DOC2FRC: Length-Consistent Document-Level Machine Translation via Fixed-Range Chunking

Xiaotian Wang[1,3] **Youyuan Lin**[2] **Zhan Shen**[1] **Hitomi Yanaka**[1,3,4]
[1]The University of Tokyo [2]Kyoto University [3]Riken [4] Tohoku University
[1]{eternaleden0321, shenzhan, hyanaka}@g.ecc.u-tokyo.ac.jp
[2]lin.youyuan.73v@st.kyoto-u.ac.jp

Abstract

Advanced large language models (LLMs) with long context windows can substantially reduce input truncation in document-level machine translation (DocMT). However, direct *Doc2Doc* translation remains prone to n-gram repetition and progressive quality degradation. A common remedy is to segment the document into finer-grained chunks. Nonetheless, conventional rule-based chunking approaches fail to handle the length distribution mismatch between training and inference. To address this, we introduce **F**ixed-**R**ange **C**hunking (**FRC**), utilizing dynamic programming to partition documents into chunks within a predefined length interval. By consistently applying FRC during training and inference, the input documents of any length are mapped to the same length distribution, substantially reducing train-test length mismatch. Centered on FRC, we propose a lightweight dual-boundary matching algorithm for chunk alignment, alongside four distinct training strategies. Experimental results show that FRC-based fine-tuning substantially improves 7B LLMs over direct *Doc2Doc* fine-tuning and outperforms existing DocMT methods on IWSLT2017. We further construct GLOBVDOC, a 10-language test set independent of mainstream DocMT training sources, and show that FRC improves out-of-distribution document translation.[1]

1 Introduction

Document-level machine translation (DocMT) has been continuously investigated for a long time (Miculicich et al., 2018; Voita et al., 2018; Maruf and Haffari, 2018; Lyu et al., 2021). Current large language models (LLMs) with long context windows have greatly alleviated the challenge of input length constraints for DocMT. However, direct *Doc2Doc* translation is still plagued by various issues, such as susceptibility to n-gram repetition during decoding (Jin et al., 2024; Peng et al., 2025a), and progressive quality degradation as sequence lengths increase (Yang et al., 2025). Accordingly, dominant strategies such as *Doc2Sent* and *Doc2Chunk* emerged, which translate documents sequentially at the level of sentences or chunks. While these methods can also mitigate the challenges associated with variable document input lengths, the design of optimal training paradigms based on them is still under active exploration.

Regarding the *Doc2Sent* strategy, early studies (Wu et al., 2024a; Li et al., 2024) utilized a fixed number of sentence pairs as contexts to facilitate training. While this approach can ensure input length alignment between training and inference, it remains rooted in sentence-level processing and thus lacks sufficient in-context information. More recently, Ramos et al. (2025) introduced a chunk-level training paradigm that partitions documents into a fixed number of sub-documents, and employs a hybrid training strategy incorporating both context-free and context-aware settings, emphasizing length-agnostic generalization. However, it cannot guarantee length distributional alignment between training and inference, which has been empirically observed to potentially lead to performance degradation (Varis and Bojar, 2021; Wan et al., 2022; Pitorro et al., 2024). While Peng et al. (2025a) proposed a solution using varying position embeddings to improve length extrapolation during training, they also prioritized enhancing translation capabilities across diverse lengths.

Driven by similar concerns regarding length distribution matching, instead of improving generalization across diverse lengths, we focus on achieving chunk-level alignment by mapping both training and inference documents into a unified length interval, thereby training exclusively on these consistent distributions. Nonetheless, conventional strategies, such as fixed-number and fixed-length

[1]The code, dataset, and fine-tuned models are publicly available at: https://github.com/ynklab/Doc2FRC.

Proceedings of the Eleventh Conference on Machine Translation, pages 782–817
October 28-29, 2026 ©2026 Association for Computational Linguistics

chunking, cannot fully reconcile length discrepancies, frequently leaving short residual fragments at the end of a document. To address this limitation, we propose **F**ixed-**R**ange **C**hunking (**FRC**), which employs a dynamic programming algorithm for document chunking, ensuring all document segments fall within a target length interval. Consequently, the length distribution of the training and inference data maintains consistently well-aligned.

In addition to introducing the method of FRC for DocMT, during the training data curation stage, we propose a lightweight chunk alignment algorithm based on dual-boundary matching, which can be applied to all document-level parallel corpora that are not fully sentence-aligned. Then, we use FRC pairs to construct five different data formats to train four types of models (as shown in Figure 1), enabling a systematic model-wise investigation.

Furthermore, aiming to probe out-of-distribution generalization and mitigate the scarcity of long-form evaluation data in the news and social domains, we manually curate a high-quality, 10-language test set named GLOBVDOC, based on the Global Voices[2]. This dataset is guaranteed to be strictly sentence-aligned and is designed to remain independent of the data distributions used in current mainstream DocMT training sets.

