequence truncation and length thresholds, the pooling mechanism itself applies uniformly across the retained sequence. It does not yet perform selective, token-level semantic routing (e.g., dynamically preserving rare keywords while aggressively dropping stop-words). This uniform spatial reduction represents a theoretical ceiling for information retention under extreme compression bounds.

Context Length Scope. Our Jasper-TC-600M model is currently trained and evaluated with a maximum context window of 1,030 tokens. This intentional constraint allows us to strictly optimize performance for standard benchmarks within compute-efficient budgets, directly reflecting the realities of lightweight, resource-constrained deployment scenarios.

To address these structural constraints, our subsequent research will explore context-adaptive, attention-guided compression policies and extend the distillation pipeline to support ultra-long context windows.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Beijing Academy of Artificial Intelligence (BAAI). 2024a. IndustryCorpus2. https://huggingface.co/datasets/BAAI/IndustryCorpus2_artificial_intelligence_machine_learning.

Beijing Academy of Artificial Intelligence (BAAI). 2024b. Infinity-Instruct. https://huggingface.co/datasets/BAAI/Infinity-Instruct.

Daniel Bolya, Cheng-Zhong Hoffman, Hao Yao, and Judy Hoffman. 2023. Token merging: Your ViT but faster. In *International Conference on Learning Representations*.

Daniel Bolya and Judy Hoffman. 2023. Token merging for fast stable diffusion. In *CVPR Workshop on Efficient Deep Learning for Computer Vision*.

Luiz Henrique Bonifacio, Vitor Jeronymo, Hugo Q Abonizio, Iman Fadaei, and Rodrigo Nogueira. 2021. mMARCO: A multilingual version of MS MARCO passage ranking dataset. *arXiv preprint arXiv:2108.13897*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Zihang Dai, Guokun Lai, Yiming Yang, and Quoc V Le. 2020. Funnel-Transformer: Filtering out sequential redundancy for efficient language processing. *Advances in Neural Information Processing Systems*, 33:4271–4282.

Tri Dao. 2023. FlashAttention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

Tri Dao, Daniel Y Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. *Advances in Neural Information Processing Systems*, 35:10482–10493.

EECS Department, UC Berkeley. 2023. Learned token pruning for efficient transformer inference. Technical Report EECS-2023-119, University of California, Berkeley.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *ACL*.

Philippe Formont, Maxime Darrin, Banafsheh Karimian, Eric Granger, and Jackie CK Cheung. 2025. Learning task-agnostic representations through multi-teacher distillation. *arXiv preprint arXiv:2510.18680*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, and Charles Foster. 2021a. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021b. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Google Research. 2024. Efficient sequence modeling for on-device ML. https://research.google/blog/efficient-sequence-modeling-for-on-device-ml/.

Saurabh Goyal, Anamitra Roy Choudhury, Sateesh Ramezani, Venkatesan Chakaravarthy, and Ashish Sabharwal. 2020. PoWER-BERT: Accelerating BERT inference via progressive word-vector elimination. In *International Conference on Machine Learning*, pages 3690–3699.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, and Shiqi Zhao. 2018. DuReader: a chinese machine reading comprehension dataset from real-world applications. In *ACL Workshop*.

Doris Hoogeveen, L. Wang, Timothy Baldwin, and Karin Verspoor. 2015. CQADupStack: A benchmark data set for community question-answering research. In *ADCS*.

Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Webb. 2019. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*.

Yupeng Hou, Jiacheng Li, Xiangjun Fu, Zhankui He, and An Yan. 2024. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. TinyBERT: Distilling BERT for natural language understanding. *arXiv preprint arXiv:1909.10351*.

Ruihan Jin, Yutao Wang, Xiaolong Chen, and 1 others. 2026. Exploring knowledge purification in multi-teacher knowledge distillation for LLMs. *arXiv preprint arXiv:2602.01064*.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Aditya Kusupati, Gantavya Bhatt, Aniket Reval, M Karvonen, Arsha Somani, Sham Kakade, and Prateek Jain. 2022. Matryoshka representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 30233–30249.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, and Ankur Parikh. 2019. Natural questions: a benchmark for question answering research. *TACL*.

Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, and Daniel Cer. 2024. Gecko: Versatile text embeddings distilled from large language models. *arXiv preprint arXiv:2403.20327*.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, and Heinrich Küttler. 2021. PAQ: 65 million probably-asked questions and what you can do with them. *TACL*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S. Weld. 2020. S2ORC: The semantic scholar open research corpus. *ACL*.

Dingkun Long, Qiong Gao, Kuan Zou, Guangwei Xu, and Pengjun Xie. 2022. Multi-CPR: A multi domain chinese dataset for passage retrieval. In *SIGIR*.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024. FineWeb-Edu: A large-scale dataset for educational content. https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu.

MOP-LIWU Community and MNBVC Team. 2023. MNBVC: Massive never-ending bt vast chinese corpus. https://github.com/esbatmop/MNBVC.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2022. MTEB: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, and Saurabh Tiwary. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *NIPS*.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *EMNLP*.

Zhijie Nie, Zilong Feng, Mengran Li, Hao Zheng, and Shengyu Zhu. 2024. When text embedding meets large language model: A comprehensive survey. *arXiv preprint arXiv:2412.09165*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Open-Source Community. Awesome-Chinese-LLM. https://github.com/HqWu-HITCS/Awesome-Chinese-LLM.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, and Alessandro Cappelli. 2024. The FineWeb datasets: Decanting the web for the finest text at scale. *arXiv preprint arXiv:2406.17557*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *ACL*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. *arXiv preprint arXiv:1908.10084*.

