

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

Haining An^{1,*}, Xingbing Chen^{2,*}, Wenbo Lyu³, Dun Zhang⁴,
Ziyang Zeng⁵, Zhenxiang Zhao³, Rui Song¹, Hao Xu^{1,†}

¹ College of Computer Science and Technology, Jilin University, Changchun, China

² University of Chinese Academy of Sciences, Beijing, China

³ Beihang University, Beijing, China

⁴ Prior Shape, Shanghai, China

⁵ Beijing University of Posts and Telecommunications, Beijing, China

*Equal contribution. †Corresponding author: xuhao@jlu.edu.cn

Abstract

Deploying text embedding models under resource constraints is challenged by the sequence-length-dependent computational cost of self-attention. However, existing sequence-reduction strategies remain predominantly static and offer limited flexibility across inference budgets. We propose **Jasper-Flash**, an inference-scalable text embedding framework combining **Elastic Token Compression (ETC)** with **Compression-Adaptive Progressive Distillation (CAPD)**. ETC uses a length threshold and a user-selected compression ratio to adjust sequence length on the fly, enabling flexible latency–quality trade-offs within a single model. To mitigate the optimization difficulty introduced by lossy sequence reduction, CAPD integrates heterogeneous multi-teacher distillation with a progressive curriculum over fixed and dynamically sampled compression ratios. We instantiate Jasper-Flash as **Jasper-Token-Compression-600M**, a bilingual text embedding model. Experiments on MTEB and C-MTEB demonstrate competitive performance across compression ratios and substantial reductions in encoding latency, particularly for longer inputs. The code and model are available at <https://huggingface.co/infgrad/Jasper-Token-Compression-600M>.

1 Introduction

While text embedding models have profoundly advanced downstream tasks such as information retrieval and document clustering (Muennighoff et al., 2023), deploying them in resource-constrained scenarios remains highly challenging. 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).

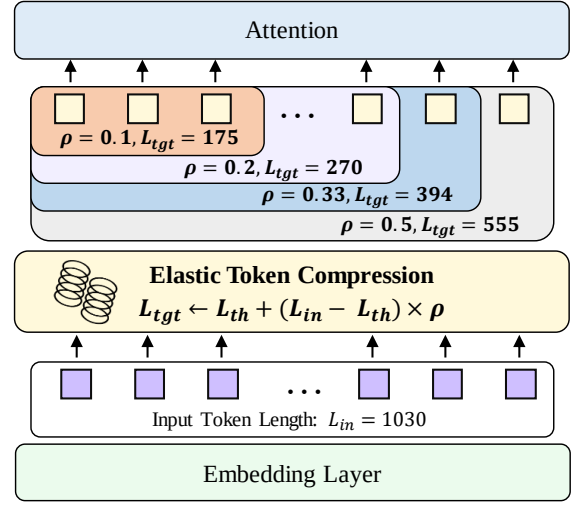


Figure 1: Illustration of the Elastic Token Compression. Positioned between the embedding and attention layers, this module utilizes a target compression ratio ρ and a length threshold (default $L_{th} = 80$) to transform a single input into multi-scale elastic target sequences.

In this work, we focus on the sequence-length-dependent computational bottleneck, which leads to substantial inference latency as input sequences grow longer.

To mitigate the sequence-length-dependent computational bottleneck, token-level sequence reduction has been explored through various techniques (Goyal et al., 2020; Dai et al., 2020; Yun et al., 2023). However, even amidst recent advancements in efficient sequence modeling (Gu and Dao, 2024; 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 bal-

ance 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 elasticity along the *sequence dimension* presents a natural 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. Because aggressive sequence reduction may degrade representation quality and complicate optimization, CAPD builds on our earlier study of multi-teacher distillation (Zhang et al., 2024) by combining heterogeneous teacher supervision with a four-stage progressive curriculum, first establishing a semantic foundation and then adapting the student to varying compression constraints.

Building upon our framework, we introduce Jasper-TC-600M, a bilingual text embedding model for which we previously released an early technical report (Zhang et al., 2025a). Experiments demonstrate strong inference elasticity: at $\rho = 0.10$, it reduces 1,024-token encoding latency from the baseline’s 24.24 ms to 4.48 ms while retaining a competitive MTEB Mean(Task) score of 74.22. Compared with its 0.6B initialization baseline, the CAPD training pipeline improves MTEB Mean(Task) from 70.47 to 74.75 at $\rho = 0.50$, primarily through progressive multi-teacher distillation and retrieval adaptation. Across compression ratios from 0.50 to 0.10, ETC preserves most of this representation quality, with Mean(Task) decreasing by only 0.53 points while providing multiple latency–quality operating points.

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 progressive training paradigm that first establishes a

strong semantic representation space through heterogeneous multi-teacher distillation and then adapts it to fixed and dynamically sampled compression ratios.

- **Jasper-Token-Compression-600M:** We train and present a bilingual inference-scalable text embedding model that substantially improves over its initialization baseline and maintains competitive performance across multiple compression ratios.

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 sequence reduction, 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 before introducing sequence compression. To integrate supervision from multiple representation spaces, we construct a unified distillation target by aggregating the similarity structures induced by different teachers, rather than constraining the student to reproduce a single teacher-specific geometry.