Our main contributions are as follows:

- We propose Fixed-Range Chunking and Dual-Boundary Matching-based Chunk Alignment algorithms to ensure length distribution consistency across training and inference with efficient chunk synchronization.
- We train four types of models across five training data formats. Our results show that FRC-based training strategies outperform the *Doc2Doc* baseline. On IWSLT2017 test data, our method achieves a d-BLEU improvement that exceeds the existing DocMT methods.
- We present GlobVDoc, a high-quality multilingual test dataset specifically designed for document translation, and establish a multilingual long-form DocMT benchmark.
- We introduce a variant of the Lexical Translation Consistency Ratio (LTCR) that estimates translation consistency based on the word alignments between the reference and the translation hypothesis.

[2] https://globalvoices.org

2 Related Work

Training-based DocMT Following the emergence of various LLM-based sentence-level translation paradigms (Xu et al., 2024; Alves et al., 2024; Guo et al., 2024), research focus has increasingly pivoted toward LLM-based DocMT. Early investigations primarily focused on *Doc2Sent* methodologies. Wu et al. (2024a) leveraged historical context from previous translations, whereas Li et al. (2024) introduced retrieval-based sentence pairs to provide auxiliary contextual information. Subsequent research focused on dataset development, utilizing training to establish performance benchmarks (Wicks et al., 2024; Alabi et al., 2025). Jin et al. (2024) extended the translation object to the chapter level for literary translation, employing direct chapter-to-chapter fine-tuning using only inter-chapter context. Compared to training exclusively on long document pairs, Ramos et al. (2025) introduced a training paradigm that segments documents into {1,2,4} parts and fine-tuned the model with MRD2D (Sun et al., 2022) and CAPT (Wang et al., 2023a) techniques, achieving performance comparable to large-scale LLMs while ensuring generalized translation ability across diverse input lengths. In contrast, our FRC approach focuses on specializing within a targeted length range.

Training-free DocMT Early DocMT works explored prompting formats (Hendy et al., 2023; Wang et al., 2023a; Karpinska and Iyyer, 2023; Wu et al., 2024a). Subsequent studies enriched prompts with document-derived signals: Liu et al. (2025) extracted summarization and entity translation knowledge from documents to construct diverse prompts and reranked candidate translations. Hu et al. (2025) investigated DocMT within a multi-turn conversation framework and showed that providing the entire source document as prior context can improve iterative refinement.

Beyond prompting, agentic systems explicitly separate roles to support DocMT: TransAgent (Wu et al., 2024b) assigns specialized agents (e.g., translator and reviewer) to achieve iterative refinement. Sent2Sent++ (Guo et al., 2025), GRAFT (Dutta et al., 2025), and DELTA (Wang et al., 2025b) utilize translation memory to retrieve contextual cues and perform step-by-step translation without debate. Finally, TransGraph (Pham et al., 2025) constructs a discourse-based knowledge graph from document chunks to guide the translation.

Since this work focuses on FRC as a solution for

ensuring distributional consistency between training and inference, we do not discuss its application to training-free DocMT in detail.

Analysis Peng et al. (2025a) showed that translation quality degrades when inference-time inputs follow length distributions unseen during training, and mitigated this shift by modifying positional embeddings. Our proposed FRC approach addresses the same length-mismatch issue by directly synchronizing the length distributions between the training and inference phases. Moreover, while many studies incorporate partial translations as context (Wang et al., 2023a; Wu et al., 2024a; Pham et al., 2025; Ramos et al., 2025), Choudhary et al. (2025) demonstrate that source-only context can also effectively capture discourse phenomena. Therefore, to ensure length regulation in FRC and efficient decoding, we use source-only context by default (see Appendix G for further discussion).

3 Training Data Curation

3.1 Fixed-Range Chunking

We propose a Two-Stage Dynamic Programming Algorithm (see Algorithm 2 in the Appendix for details) designed to optimize the partitioning of a set of pre-segmented basic units $U = \{u_i\}_{i=1}^{n}$ into chunks that strictly fall within a fixed range $[m, M]$, where m and M denote the minimum and maximum token limits, respectively.

We also design a piecewise cost function $C(l)$. It encourages lengths to cluster within $[m, M]$ when valid, and minimizes the magnitude of deviation when boundary violations are inevitable.

Pre-processing Any single unit u_i exceeding M tokens is isolated prior to the DP process.

Stage 1: Constraint Optimization We search for a segmentation in which every chunk's length l strictly satisfies $m \leq l \leq M$. Let $f[i]$ denote the minimum cost to pack the first i units. If $f[n] < \infty$, the result is returned.