Noam Shazeer. 2020. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Xiaofeng Shi, Lulu Zhao, Hua Zhou, Donglin Hao, and Yonghua Lin. 2024. IndustryInstruction. https://huggingface.co/datasets/BAAI/IndustryInstruction.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. MobileBERT: a compact task-agnostic BERT for resource-limited devices. *Proceedings of the Association for Computational Linguistics*, pages 2158–2170.

Hongxuan Tang, Hongyu Li, Jing Liu, Yu Hong, and Hua Wu. 2021. DuReader_robust: A chinese dataset towards evaluating robustness and generalization of machine reading comprehension in real-world applications. In *ACL*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, and Thomas Wolf. 2024. Smoltalk: A dataset for SmolLM training. https://huggingface.co/datasets/HuggingFaceTB/smoltalk.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, volume 30.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Xiaohui Xie, Qian Dong, Bingning Wang, Feiyang Lv, and Ting Yao Zhang. 2023. T2Ranking: A large-scale chinese benchmark for passage ranking. In *SIGIR*.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, and Chenjie Cao. 2020. CLUE: A chinese language understanding evaluation benchmark. In *COLING*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, and William W. Cohen. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, and Zhengying Liu. 2023. MetaMath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*.

Peng Yu, En Xu, Bo Chen, Hao Chen, and Yan Xu. 2025. QZhou-Embedding technical report: Robust text representation. *arXiv preprint arXiv:2508.21632*.

Sha Yuan, Hanyu Zhao, Zhengxiao Du, Ming Ding, and Xiao Liu. 2021. WuDaoCorpora: A super large-scale chinese corpora for pre-training language models. *AI Open*.

Jungmin Yun, Mihyeon Kim, and Youngbin Kim. 2023. Focus on the core: Efficient attention via pruned token compression for document classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13617–13628.

Sheng Zhang, Xin Zhang, Hui Wang, Jiajun Cheng, and Pei Li. 2018. Multi-scale attentive interaction networks for chinese medical question answer selection. *IEEE Access*.

Yanzhao Zhang, Mengran Li, Dingkun Long, Yuchen Ren, and Lei Zhang. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Xinping Zhao, Xiaofan Hu, Zeqi Shan, Yuzheng Du, and Yutong Hao. 2025. KaLM-Embedding-V2: Superior training techniques and data inspire a versatile embedding model. *arXiv preprint arXiv:2506.20923*.

Tianyu Zheng, Ge Zhang, Tianhao Shen, Xueling Liu, and Bill Yuchen Lin. 2024. OpenCodeInterpreter: Integrating code generation with execution and refinement. In *ACL*.

A Proof of the Capability-Averaging Property

In Section 2.2, we introduce a heterogeneous multi-teacher distillation paradigm. To mathematically ground the capability-averaging property of concatenating multiple teacher models, we prove that the global dot product of concatenated and subsequently normalized representations is strictly equivalent to the arithmetic mean of their individual sub-vector dot products.

Lemma 1 (Dot-Product Equivalence)**.** *Given n independently l_2-normalized sub-vectors $u_i, v_i \in \mathbb{R}^d$ for $i \in \{1, 2, \ldots, n\}$, satisfying $||u_i||_2 = 1$ and $||v_i||_2 = 1$. Let the unnormalized concatenated vectors be defined as $\tilde{U} = [u_1; u_2; \ldots; u_n]$ and $\tilde{V} = [v_1; v_2; \ldots; v_n]$, where $[\cdot;\cdot]$ denotes feature concatenation along the channel dimension. If the global representations U and V are obtained by*

globally l_2-normalizing $\tilde{U}$ and $\tilde{V}$ respectively, it strictly holds that:

$$U \cdot V = \frac{1}{n} \sum_{i=1}^{n} (u_i \cdot v_i) \tag{6}$$

Proof. First, we calculate the l_2 norm of the unnormalized concatenated vector $\tilde{U}$. By the definition of the l_2 norm for concatenated spaces, the squared norm of $\tilde{U}$ is the sum of the squared norms of its constituent sub-vectors:

$$||\tilde{U}||_2^2 = \sum_{i=1}^{n} ||u_i||_2^2 = \sum_{i=1}^{n} 1 = n \tag{7}$$

Thus, the norm is $||\tilde{U}||_2 = \sqrt{n}$. By the same logic, $||\tilde{V}||_2 = \sqrt{n}$.

Next, we formulate the globally normalized vectors U and V by scaling the concatenated vectors:

$$U = \frac{\tilde{U}}{||\tilde{U}||_2} = \frac{1}{\sqrt{n}}\tilde{U}, \quad V = \frac{\tilde{V}}{||\tilde{V}||_2} = \frac{1}{\sqrt{n}}\tilde{V} \tag{8}$$

Now, we expand the global dot product of U and V by extracting the scalar scaling factors:

$$U \cdot V = \left(\frac{1}{\sqrt{n}}\tilde{U}\right) \cdot \left(\frac{1}{\sqrt{n}}\tilde{V}\right) = \frac{1}{n}(\tilde{U} \cdot \tilde{V}) \tag{9}$$

Since the dot product of two concatenated vectors is exactly the sum of the dot products of their corresponding aligned sub-vectors, we have:

$$\tilde{U} \cdot \tilde{V} = \sum_{i=1}^{n} (u_i \cdot v_i) \tag{10}$$

Substituting this decomposition back into the expanded global dot product yields the final equivalence:

$$U \cdot V = \frac{1}{n} \sum_{i=1}^{n} (u_i \cdot v_i) \tag{11}$$

□

B Theoretical Justification of Dimensionality Reduction for Non-MRL Teachers

In Section 2.2, we compress the 3072-dimensional representation of the QZhou model into a 1024-dimensional vector via parameter-free block-wise summation. To theoretically justify why this operation preserves the semantic topology of the original space without causing catastrophic information loss, we provide the following mathematical proof.