Mathematically, given independently ℓ_2 -normalized sub-vectors u_i and v_i from n teacher models, their globally normalized concatenated representations, defined as $U = \text{Norm}(u_1 \parallel \dots \parallel u_n)$

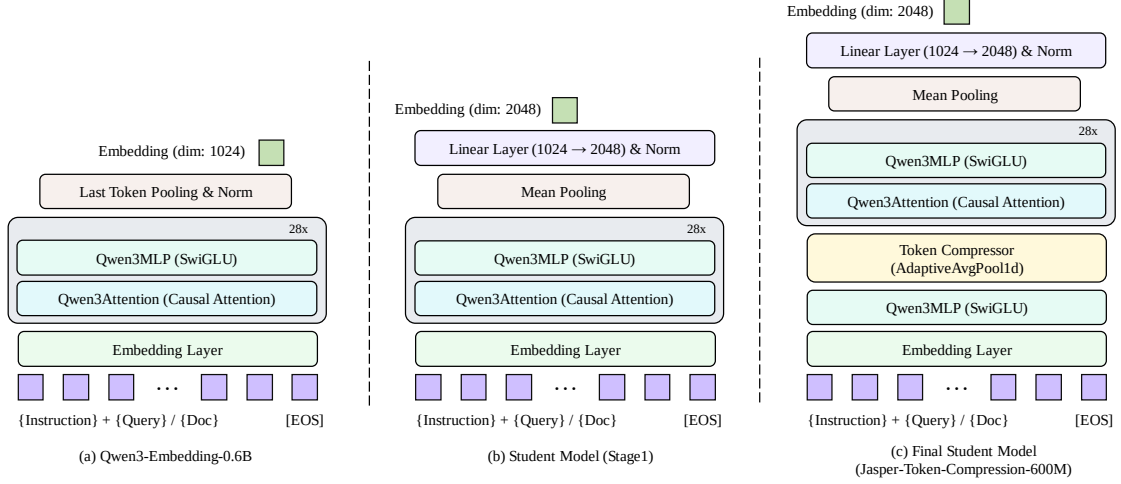


Figure 2: **Model architecture evolution.** (a) Qwen3-Embedding-0.6B, the original base model. (b) Student Model (Stage 1), adapted for unified teacher distillation by replacing last-token pooling with mean pooling and introducing a linear projection layer. (c) Final Student Model (Jasper-Token-Compression-600M),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

and $V = \text{Norm}(v_1 \parallel \dots \parallel v_n)$, satisfy:

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

Because U , V , u_i , and v_i are all ℓ_2 -normalized, their dot products correspond to cosine similarities. Equation 1 therefore establishes that the cosine similarity in the fused representation space is exactly the arithmetic mean of the cosine similarities produced by the individual teachers. A concise proof is provided in Appendix A.

Based on this equal-weight similarity aggregation, we introduce a heterogeneous multi-teacher distillation paradigm that exploits *capability complementarity*. Specifically, teachers with different task strengths provide complementary supervision, allowing the fused target to reflect multiple similarity structures instead of depending exclusively on a single teacher. Assuming the output vector of the i -th teacher model is E_{t_i} , the target teacher representation E_t is formulated as:

$$E_t = \text{Norm}\left(\text{Norm}(E_{t_1}) \parallel \dots \parallel \text{Norm}(E_{t_n})\right).$$

To exploit capability complementarity while maintaining computational efficiency, we select two models with different task strengths from the MTEB leaderboard: Qwen3-Embedding-8B (Zhang et al., 2025b), which excels in asymmetric retrieval tasks (69.44), and QZhou-Embedding-7B (Yu et al., 2025a), which performs strongly in STS tasks (91.65). To prevent direct concatenation

from producing an excessively high-dimensional target vector, we reduce both teacher representations to 1024 dimensions. Leveraging its native Matryoshka Representation Learning capability, we directly extract the first 1024 dimensions of Qwen3-Embedding-8B, denoted as E_{qwen} . For QZhou-Embedding-7B, we reduce its 3072-dimensional representation to 1024 dimensions using parameter-free block-wise summation, denoted as E_{qzhou} . This simple transformation matches the dimensionality of the two teacher representations without introducing an additional learnable projection layer. The resulting 2048-dimensional dual-teacher fusion vector is formulated as:

$$E_t = \text{Norm}\left(\text{Norm}(E_{qwen}) \parallel \text{Norm}(E_{qzhou})\right).$$

For the student model, we adopt Qwen3-Embedding-0.6B (Figure 2(a)) as the initialization backbone, which produces a 1024-dimensional output. To 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 project the student features to 2048 dimensions. Following ℓ_2 normalization to obtain the student vector E_s , the learning objective is defined as the cosine distance between E_s and E_t :

$$\mathcal{L}_{\cosine} = 1 - E_s \cdot E_t. \quad (2)$$

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}

```

1: if  $L_{in} \leq L_{th}$  then
2:    $L_{tgt} \leftarrow L_{in}$   $\triangleright$  No compression below the threshold
3: else
4:    $L_{tgt} \leftarrow L_{th} + (L_{in} - L_{th}) \times \rho$ 
5: end if
6: return  $L_{tgt}$ 

```

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 token embeddings first undergo a token-wise feature transformation via a randomly initialized SwiGLU network (Shazeer, 2020) (i.e., Qwen3MLP), which preserves the sequence length and hidden dimension 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 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

Algorithm 2 Compression Ratio Sampling

Input: Baseline compression ratio ρ

Output: Sampled compression ratio r

```

1:  $p \leftarrow \text{Uniform}(0, 1)$ 
2: if  $p < 0.2$  then
3:    $r \leftarrow \text{Uniform}(0.1, \rho)$ 
4: else if  $p < 0.6$  then
5:    $r \leftarrow \rho$ 
6: else if  $p < 0.8$  then
7:    $r \leftarrow \text{Uniform}(\rho, 2\rho)$ 
8: else
9:    $r \leftarrow \text{Uniform}(2\rho, 1.0)$ 
10: end if
11: return  $r$ 

```

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 adaptation for retrieval via contrastive learning, instantiate $\mathcal{L}_{task}$ using the InfoNCE loss (van den 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 at a 1:1 ratio between English and Chinese, followed by 2.5 million fine-tuning instances in the final stage to elicit specialized retrieval capabilities. We set the maximum input sequence length to 1,030 tokens throughout training and all downstream effectiveness evaluations. Detailed dataset compositions and comprehensive hyperparameter settings are provided in Appendix B and Appendix C. For elastic token compression, we empirically set the length threshold L_{th} to 80 and the baseline compression ratio ρ_{base} to 0.33. Under this configuration, inputs no longer than the threshold remain uncompressed, while inputs exceeding the threshold are reduced according to the specified compression ratio. 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 D.

At $\rho = 0.50$, Jasper-TC-600M achieves a Mean(Task) score of 74.75 on English MTEB, improving upon its initialization model, Qwen3-Embedding-0.6B, by 4.28 points at the same backbone scale. It also slightly outperforms the substantially larger Qwen3-Embedding-4B model, which obtains 74.61, demonstrating that the proposed framework produces a strong embedding model without increasing the backbone size.