Assumption 1 (Isotropic and Zero-Mean Features)**.** *Following standard theoretical simplifications in representation learning, we model the high-dimensional embedding features as approximately zero-mean and isotropic (i.e., dimensions are orthogonal and independent). This assumption is empirically grounded: (1) the ubiquitous use of Layer Normalization (Ba et al., 2016) enforces a near-zero mean across hidden states; (2) modern contrastive fine-tuning objectives specifically optimize for representation uniformity (Gao et al., 2021b; Wang and Isola, 2020), effectively mitigating anisotropy and pushing the high-dimensional semantic space to be isotropic. Under this premise, the expected product of mismatched dimensions approaches zero, i.e., $\mathbb{E}[x_a y_b] \approx 0$ for $a \neq b$.*

Proof. Let $x, y \in \mathbb{R}^D$ be two original high-dimensional embedding vectors (e.g., $D = 3072$), and $\hat{x}, \hat{y} \in \mathbb{R}^d$ be the reduced vectors (e.g., $d = 1024$). The block size is $k = D/d = 3$. The block-wise summation for the i-th dimension of $\hat{x}$ is defined as:

$$\hat{x}_i = \sum_{j=1}^{k} x_{(i-1)k+j} \tag{12}$$

The dot product of the reduced vectors in the d-dimensional space is calculated as:

$$\begin{aligned} \hat{x} \cdot \hat{y} &= \sum_{i=1}^{d} \hat{x}_i \hat{y}_i \\ &= \sum_{i=1}^{d} \left(\sum_{j=1}^{k} x_{(i-1)k+j} \right) \left(\sum_{l=1}^{k} y_{(i-1)k+l} \right) \end{aligned} \tag{13}$$

Expanding the product, we can decouple the terms into diagonal components (where the original dimensions match) and cross components (where they are mismatched):

$$\begin{aligned} \hat{x} \cdot \hat{y} = &\sum_{i=1}^{d} \sum_{j=1}^{k} x_{(i-1)k+j} y_{(i-1)k+j} \\ &+ \sum_{i=1}^{d} \sum_{j \neq l}^{k} x_{(i-1)k+j} y_{(i-1)k+l} \end{aligned} \tag{14}$$

Notice that the first term is exactly equivalent to the dot product in the original D-dimensional

space:

$$\sum_{i=1}^{d}\sum_{j=1}^{k} x_{(i-1)k+j}y_{(i-1)k+j} = \sum_{m=1}^{D} x_m y_m = x \cdot y \tag{15}$$

Let the second term be defined as the cross-dimensional noise Δ_{cross}. Thus, we have $\hat{x} \cdot \hat{y} = x \cdot y + \Delta_{cross}$.

Applying Assumption 1, since the dimensions are modeled as zero-mean and independent, the expectation of the cross-dimensional product is zero:

$$\mathbb{E}[\Delta_{cross}] = \sum_{i=1}^{d}\sum_{j\neq l}^{k} \mathbb{E}[x_{(i-1)k+j}y_{(i-1)k+l}] \approx 0 \tag{16}$$

Taking the expectation of the entire equation, the noise term cancels out, yielding:

$$\mathbb{E}[\hat{x} \cdot \hat{y}] \approx x \cdot y \tag{17}$$

Therefore, we conclude that the block-wise summation operation unbiasedly preserves the expected dot product (and subsequently the relative cosine similarity after global l_2 normalization) of the original high-dimensional space. This ensures high-fidelity knowledge distillation without introducing uninitialized learnable parameters. □

C Training Detail

Regarding implementation details, the core training is executed on 4 RTX 4090 GPUs, utilizing the Adam optimizer (Kingma and Ba, 2014) and FlashAttention-2 (Dao, 2023). To ensure the full reproducibility of our experiments, we detail the comprehensive hyperparameter configurations utilized across the four-stage progressive training pipeline in Table 4.

D Dataset Construction and Mining Methodology

This research employs a dual-stage training framework designed to systematically enhance the bilingual representation capabilities of our text embedding model.

D.1 Stage 1: Knowledge Distillation

The objective of this stage is to establish a robust general semantic space by extracting features from large-scale heterogeneous corpora. We integrated approximately 12 million training pairs, comprising queries (q), short sentences (s), and long passages (p), with a strict 1:1 ratio between English and Chinese data. Mining is conducted via two strategies: (1) **Active Ratio Strategy**, which extracts queries from public SFT and QA datasets while matching corresponding s and p from pre-training corpora to ensure balanced domain distribution; and (2) **Heuristic Rule Extraction**, which utilizes predefined patterns (e.g., interrogative structures, instruction verbs) to automatically identify potential query contexts from massive unsupervised corpora such as WuDao (Yuan et al., 2021). Data cleaning follows the principle of "preserving original diversity," removing only obvious formatting noise to maximize feature generalization.