More importantly, Jasper-TC-600M maintains stable performance across its inference-time compression ratios. Reducing ρ from 0.50 to 0.10 decreases the Mean(Task) score by only 0.53 points, from 74.75 to 74.22. This result demonstrates that ETC provides multiple sequence-length operating points within a single model while preserving most

of its representation quality. The respective contributions of representation learning and compression adaptation are further examined through the stage-wise analysis in Section 4.3.

3.3 Inference Performance & Efficiency

To quantify the efficiency gains of ETC, we systematically evaluate encoding latency across various input lengths on RTX 4090 GPUs, as reported in Table 2. ETC provides increasingly substantial efficiency gains as the measured input length increases. At input lengths of 1,024 and 2,048 tokens, latency is 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 acceleration is facilitated by the dense matrix operations used in ETC, which are well suited to hardware parallelization. All downstream effectiveness results follow the maximum-length setting of 1,030 tokens, while latency measurements at additional input lengths are reported to characterize computational scaling.

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 evaluate the effect of capability complementarity, we construct a single-teacher baseline supervised only by Qwen3-Embedding-8B while retaining the same student architecture, training data, and optimization budget. This controlled comparison examines whether incorporating an additional teacher with complementary strengths provides a more effective distillation target than Qwen-only supervision. As shown in Table 3, heterogeneous teacher fusion improves the MTEB-subset Mean(Task) score from 68.91 to 70.11. In particular, the heterogeneous configuration improves Classification, Clustering, Pair Classification, Retrieval, STS, and Summarization, while the Qwen-only baseline remains slightly stronger on Reranking. These results support the empirical benefit of capability complementarity under our training setting: the additional teacher improves several task

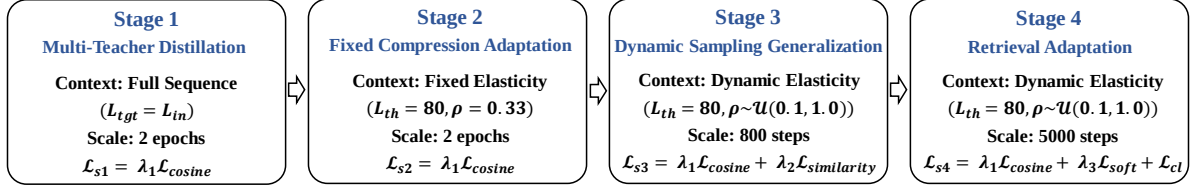


Figure 3: Overview of the four-stage CAPD training pipeline.

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**.

categories without requiring separate task-specific student models.

4.2 Ablation on Token Compression Module

To evaluate ETC under a controlled setting, we compare it with **Hard Truncation**, **Token Pruning**, **Token Merging (ToMe)**, and **Learnable Conv1D** at $\rho = 0.33$. All variants are initialized from the same Stage-1 checkpoint and adapted using the same one-epoch Stage-2 training budget. All variants retain the same preceding MLP and replace only the sequence-reduction operator. Hard Truncation, ℓ_2 -norm-based Token Pruning, and similarity-based Token Merging are parameter-free operators, whereas Conv1D introduces newly initialized trainable parameters. For all variants, the shared backbone, projection layer, and MLP are updated end-to-end during adaptation.

As shown in Table 3, our MLP-AdaptiveAvgPool configuration achieves 66.28 Mean(Task), compared with 51.62 for Hard Truncation, 55.38 for Token Pruning, 54.55 for Token Merging, and 46.90 for Learnable Conv1D. The differences are particularly evident on Retrieval, where ETC obtains 49.58, while the alternative operators range from 10.21 to 29.78. These results demonstrate the effectiveness of ETC under the matched one-epoch adaptation setting, but do not establish the performance ceiling of the alternative

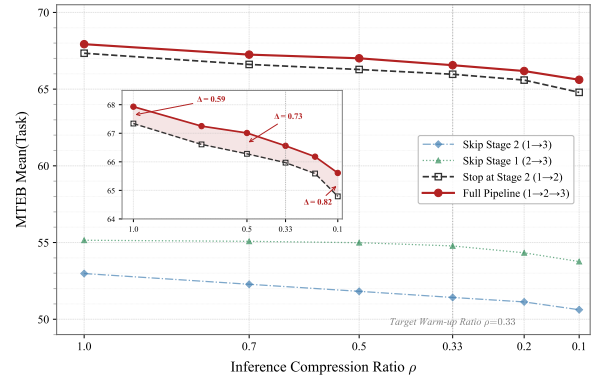


Figure 4: Ablation study of the progressive distillation pipeline across varying compression ratios.

operators under longer or individually optimized training schedules. In a separately conducted operator-level latency benchmark, ETC requires 22.57 ms at an input length of 4096, compared with 22.83 ms for Hard Truncation and 55.25 ms for Token Merging. Complete results from this separate benchmark are provided in Appendix F.

4.3 Ablation on Progressive Distillation Pipeline

Figure 4 compares the complete progressive pipeline with variants that stop after Stage 2 or bypass individual adaptation stages. Compared with stopping at fixed-ratio training, the complete pipeline maintains a more stable performance pro-

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
<i>Ablation on Heterogeneous Multi-Teacher Distillation (MTEB-Subset)</i>										
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
<i>Ablation on Token Compression Module (MTEB-Subset)</i>										
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
<i>Ablation of Retrieval Adaptation (Full MTEB)</i>										
w/o Retrieval Adaptation	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.

file across inference compression ratios, suggesting that dynamic-ratio training improves robustness to varying sequence budgets. Skipping Stage 2 produces more pronounced degradation under aggressive compression, supporting the use of fixed-ratio adaptation before dynamic compression training.

The controlled stage-wise results in Appendix F further clarify the respective roles of distillation and compression adaptation. Stage 1 establishes a strong uncompressed representation space with an MTEB Mean(Task) score of 74.73. At $\rho = 0.33$, fixed-compression training in Stage 2 achieves 74.34, while subsequent dynamic-ratio training in Stage 3 improves the score to 74.67. The final difference of only 0.06 points relative to the uncompressed Stage 1 indicates that ETC primarily preserves the representation quality established through distillation while enabling elastic sequence reduction.