D.2 Stage 2: Supervised Fine-tuning (SFT)

Following the distillation phase, the model is refined using approximately 2.5 million high-quality supervised samples processed via the **KaLM** framework (Zhao et al., 2025). This stage focuses on enhancing the model's discriminative power for fine-grained semantic boundaries. All tasks are reformulated into a unified retrieval format with task-specific instructions. Two core techniques are applied: (1) **Ranking Consistency Filtering**, which validates query-positive alignment by checking similarity rankings and discarding false negatives; and (2) **KaLM Hard Negative Mining**, which manually selects documents ranked between 50 and 100 in the initial retrieval pool as hard negatives. This forces the model to learn more precise decision boundaries compared to random negative sampling. For datasets that appear in both our training mixture and downstream evaluation benchmarks (e.g., MS MARCO, NQ, and T2Ranking), we strictly utilize their official training splits for supervised fine-tuning.

E Full Results

In this section, we expand upon the main experiments by providing a more exhaustive evaluation. Specifically, Table 7 presents a comprehensive comparative analysis of our model across varying compression ratios against baseline models on both the MTEB and C-MTEB benchmarks. Table 8 and Table 9 present an extended comparative analysis against a broader range of baseline models on both the MTEB (English) and C-MTEB (Chinese) benchmarks. Furthermore, Table 10 and Table 11 detail the fine-grained, dataset-specific

Category	Hyperparameter	Stage 1	Stage 2	Stage 3	Stage 4
Training	Training Duration	2 epochs	2 epochs	800 steps	5,000 steps
	Optimizer	Adam	Adam	Adam	Adam
	Learning Rate	1e-4	7e-5	7e-5	2e-5
	LR Scheduling	Cosine	Cosine	Cosine	Cosine
	Warmup Ratio	0.005	0.005	0.005	-
	Max Seq. Length	1,030	1,030	1,030	-
	Global Batch Size	256	256	512	64
Compression	Length Threshold (L_{th})	-	80	80	80
	Baseline Ratio (ρ)	-	0.33	0.33	0.33
	Actual Mode	-	Fixed	Dynamic	Dynamic
Loss & Task	Cosine Weight (λ_1)	10	10	10	10
	Similarity Weight (λ_2)	-	-	100	-
	KL Weight (λ_3)	-	-	-	16
	Hard Negatives (K)	-	-	-	3
	Temp. (τ / α)	-	-	-	0.3 / 0.1

Table 4: Comprehensive summary of hyperparameters across the four training stages. λ_1, λ_2, and λ_3 denote the balancing weights for cosine similarity, pairwise similarity, and KL divergence losses, respectively, as defined in the methodology.

Table 5: Summary of Knowledge Distillation Datasets (approx. 12M pairs)

Source Dataset	Format	Language
The Pile (Gao et al., 2021a)	s, p	English
FineWeb / FineWeb-Edu (Penedo et al., 2024)	s, p	English
IndustryCorpus2 (Beijing Academy of Artificial Intelligence (BAAI), 2024a)	s, p	Multilingual
FineWeb-Edu-Chinese (Lozhkov et al., 2024)	s, p	Chinese
Sentence-transformers Training Data (Reimers and Gurevych, 2019)	q	English
BGE-M3-Data (Chen et al., 2024)	q	Multilingual
Industry Instruction (Shi et al., 2024)	q	English
Infinity-Instruct (Beijing Academy of Artificial Intelligence (BAAI), 2024b)	q	English
Smoltalk (Tunstall et al., 2024)	q	English
Smoltalk-Chinese (Tunstall et al., 2024)	q	Chinese
MNBVC (MOP-LIWU Community and MNBVC Team, 2023)	q	Chinese
CLUE-WebQA & Encyclopedia QA (Xu et al., 2020)	q	Chinese
Awesome-LLM-Datasets (ZH) (Open-Source Community)	q	Chinese
WuDao Corpora (Queries) (Yuan et al., 2021)	q	Chinese

Note: q denotes Query; s denotes Sentence; p denotes Passage.

performance breakdown of our Jasper-TC-600M model across all constituent task categories.

F Details of the MTEB Subset

Evaluation Settings: To manage computational costs, the ablations in Sections 4.1-4.4 employ a scaled-down setting (1 epoch) evaluated on a streamlined MTEB subset (23 datasets, detailed in Appendix F), which efficiently reveals consistent relative performance trends. Conversely, Section 4.5 utilizes the full training configuration (2 epochs) on the complete benchmark to ensure direct comparability with our main results.

G Extended Qualitative Analysis

To provide a comprehensive view of our Elastic Token Compression mechanism, we present extended token-level similarity heatmaps across diverse retrieval scenarios in Figure 7(a)–(d), with the exact text pairs detailed in Table 13.

H Extended Qualitative Analysis

To provide a comprehensive view of our ETC module and mitigate potential cherry-picking bias in our experimental evaluations, we present an extended qualitative analysis across a total of five diverse and challenging retrieval scenarios: (1) Extreme Tail, (2) Lost in the Middle, (3) Scattered Evidence, (4) Dense Distraction, and (5) Temporal Update. Table 13 details the original queries and the highly compressed document contexts for all five cases. To visually demonstrate our model's alignment capabilities, the corresponding token-level similarity heatmaps are distributed across the

Table 6: Summary of Supervised Fine-tuning Datasets (approx. 2.5M pairs)

Source Dataset	Format	Language
MS MARCO (Nguyen et al., 2016)	s2p	English
Natural Questions (NQ) (Kwiatkowski et al., 2019)	s2p	English
SQuAD 2.0 (Rajpurkar et al., 2018)	s2p	English
HotpotQA (Yang et al., 2018)	s2p	English
TriviaQA (Joshi et al., 2017)	s2p	English
ELI5 (Fan et al., 2019)	s2p	English
PAQ (Part 2) (Lewis et al., 2021)	s2p	English
T2Ranking (Xie et al., 2023)	s2p	Chinese
DuReader / DuReader Checklist (He et al., 2018; Tang et al., 2021)	s2p	Chinese
mMARCO (zh) (Bonifacio et al., 2021)	s2p	Chinese
Multi-CPR (Ecom/Medical/Video) (Long et al., 2022)	s2p	Chinese
S2ORC (Lo et al., 2020)	s2p	English
CMedQA-V2.0 (Zhang et al., 2018)	s2p	Chinese
Amazon QA / Amazon Reviews (Ni et al., 2019; Hou et al., 2024)	s2p	English
CodeFeedback (Zheng et al., 2024)	s2p	English
StackExchange DupQuestions (Hoogeveen et al., 2015)	s2s	English
MetaMathQA (Yu et al., 2023)	s2p	English

Note: s2p denotes sentence-to-passage (asymmetric); s2s denotes sentence-to-sentence (symmetric).

paper: the heatmap for Case 1 is presented in the main text (Figure 6), while the heatmaps for Cases 2 through 5 correspond to subfigures (a) through (d) in Figure 7 of this appendix, respectively.

Within Table 13, the text highlighted in **bold** corresponds to the core evidence tokens that exhibited the highest activation scores against the query tokens in our model's compressed semantic space. As illustrated in the accompanying heatmaps, the red dashed line denotes the strict sequence-length budget (i.e., the Hard Truncation Point). While standard baselines are forced to mechanically discard tokens beyond this boundary, our Jasper-TC-600M model equipped with the ETC module elastically compresses the entire document context into the exact same budget. The heatmaps reveal that our model successfully preserves the highly activated core evidence (indicated by the prominent dark blue clusters). This confirms that our framework conducts robust, position-agnostic semantic distillation during the retrieval process, successfully preserving indispensable evidence regardless of its absolute physical position, lexical overlaps, or temporal shifts in the original text.