We additionally evaluate a **Skip Stage 1** variant that directly performs the same two-epoch compres-

sion training at $\rho = 0.33$. Its Mean(Task) score decreases to 56.36, below the initialization baseline of 70.47, while Retrieval drops from 61.83 to 28.15. Under our training setting, this substantial degradation suggests that simultaneously learning multi-teacher representation alignment and lossy sequence compression creates considerable optimization difficulty. Together, these results support separating initial representation alignment from subsequent compression adaptation; complete results are reported in Appendix F.

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 that $L_{th} = 1$ destructively dilutes short

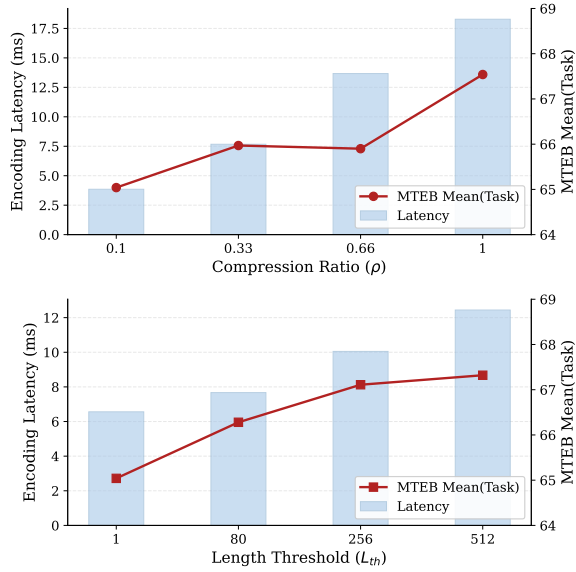


Figure 5: Parameter sensitivity analysis. (a) Impact of the compression ratio ρ ($L_{th} = 80$). (b) Impact of the length threshold L_{th} ($\rho = 0.33$).

queries, whereas $L_{th} = 80$ sharply recovers performance with a negligible latency penalty; further relaxation (e.g., 256) inflates latency for marginal gains. These results support our configuration ($\rho = 0.33, L_{th} = 80$) as an effective trade-off that reduces sequence-length-dependent computational overhead within the evaluated setting while safeguarding short-query semantics.

4.5 Ablation of Retrieval Adaptation

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 changes only marginally, the improvement is concentrated in Retrieval, consistent with the targeted objective of Stage 4. This result suggests that Stage 4 improves retrieval-oriented representations without substantially degrading the broader 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 failure, our **ETC** module maintains robust query alignment. The token-level heatmap reveals that

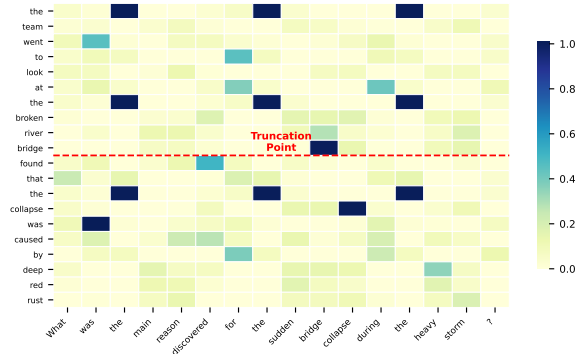


Figure 6: Token-level similarity heatmap (vertical axis: compressed document tokens; horizontal axis: query tokens). It illustrates how ETC’s semantic distillation preserves core tail evidence beyond the hard truncation point (red dashed line).

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 G.

6 Related Work

6.1 Text Embeddings and Knowledge Distillation

Text embedding research has progressed from sentence-level representation learning with models such as SBERT (Reimers and Gurevych, 2019), through weakly supervised contrastive pre-training exemplified by E5 (Wang et al., 2022), to recent LLM-based embedding models including Qwen3 (Zhang et al., 2025b), QZhou (Yu et al., 2025a), and KaLM (Zhao et al., 2026). Knowledge distillation has been widely used to obtain compact language encoders, including TinyBERT (Jiao et al., 2020), MobileBERT (Sun et al., 2020), and MiniLM (Wang et al., 2020). Gecko further combines LLM-generated data and relabeling to distill versatile text embeddings (Lee et al., 2024). Multi-teacher distillation has also been studied as a means of combining complementary representation spaces (Foromont et al., 2025). In contrast to these approaches, CAPD couples heterogeneous teacher supervision with a progressive curriculum that first establishes the student representation space and then adapts it to fixed and dynamic sequence compression.

6.2 Efficient Inference and Token Compression

The quadratic complexity of Transformer self-attention motivates both hardware-aware attention kernels such as FlashAttention (Dao et al., 2022) and alternative linear-time architectures such as Mamba (Gu and Dao, 2024). Within standard Transformer encoders, PoWER-BERT progressively eliminates word vectors (Goyal et al., 2020), Funnel-Transformer compresses intermediate hidden sequences (Dai et al., 2020), and Focus on the Core combines token pruning and compression (Yun et al., 2023). Token merging methods such as ToMe combine similar token representations to reduce sequence length (Bolya et al., 2023). ETC differs in targeting text embedding inference with a user-selected compression ratio and a shared model that exposes multiple latency–quality operating points without retraining.

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 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 between our 0.6B student and the 8B retrieval-specialized teacher. We attribute this gap to three main factors: (1) although heterogeneous teacher fusion exploits capability complementarity, our current equal-weight construction cannot adapt the contribution of each teacher to individual inputs or

downstream tasks and may therefore underemphasize the retrieval-specialized teacher in some cases; (2) the 0.6B student has substantially lower representational capacity than the 7B and 8B teachers, limiting its ability to reproduce all of their task-specific knowledge; and (3) while the average-pooling compression module efficiently preserves global semantic information, it may smooth over fine-grained entity-level signals required for exact-match asymmetric retrieval.