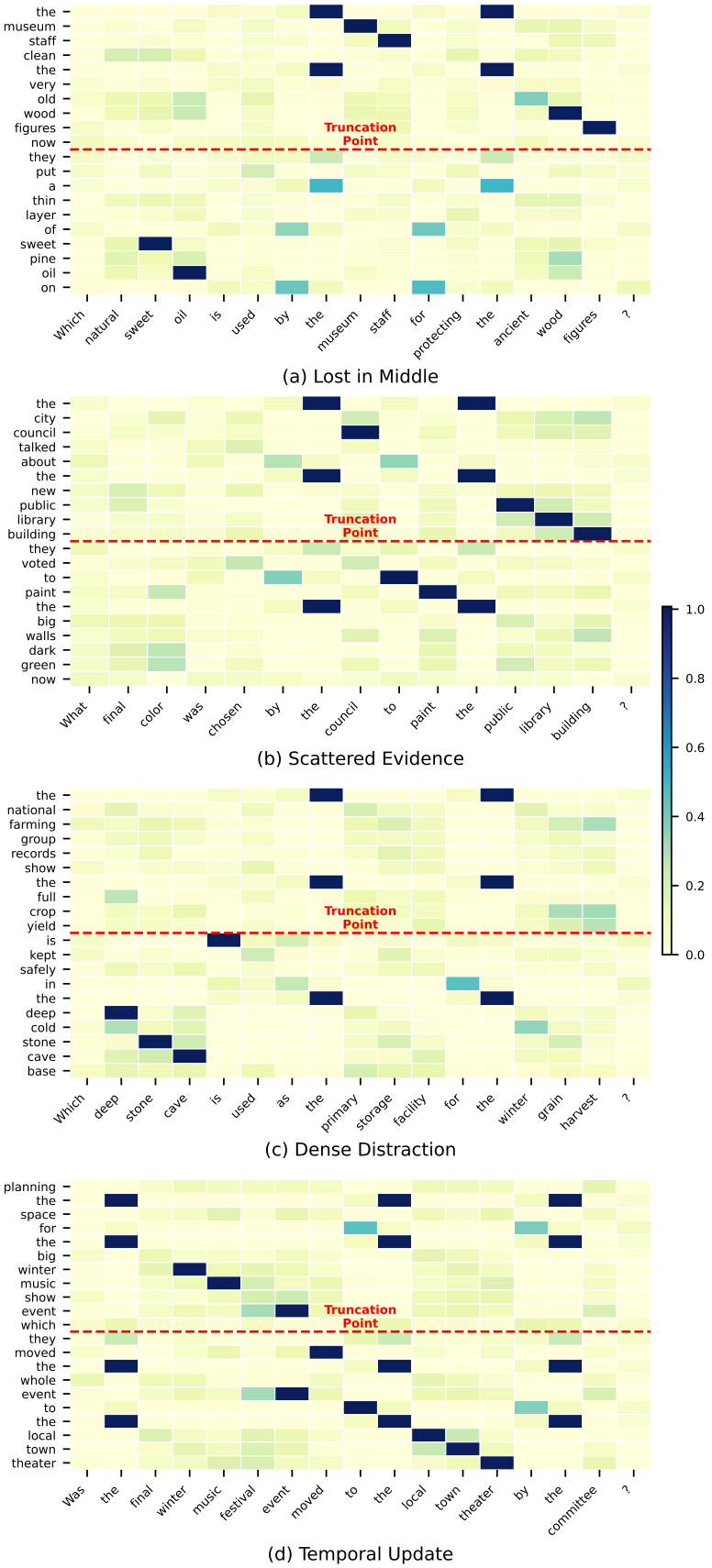

Figure 7: Token-level similarity heatmaps for four extended retrieval scenarios: (a) Lost in Middle, (b) Scattered Evidence, (c) Dense Distraction, and (d) Temporal Update. The red dashed line denotes the hard truncation point. Dark blue clusters indicate the highly activated core evidence successfully preserved by our ETC module.

Model	Params	Dim.	M.Task	M.Type	Clas.	Clus.	Pair.	Rank.	Retr.	STS	Sum.
MTEB (English)											
QZhou-Embedding	7B	3584	**75.97**	**69.52**	88.97	**61.65**	**92.43**	**51.77**	67.12	**91.65**	33.05
Qwen3-Embedding-8B	8B	4096	75.23	68.71	**90.43**	58.57	87.52	51.56	**69.44**	88.58	**34.83**
Qwen3-Embedding-4B	4B	2560	74.61	68.10	89.84	57.51	87.01	50.76	68.46	88.72	34.39
Qwen3-Embedding-0.6B	595M	1024	70.47	64.72	84.58	54.05	84.37	48.18	61.83	86.57	33.43
Jasper-TC-600M ($\rho = 0.50$)	595M	2048	74.75	68.46	90.35	59.44	90.15	50.60	66.19	88.79	33.66
Jasper-TC-600M ($\rho = 0.33$)	595M	2048	74.71	68.43	90.37	59.33	90.15	50.60	66.17	88.71	33.66
Jasper-TC-600M ($\rho = 0.20$)	595M	2048	74.58	68.33	90.32	59.22	90.15	50.59	65.77	88.75	33.53
Jasper-TC-600M ($\rho = 0.10$)	595M	2048	74.22	67.93	90.27	58.80	90.15	50.59	64.76	88.77	32.16
C-MTEB (Chinese)											
QZhou-Embedding-Zh	7B	1792	**78.52**	**80.29**	78.57	**81.76**	**95.49**	72.37	**80.24**	**73.33**	–
QZhou-Embedding	7B	3584	76.99	78.58	**79.99**	70.91	95.07	**74.85**	78.80	71.89	–
Qwen3-Embedding-8B	8B	4096	73.84	75.00	76.97	80.08	84.23	66.99	78.21	63.53	–
Qwen3-Embedding-4B	4B	2560	72.27	73.51	75.46	77.89	83.34	66.05	77.03	61.26	–
Qwen3-Embedding-0.6B	595M	1024	66.33	67.45	71.40	68.74	76.42	62.58	71.03	54.52	–
Jasper-TC-600M ($\rho = 0.50$)	595M	2048	73.51	75.00	77.72	77.45	85.38	69.95	75.64	63.88	–
Jasper-TC-600M ($\rho = 0.33$)	595M	2048	73.51	75.03	77.64	77.69	85.38	70.11	75.48	63.88	–
Jasper-TC-600M ($\rho = 0.20$)	595M	2048	73.38	74.92	77.60	77.47	85.38	70.06	75.16	63.88	–
Jasper-TC-600M ($\rho = 0.10$)	595M	2048	73.17	74.76	77.41	77.42	85.38	69.83	74.63	63.88	–

Table 7: Comprehensive results on MTEB (English) and C-MTEB (Chinese) benchmarks. Abbreviations: Params (Parameters), Dim. (Dimension), Clas. (Classification), Clus. (Clustering), Pair. (Pair Classification), Rank. (Reranking), Retr. (Retrieval), Sum. (Summarization). M.Task and M.Type denote Mean(Task) and Mean(TaskType), respectively. The rows highlighted in gray represent our best-performing configurations ($\rho = 0.50$). Best scores across all models within each respective benchmark are **bolded**. Note: C-MTEB does not include a Summarization task, indicated by '–'.

Model	Size	Dimension	Mean (Task)	Mean (Task Type)
stella_en_400M_v5	435M	4096	69.39	64.84
KaLM-embedding-mini-v2.5	494M	896	71.29	65.31
Qwen3-Embedding-0.6B	595M	1024	70.47	64.72
Jasper-Token-Compression-600M	600M	2048	74.75	68.46
jasper_en_vision_language_v1	1.5B	8960	71.41	66.65
F2LLM-1.7B	1.7B	2560	72.01	65.67
Qwen3-Embedding-4B	4B	2560	74.60	68.10
F2LLM-4B	4B	2560	73.67	67.29
QZhou-Embedding	7B	3584	75.97	69.52
LGAI-Embedding-Preview	7B	4096	74.12	68.40
Linq-Embed-Mistral	7B	4096	69.80	65.29
SFR-Embedding-Mistral	7B	4096	69.31	64.94
NV-Embed-v2	7B	4096	69.81	65.00
Qwen3-Embedding-8B	8B	4096	75.22	68.71
Seed 1.5-Embedding	-	2048	74.76	68.56
Seed 1.6-Embedding	-	2048	74.07	67.98
gemini-embedding-001	-	3072	73.30	67.67

Table 8: Results on the MTEB (English). Rows highlighted in light blue (QZhou-Embedding and Qwen3-Embedding-8B) correspond to the teacher models. The row highlighted in light orange (Qwen3-Embedding-0.6B) corresponds to the initialized student model. The row highlighted in gray represents our proposed Jasper-Token-Compression-600M model.

Model	Size	Dimension	Mean (Task)	Mean (Task Type)
Yinka	326M	1024	70.70	71.73
stella-mrl-large-zh-v3.5-1792d	326M	1792	68.60	69.30
ritrieve_zh_v1	326M	1792	72.71	73.85
xiaobu-embedding-v2	326M	768	72.36	73.48
Conan-embedding-v1	326M	768	72.50	73.65
zpoint_large_embedding_zh	326M	1792	71.81	72.82
KaLM-embedding-mini-v2.5	494M	896	70.89	72.43
Qwen3-Embedding-0.6B	595M	1024	66.33	67.45
Jasper-Token-Compression-600M	600M	2048	73.51	75.00
Youtu-Embedding	2B	2048	77.58	78.86
Qwen3-Embedding-4B	4B	2560	72.27	73.51
QZhou-Embedding-Zh	7B	1792	78.52	80.29
QZhou-Embedding	7B	3584	76.99	78.58
gte-Qwen2-7B-instruct	7B	3584	71.62	72.19
Qwen3-Embedding-8B	8B	4096	73.84	75.00
Seed1.5-Embedding	-	2048	74.87	76.01
Seed 1.6-embedding	-	2048	75.63	76.68
Conan-embedding-v2	-	3584	74.24	75.99
piccolo-large-zh-v2	-	1024	70.86	71.94

Table 9: Results on the C-MTEB (Chinese). Rows highlighted in light blue (QZhou-Embedding and Qwen3-Embedding-8B) correspond to the teacher models. The row highlighted in light orange (Qwen3-Embedding-0.6B) corresponds to the initialized student model. The row highlighted in gray represents our proposed Jasper-Token-Compression-600M model.

Task Type	Testset	Jasper-Token-Compression-600M
Classification	AmazonCounterfactualClassification	93.52
Classification	Banking77Classification	87.46
Classification	ToxicConversationsClassification	91.06
Classification	MassiveIntentClassification	85.29
Classification	MassiveScenarioClassification	90.85
Classification	TweetSentimentExtractionClassification	78.57
Classification	MTOPDomainClassification	98.97
Classification	ImdbClassification	97.11
Clustering	StackExchangeClusteringP2P.v2	53.54
Clustering	MedrxivClusteringS2S.v2	46.23
Clustering	ArXivHierarchicalClusteringP2P	66.10
Clustering	TwentyNewsgroupsClustering.v2	67.61
Clustering	BiorxivClusteringP2P.v2	51.96
Clustering	ArXivHierarchicalClustering S2S	63.87
Clustering	MedrxivClusteringP2P.v2	49.01
Clustering	StackExchangeClustering.v2	77.23
Pair Classification	TwitterURLCorpus	88.91
Pair Classification	SprintDuplicateQuestions	98.34
Pair Classification	TwitterSemEval2015	83.20
Reranking	MindSmallReranking	32.77
Reranking	AskUbuntuDupQuestions	68.44
Retrieval	TRECCOVID	89.92
Retrieval	Touche2020Retrieval.v3	67.94
Retrieval	SCIDOCS	25.51
Retrieval	HotpotQAHardNegatives	76.67
Retrieval	CQADupstackUnixRetrieval	59.13
Retrieval	CQADupstackGamingRetrieval	70.85
Retrieval	FiQA 2018	53.89
Retrieval	FEVERHardNegatives	93.66
Retrieval	ClimateFEVERHardNegatives	46.03
Retrieval	ArguAna	78.28
STS	STS13	93.35
STS	STS14	89.63
STS	STS12	86.07
STS	STS17	93.42
STS	STSBenchmark	92.89
STS	STS22.v2	73.30
STS	SICK-R	85.30
STS	BIOSSES	91.68
STS	STS15	93.47
Summarization	SummEvalSummarization.v2	33.66

Table 10: Results for each dataset in the English portion of the MTEB benchmark.

Task Type	Testset	Jasper-Token-Compression-600M
Classification	JDReview	88.39
Classification	OnlineShopping	94.56
Classification	TNews	57.04
Classification	MultilingualSentiment	81.23
Classification	Waimai	90.05
Classification	IFlyTek	55.03
Clustering	CLSClusteringP2P	69.94
Clustering	CLSClusteringS2S	67.98
Clustering	ThuNewsClusteringS2S	84.07
Clustering	ThuNewsClusteringP2P	87.82
Pair Classification	Ocnli	83.00
Pair Classification	Cmnli	87.76
Reranking	CMedQAv2 - reranking	88.59
Reranking	T2Reranking	67.54
Reranking	CMedQAv1 - reranking	88.44
Reranking	MMarcoReranking	35.24
Retrieval	VideoRetrieval	76.10
Retrieval	MMarcoRetrieval	83.93
Retrieval	EcomRetrieval	68.93
Retrieval	CmedqaRetrieval	47.54
Retrieval	T2Retrieval	86.14
Retrieval	CovidRetrieval	87.52
Retrieval	MedicalRetrieval	65.37
Retrieval	DuRetrieval	89.54
STS	QBQTC	47.04
STS	AFQMC	56.07
STS	ATEC	53.87
STS	LCQMC	80.68
STS	STSB	88.66
STS	BQ	72.64
STS	PAWSX	48.17

Table 11: Results for each dataset in the Chinese portion of the MTEB benchmark.