Granularity of Token Compression. While our MLP-AvgPool compression module achieves sequence-level elasticity through dynamic sequence reduction 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. We set the maximum input sequence length to 1,030 tokens throughout training and all downstream effectiveness evaluations. Although the latency analysis additionally considers other input lengths to characterize the computational scaling behavior of ETC, our empirical conclusions regarding representation quality are based on and limited to the 1,030-token evaluation setting.

Potential Data Contamination. Given the use of large-scale Internet data, unintended overlap between our training corpus and downstream evaluation datasets cannot be fully ruled out.

Future work will explore more selective and context-adaptive compression mechanisms to better preserve fine-grained token-level information under aggressive compression.

Acknowledgments

We thank the anonymous reviewers and the area chair for their constructive feedback. This work is supported by the National Natural Science Foundation of China (NSFC): “Research on Understanding Ancient Characters Based on Multi-modal Large Models” (No. 62476111), the Department of Science and Technology of Jilin Province: “Research and Application of Key Technologies for Cognitive Intelligence-Oriented Smart Education” (No.

20230201086GX), the Science and Technology Bureau of Changchun: “AI + Education: Large Model for Educational Evaluation” (No. 24GNYZ11), and the Industry-University-Research Innovation Fund for Chinese Universities: “Research and Application of an Integrated Teaching Model for Human centered Artificial Intelligence” (No. 2022XF017).

References

- Baidu DuReader Team. 2021. [DuReader Checklist](#). GitHub repository.
- Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2016. [MS MARCO: A human generated machine reading comprehension dataset](#). *Preprint*, arXiv:1611.09268.
- Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, and 3 others. 2025. [SmolLM2: When smol goes big – data-centric training of a small language model](#). *Preprint*, arXiv:2502.02737.
- Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. 2023. [Token merging: Your ViT but faster](#). In *International Conference on Learning Representations*.
- Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2021. [mMARCO: A multilingual version of the MS MARCO passage ranking dataset](#). *Preprint*, arXiv:2108.13897.
- Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [M3-Embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). *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](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 4271–4282. Curran Associates, Inc.
- Tri Dao. 2024. [FlashAttention-2: Faster attention with better parallelism and work partitioning](#). In *The Twelfth International Conference on Learning Representations*.
- Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. [FlashAttention: Fast and memory-efficient exact attention with IO-awareness](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 16344–16359. Curran Associates, Inc.
- Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. [ELI5: Long form question answering](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.
- Philippe Formont, Maxime Darrin, Banafsheh Karimian, Jackie C. K. Cheung, Eric Granger, Ismail Ben Ayed, Mohammadhadi Shateri, and Pablo Piantanida. 2025. [Learning task-agnostic representations through multi-teacher distillation](#). In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. [The pile: An 800gb dataset of diverse text for language modeling](#). *Preprint*, arXiv:2101.00027.
- Saurabh Goyal, Anamitra Roy Choudhury, Saurabh Raje, Venkatesan Chakaravarthy, Yogish Sabharwal, and Ashish Verma. 2020. [PoWER-BERT: Accelerating BERT inference via progressive word-vector elimination](#). In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3690–3699. PMLR.
- Albert Gu and Tri Dao. 2024. [Mamba: Linear-time sequence modeling with selective state spaces](#). In *Proceedings of the First Conference on Language Modeling*. OpenReview.
- Mansi Gupta, Nitish Kulkarni, Raghuveer Chanda, Anirudha Rayasam, and Zachary C. Lipton. 2019. [AmazonQA: A review-based question answering task](#). In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 4996–5002. International Joint Conferences on Artificial Intelligence Organization.
- Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2018. [DuReader: A chinese machine reading comprehension dataset from real-world applications](#). In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 37–46. Association for Computational Linguistics.
- Doris Hoogeveen, Karin M. Verspoor, and Timothy Baldwin. 2015. [CQADupStack: A benchmark data set for community question-answering research](#). In *Proceedings of the 20th Australasian Document Computing Symposium, ADCS ’15*, pages 3:1–3:8, New York, NY, USA. Association for Computing Machinery.

- Yupeng Hou, Jiacheng Li, Xiangjun Fu, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2026. [Bridging language and items for retrieval and recommendation: Benchmarking LLMs as semantic encoders](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3251–3265, San Diego, California, United States. Association for Computational Linguistics.
- Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. [TinyBERT: Distilling BERT for natural language understanding](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.
- Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. [TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.
- Diederik P. Kingma and Jimmy Ba. 2014. [Adam: A method for stochastic optimization](#). *Preprint*, arXiv:1412.6980. Published at the 3rd International Conference on Learning Representations, ICLR 2015.
- Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. 2022. [Metryoshka representation learning](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 30233–30249. Curran Associates, Inc.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. [Natural questions: A benchmark for question answering research](#). *Transactions of the Association for Computational Linguistics*, 7:452–466.
- Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R. Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, Yi Luan, Sai Meher Karthik Duddu, Gustavo Hernandez Abrego, Weiqiang Shi, Nithi Gupta, Aditya Kusupati, Prateek Jain, Siddhartha Reddy Jonnalagadda, Ming-Wei Chang, and Iftexhar Naim. 2024. [Gecko: Versatile text embeddings distilled from large language models](#). *Preprint*, arXiv:2403.20327.
- Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. [PAQ: 65 million probably-asked questions and what you can do with them](#). *Transactions of the Association for Computational Linguistics*, 9:1098–1115.
- Jijie Li, Li Du, Hanyu Zhao, Bo-wen Zhang, Liangdong Wang, Boyan Gao, Guang Liu, and Yonghua Lin. 2025. [Infinity Instruct: Scaling instruction selection and synthesis to enhance language models](#). *Preprint*, arXiv:2506.11116.
- Peng Li, Wei Li, Zhengyan He, Xuguang Wang, Ying Cao, Jie Zhou, and Wei Xu. 2016. [Dataset and neural recurrent sequence labeling model for open-domain factoid question answering](#). *Preprint*, arXiv:1607.06275.
- Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. [S2ORC: The semantic scholar open research corpus](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.
- Dingkun Long, Qiong Gao, Kuan Zou, Guangwei Xu, Pengjun Xie, Ruijie Guo, Jian Xu, Guanjuan Jiang, Luxi Xing, and Ping Yang. 2022. [Multi-CPR: A multi domain chinese dataset for passage retrieval](#). In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3046–3056. Association for Computing Machinery.
- 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. 2023. [MTEB: Massive text embedding benchmark](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037. Association for Computational Linguistics.
- Zhijie Nie, Zhangchi Feng, Mingxin Li, Cunwang Zhang, Yanzhao Zhang, Dingkun Long, and Richong Zhang. 2024. [When text embedding meets large language model: A comprehensive survey](#). *Preprint*, arXiv:2412.09165.
- Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. 2024. [The FineWeb datasets: Decanting the web for the finest text data at scale](#). In *Advances in Neural Information Processing Systems*, volume 37. Curran Associates, Inc.
- Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. [Know what you don’t know: Unanswerable questions for SQuAD](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*

- and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Sentence Transformers. 2021. [Embedding training data](#). Hugging Face dataset repository.
- Noam Shazeer. 2020. [GLU variants improve transformer](#). *Preprint*, arXiv:2002.05202.
- Xiaofeng Shi, Lulu Zhao, Hua Zhou, and Donglin Hao. 2024a. [IndustryCorpus2](#).
- Xiaofeng Shi, Lulu Zhao, Hua Zhou, Donglin Hao, and Yonghua Lin. 2024b. [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](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2158–2170, Online. Association for Computational Linguistics.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. [Representation learning with contrastive predictive coding](#). *Preprint*, arXiv:1807.03748.
- 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, pages 5998–6008. Curran Associates, Inc.
- 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](#). *Preprint*, arXiv:2212.03533.
- 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](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788. Curran Associates, Inc.
- Xiaohui Xie, Qian Dong, Bingning Wang, Feiyang Lv, Ting Yao, Weinan Gan, Zhijing Wu, Xiangsheng Li, Haitao Li, Yiqun Liu, and Jin Ma. 2023. [T2Ranking: A large-scale chinese benchmark for passage ranking](#). In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2681–2690. Association for Computing Machinery.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. [HotpotQA: A dataset for diverse, explainable multi-hop question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.
- Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024. [MetaMath: Bootstrap your own mathematical questions for large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Peng Yu, En Xu, Bin Chen, Haibiao Chen, and Yinfei Xu. 2025a. [QZhou-Embedding technical report](#). *Preprint*, arXiv:2508.21632.
- Yijiong Yu, Ziyun Dai, Zekun Wang, Wei Wang, Ran Chen, and Ji Pei. 2025b. [OpenCSG Chinese Corpus: A series of high-quality chinese datasets for LLM training](#). *Preprint*, arXiv:2501.08197.
- Sha Yuan, Hanyu Zhao, Zhengxiao Du, Ming Ding, Xiao Liu, Yukuo Cen, Xu Zou, Zhilin Yang, and Jie Tang. 2021. [WuDaoCorpora: A super large-scale chinese corpora for pre-training language models](#). *AI Open*, 2:65–68.
- 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, Singapore. Association for Computational Linguistics.
- Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2024. [Jasper and stella: Distillation of SOTA embedding models](#). *Preprint*, arXiv:2412.19048.
- Dun Zhang, Ziyang Zeng, Yudong Zhou, and Shuyang Lu. 2025a. [Jasper-token-compression-600m technical report](#). *Preprint*, arXiv:2511.14405.
- Sheng Zhang, Xin Zhang, Hui Wang, Lixiang Guo, and Shanshan Liu. 2018. [Multi-scale attentive interaction networks for chinese medical question answer selection](#). *IEEE Access*, 6:74061–74071. Dataset available at <https://github.com/zhangsheng93/cMedQA2>.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025b. [Qwen3 Embedding: Advancing text embedding and reranking through foundation models](#). *Preprint*, arXiv:2506.05176.
- Xinping Zhao, Xinshuo Hu, Zifei Shan, Shouzheng Huang, Yao Zhou, Xin Zhang, Zetian Sun, Zhenyu Liu, Dongfang Li, Xinyuan Wei, Youcheng Pan, Yang Xiang, Meishan Zhang, Haofen Wang, Jun Yu, Baotian Hu, and Min Zhang. 2026. [KaLM-Embedding-V2: Superior training techniques and data inspire a versatile embedding model](#). In *The Fourteenth International Conference on Learning Representations*.
- Tianyu Zheng, Ge Zhang, Tianhao Shen, Xueling Liu, Bill Yuchen Lin, Jie Fu, Wenhui Chen, and Xiang Yue. 2024. [OpenCodeInterpreter: Integrating code generation with execution and refinement](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12834–12859.

A Equal-Weight Aggregation of Teacher Similarities

Lemma 1 (Equal-Weight Cosine Similarity Aggregation). *Let u_i and v_i be the independently ℓ_2 -normalized representations produced by the i -th teacher for two inputs, where $i \in \{1, \dots, n\}$. Thus,*

$$\|u_i\|_2 = \|v_i\|_2 = 1.$$

Define the globally normalized concatenated representations as

$$\begin{aligned} U &= \text{Norm}(u_1 \parallel \dots \parallel u_n), \\ V &= \text{Norm}(v_1 \parallel \dots \parallel v_n). \end{aligned}$$

Then their cosine similarity satisfies

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

Proof. Let $\tilde{U} = u_1 \parallel \dots \parallel u_n$ and $\tilde{V} = v_1 \parallel \dots \parallel v_n$ denote the unnormalized concatenated representations. Because every component is independently normalized, their squared norms are

$$\begin{aligned} \|\tilde{U}\|_2^2 &= \sum_{i=1}^n \|u_i\|_2^2 = n, \\ \|\tilde{V}\|_2^2 &= \sum_{i=1}^n \|v_i\|_2^2 = n. \end{aligned}$$

Consequently,

$$U = \frac{\tilde{U}}{\sqrt{n}}, \quad V = \frac{\tilde{V}}{\sqrt{n}}.$$

The dot product of the normalized concatenated representations is therefore

$$\begin{aligned} U \cdot V &= \frac{1}{n} (\tilde{U} \cdot \tilde{V}) \\ &= \frac{1}{n} \sum_{i=1}^n (u_i \cdot v_i). \end{aligned}$$

Since U , V , u_i , and v_i are all ℓ_2 -normalized, the corresponding dot products are cosine similarities. Hence, the cosine similarity in the fused representation space is the arithmetic mean of the cosine similarities produced by the individual teachers. $\square$

B 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, 2024). 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.