Category	Dataset	Domain / Text Type	Main Metric
Retrieval	ArguAna	Argument Retrieval (Long Query/Doc)	nDCG@10
	SCIDOCS	Scientific Literature Citation	nDCG@10
	Hotpot QA Hard Negatives	Multi-hop Question Answering	nDCG@10
	TRECCOVID	Medical/Pandemic Search	nDCG@10
	FIQA2018	Financial Opinion Mining & QA	nDCG@10
	FEVERHardNegatives	Fact Extraction and Verification	nDCG@10
Classification	Banking77Classification	Customer Service Intent Detection	Accuracy
	ImdbClassification	Movie Reviews (Long Text Sentiment)	Accuracy
	AmazonCounterfactualClassification	Counterfactual Detection (Product Reviews)	Accuracy
	ToxicConversationsClassification	Social Media Comment Moderation	Accuracy
Clustering	TwentyNewsgroupsClustering.v2	News Articles (Classic Benchmark)	V-Measure
	ArXivHierarchicalClusteringP2P	Paragraph-to-Paragraph Academic Texts	V-Measure
	MedrxivClusteringP2P.v2	Medical Research Paragraphs	V-Measure
	StackExchangeClusteringP2P.v2	Technical Community QA Threads	V-Measure
STS	STSBenchmark	Standard Semantic Textual Similarity	Spearman corr.
	SICK-R	Entailment and Relatedness	Spearman corr.
	STS12	Historical Similarity Baseline (2012)	Spearman corr.
	STS14	Historical Similarity Baseline (2014)	Spearman corr.
	STS15	Historical Similarity Baseline (2015)	Spearman corr.
Reranking	AskUbuntuDupQuestions	Technical Forum Duplicate Re-ranking	MAP / MRR
Pair Classification	SprintDuplicateQuestions	Community QA Duplicate Detection	Cosine F1
	TwitterSemEval2015	Paraphrase Detection (Social Media)	Cosine F1
Summarization	SummEvalSummarization.v2	Text Summarization Evaluation	Spearman corr.

Table 12: Summary of the streamlined MTEB subset (23 datasets) utilized in our ablation studies. This subset spans all 7 core task categories to balance evaluation comprehensiveness and computational efficiency.

Scenario	Query	Document Snippet (Omitted for brevity)	Mechanism Highlight
1. Extreme Tail	What was the main reason discovered for the sudden bridge collapse during the heavy storm?	Following the unprecedented structural disaster... *[... 500 words of meteorological and traffic logs omitted ...]* After testing the chemical makeup of the broken pieces, the dedicated researchers ultimately found that the **collapse** was caused by **deep red rust** which had slowly compromised the core anchor integrity over several decades.	Preserves critical evidence at the absolute sequence boundary, preventing hard-truncation failure.
2. Lost in Middle	Which natural sweet oil is used by the museum staff for protecting the ancient wood figures?	The global cultural heritage foundation recently announced a major upcoming exhibition... *[... 300 words of logistic preparations omitted ...]* Instead, drawing inspiration from traditional preservation methods used by ancient boat builders, they put a thin layer of **sweet pine oil** on across the entire exterior surface to create a breathable hydrophobic barrier. *[... 300 words of security measures omitted ...]*	Successfully isolates and retains core facts buried deep within homogenous, redundant context.
3. Scattered	What final color was chosen by the council to paint the public library building?	The primary agenda item was heavily debated as the city council talked about the new public library building... *[... 500 words detailing architectural disputes and budgets omitted ...]* After reviewing all the public survey results regarding the visual integration of the facility, they voted to **paint** the big walls **dark green** now to ensure perfect harmony with the surrounding ancient oak trees.	Implicitly links the target entity (front) with its corresponding temporal state/value (tail) in the compressed space.
4. Distraction	Which deep stone cave is used as the primary storage facility for the winter grain harvest?	During the early spring months, vegetables are temporarily housed in the large wooden barn... the massive corn and wheat surplus is deposited into the towering steel silo... *[... 400 words of transportation logs omitted ...]* To protect the most vital nutritional resources from freezing temperatures and external threats, the massive winter grain harvest is kept safely in the **deep cold stone cave** base located deep beneath the northern mountain range.	Resists lexical overlapping traps (e.g., "wooden barn", "steel silo", "brick warehouse" appearing in wrong contexts) to find the true semantic match.
5. Temporal Update	Was the final winter music festival event moved to the local town theater by the committee?	The board proudly announced... that the gathering would be hosted at the massive outdoor central stadium. *[... 400 words detailing stadium stage preparations omitted ...]* Realizing that proceeding with the original outdoor plan would be incredibly dangerous due to the blizzard, they **moved** the whole event to the local **town theater** to guarantee absolute safety.	Captures dynamic state changes and prevents anchoring bias toward obsolete or superseded information mentioned early in the document.

Table 13: Detailed query and context pairs for the extended qualitative analysis. Bold text indicates the token spans that maintained the highest attention/similarity signals in the ETC compressed representations, successfully bypassing positional, lexical, and temporal biases.