C 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.

C.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.

C.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., 2026). 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 neg-

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.

atives. 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.

D 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. Tables 8 and 9 extend this comparison to a broader range of baseline models, while Tables 10 and 11 provide dataset-level performance breakdowns across the constituent task categories.

E 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 E), 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.

F Ablation Details and Additional Results

Compression-operator implementations. All compression-operator ablations start from the same Stage 1 checkpoint and use the same one-epoch Stage 2 adaptation budget. All variants retain the same preceding MLP and replace only the sequence-reduction operator. Hard Truncation keeps the prefix up to the target length. Token Pruning ranks token representations by their ℓ_2 norms, retains the highest-scoring tokens up to the target length, and restores their original order. Token Merging (ToMe) computes similarity scores between adjacent token representations, selects pairs according to the reduction budget, and averages the selected pairs. These pruning and merging variants are controlled operator-level implementations rather than reproductions of a specific published system. Learnable Conv1D replaces the parameter-free reduction operator with a randomly initialized trainable one-dimensional convolution. All trainable components follow the same optimization setting across variants.

Operator-level latency. Table 13 reports a separately conducted operator-level latency benchmark on an RTX 4090 with a batch size of 32. Because this benchmark uses a controlled operator-ablation setup, its absolute latency values should not be directly compared with those reported in the main end-to-end experiments.

Additional progressive-training results. Table 14 reports full-MTEB results from a controlled run using the same training configuration.

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

Source Dataset	Format	Language
The Pile (Gao et al., 2020)	s, p	English
FineWeb / FineWeb-Edu (Penedo et al., 2024)	s, p	English
IndustryCorpus2 (Shi et al., 2024a)	s, p	Multilingual
FineWeb-Edu-Chinese (Yu et al., 2025b)	s, p	Chinese
Sentence-transformers Training Data (Sentence Transformers, 2021)	q	English
BGE-M3-Data (Chen et al., 2024)	q	Multilingual
Industry Instruction (Shi et al., 2024b)	q	English
Infinity-Instruct (Li et al., 2025)	q	English
Smoltalk (Ben Allal et al., 2025)	q	English
Smoltalk-Chinese (Yu et al., 2025b)	q	Chinese
MNBVC (MOP-LIWU Community and MNBVC Team, 2023)	q	Chinese
WebQA (Li et al., 2016)	q	Chinese
WuDao Corpora (Queries) (Yuan et al., 2021)	q	Chinese

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

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

Source Dataset	Format	Language
MS MARCO (Bajaj 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; Baidu DuReader Team, 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 (Gupta et al., 2019; Hou et al., 2026)	s2p	English
CodeFeedback (Zheng et al., 2024)	s2p	English
StackExchange DupQuestions (Hoogeveen et al., 2015)	s2s	English
MetaMathQA (Yu et al., 2024)	s2p	English

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

Stage 1 reaches a Mean(Task) score of 74.73 without compression, while Stage 3 retains 74.67 at $\rho = 0.33$. Directly combining teacher alignment and sequence compression without Stage 1 reduces Mean(Task) to 56.36 and Retrieval to 28.15, compared with 70.47 and 61.83 for the initialization baseline. These results support the optimization benefit of the progressive curriculum under our experimental setting.

G 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 15 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 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 15, 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-

Model	Params	Dim.	M.Task	M.Type	Clas.	Clus.	Pair.	Rank.	Retr.	STS	Sum.
<i>MTEB (English)</i>											
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
<i>C-MTEB (Chinese)</i>											
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 ‘–’.

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.

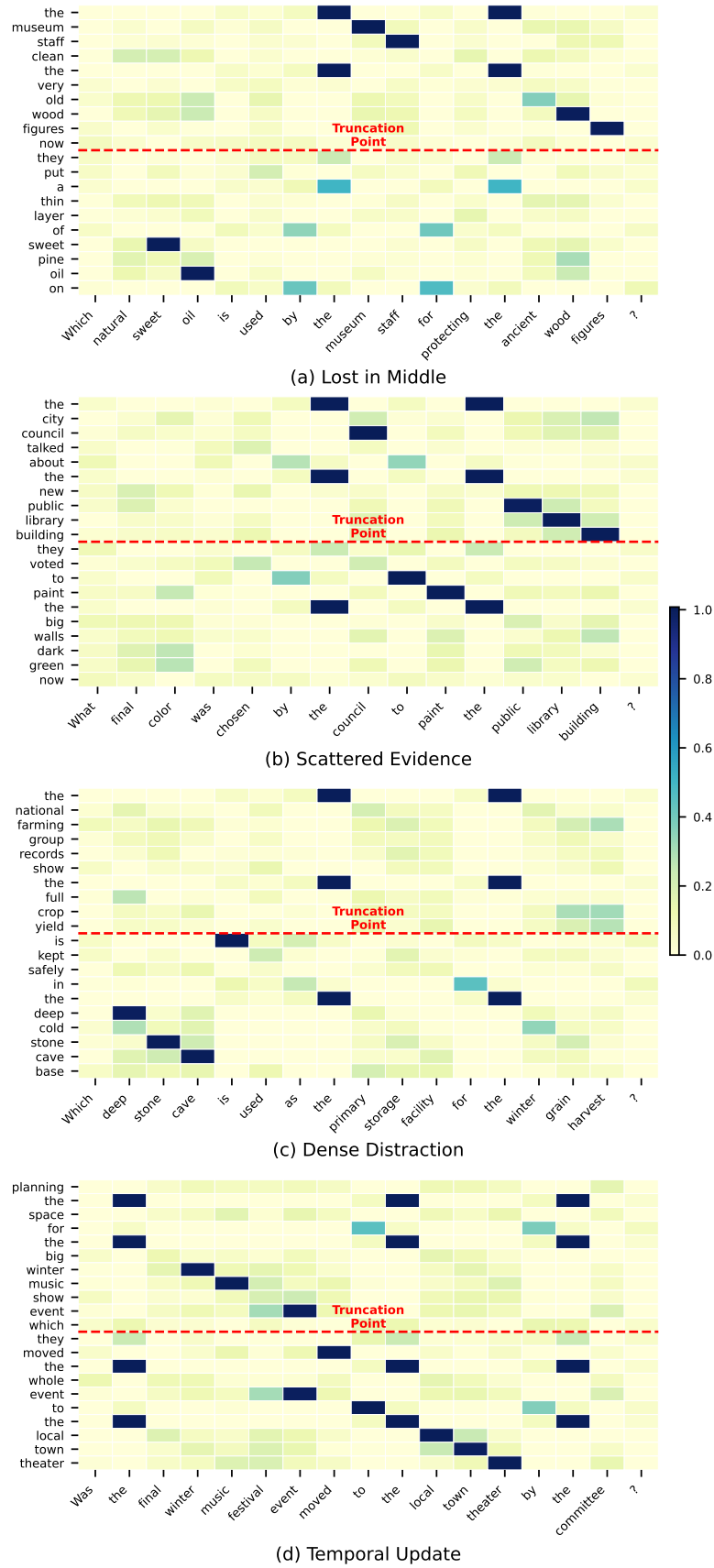


Figure 7: Token-level similarity heatmaps for four extended retrieval scenarios: (a) Lost in Middle, (b) Scattered Evidence, (c) Dense Distraction, and (d) Temporal Update. The red dashed line denotes the hard truncation point. Dark blue clusters indicate the highly activated core evidence successfully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

Model	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.

Model / Operator	Encoding Latency (ms) vs. Input Length					
	128	256	512	1024	2048	4096
Qwen3-Embedding-0.6B	1.36	3.05	7.35	15.71	32.80	71.98
Jasper-TC-600M ($\rho = 0.33$)	1.05	1.58	2.74	5.49	11.29	22.57
Hard Truncation ($\rho = 0.33$)	1.07	1.60	2.76	5.56	11.45	22.83
Token Pruning ($\rho = 0.33$)	1.20	1.73	2.90	5.73	11.60	23.09
Token Merging ($\rho = 0.33$)	1.93	3.58	6.69	13.49	27.40	55.25
Learnable Conv1D ($\rho = 0.33$)	1.34	1.88	3.02	5.83	11.75	23.22

Table 13: Encoding latency of different compression operators across input lengths. This operator-level benchmark was conducted separately from the main latency evaluation on an RTX 4090 with a batch size of 32. All compression variants use $\rho = 0.33$. The highlighted row denotes our ETC configuration, and the best latency at each input length is shown in **bold**.

Model / Training Phase	Params	Dim.	M.Task	M.Type	Clas.	Clus.	Pair.	Rank.	Retr.	STS	Sum.
<i>Initialization Baseline</i>											
Qwen3-Embedding-0.6B	595M	1024	70.47	64.72	84.58	54.05	84.37	48.18	61.83	86.57	33.43
<i>Standard Progressive Training</i>											
Jasper-TC-600M (Stage 1, no compression)	595M	2048	74.73	68.51	90.43	60.01	89.92	50.72	65.73	88.63	34.14
Jasper-TC-600M (Stage 2, $\rho = 0.50$)	595M	2048	74.05	67.80	90.34	58.18	89.82	50.47	64.71	88.66	32.45
Jasper-TC-600M (Stage 2, $\rho = 0.33$)	595M	2048	74.34	68.05	90.37	59.64	89.82	50.47	64.70	88.64	32.73
Jasper-TC-600M (Stage 2, $\rho = 0.20$)	595M	2048	73.69	67.79	90.33	56.86	89.82	50.47	64.08	88.69	34.25
Jasper-TC-600M (Stage 2, $\rho = 0.10$)	595M	2048	73.56	67.54	90.30	58.48	89.82	50.47	62.82	88.30	32.58
Jasper-TC-600M (Stage 3, $\rho = 0.50$)	595M	2048	74.46	68.18	90.55	58.34	90.15	50.71	65.75	88.82	32.92
Jasper-TC-600M (Stage 3, $\rho = 0.33$)	595M	2048	74.67	68.33	90.54	60.11	90.15	50.71	65.24	88.78	32.82
Jasper-TC-600M (Stage 3, $\rho = 0.20$)	595M	2048	74.48	68.44	90.50	59.36	90.15	50.71	64.92	88.78	34.63
Jasper-TC-600M (Stage 3, $\rho = 0.10$)	595M	2048	74.08	67.95	90.49	58.60	90.15	50.71	64.11	88.76	32.83
<i>Progressive-Pipeline Ablation</i>											
Jasper-TC-600M (Skip Stage 1, $\rho = 0.33$)	595M	2048	56.36	53.92	78.27	46.60	76.49	43.83	28.15	76.17	27.90

Table 14: Full-MTEB results for the standard progressive training pipeline and the Skip Stage 1 ablation. Stage 1 is evaluated without compression, Stage 2 uses fixed compression ratios, and Stage 3 is trained with dynamic compression-ratio sampling and evaluated at the indicated ratio. The gray row denotes the representative compressed configuration ($\rho = 0.33$). Best scores across the reported standard training stages are shown in **bold**. M.Task and M.Type denote Mean(Task) and Mean(TaskType), respectively. These results come from a controlled supplementary run using the same training configuration and are reported as a pipeline analysis rather than as intermediate checkpoints of the main experimental run.

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... [<i>... 500 words of meteorological and traffic logs omitted ...</i>] 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... [<i>... 300 words of logistic preparations omitted ...</i>] 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. [<i>... 300 words of security measures omitted ...</i>]	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... [<i>... 500 words detailing architectural disputes and budgets omitted ...</i>] 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... [<i>... 400 words of transportation logs omitted ...</i>] 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. [<i>... 400 words detailing stadium stage preparations omitted ...</i>] 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 15: 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.