Stage 2: Relaxed Optimization If strict segmentation is infeasible, we relax the constraints. We allow at most K chunks to violate the boundary $[m, M]$ and extend the DP state to $f[v][i]$, representing the minimum cost for the first i units using exactly v violations. The final solution is chosen by minimizing the cost over all $v \in [1, K]$.

3.2 Chunk Alignment

While existing sentence alignment methods (Sennrich and Volk, 2011; Thompson and Koehn, 2019) can provide precise correspondences and achieve global optimality for chunk alignment, the intensive preprocessing required for large-scale corpora is often computationally prohibitive. Additionally, perfect sentence alignment is often unattainable in real-world scenarios (O'Brien et al., 2025).

Therefore, we obviate strict sentence alignment by proposing a lightweight Dual-Boundary Matching based Chunk Alignment method that leverages the two boundary units of a source chunk to match high-scoring target units located in proximal relative positions within the target document; the span delimited by these two target units is designated as the corresponding target chunk. Although this heuristic may lead to marginal data loss due to algorithmic constraints (see Table 9 in the Appendix for details), it offers a highly scalable and efficient alternative for processing massive datasets.

Algorithm 1 details the proposed chunk alignment strategy. We first define SIMWRP, a hybrid scoring function that aggregates similarity and relative position information with hyperparameters λ and σ. For each source chunk c_i, we calculate an initial match score against all target units $t_j \in \mathcal{T}$ based on the chunk's end boundary c_i^{end} and identify the top-k candidates. To reduce matching ambiguity, we perform dual verification, which refines their scores by incorporating the corresponding match of the subsequent chunk's start boundary unit c_{i+1}^{first}. Finally, a DP approach searches for the optimal path that maximizes the total score. If the boundary constraints remain unsatisfied, the candidate window size k is incrementally expanded until a feasible alignment is identified or the search space is exhausted.

3.3 FRC-based Training Formats

Source documents are partitioned into fixed-range chunks using the method outlined in Section 3.1, and aligned via the algorithm described in Section 3.2 to generate FRC pairs. Utilizing these pairs, we develop various training strategies. As shown in Figure 1, we define five distinct data formats: (1) **Doc**: The original, full-length document; (2) **0c1t**: A standalone chunk pair without additional context; (3) **1c1t**: Uses the preceding chunk as context (the 1st chunk serves exclusively as the contextual input); (4) **2c1t**: Uses the two preceding chunks as context (the 1st and 2nd chunks serve exclusively as contextual inputs); (5) **FS4** (Fixed-Size 4): Chunks are merged into four segments, utilizing all previous chunks as context in an incremental

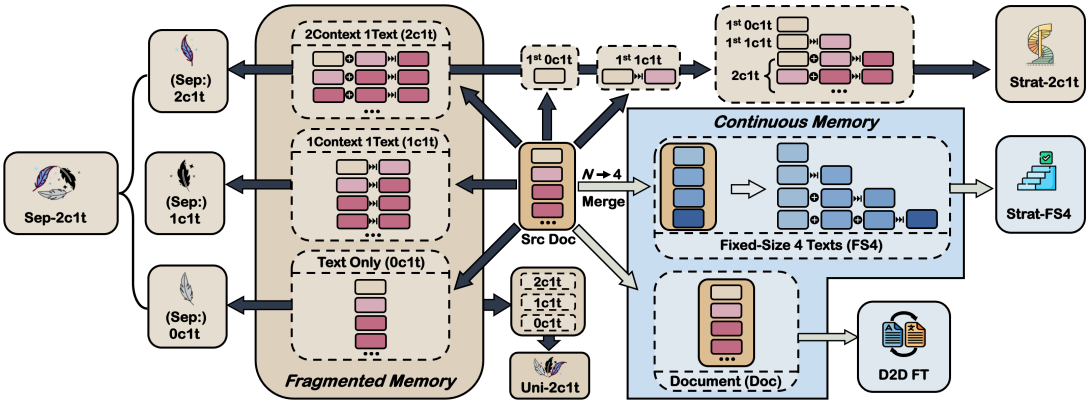


Figure 1: Construction scheme of training data for four model configurations derived from a single document.

manner. Leveraging these training data formats, we train four different types of models:

D2DFT A baseline model fine-tuned directly on document pairs.

SEP Independent models trained separately using 0c1t, 1c1t, and 2c1t formats. Their combined implementation is referred to as the SEP-2c1t.

STAIR This category includes STAIR-FS4, which utilizes the incrementally stratified data, and STAIR-2c1t, which employs a "spiral" staircase approach that begins with 0c1t and 1c1t samples before maintaining the 2c1t training format.

UNI A single model trained on a comprehensive mixture of 0c1t, 1c1t, and 2c1t data, where the three formats exhibit a clustered distribution, collectively forming three phases during training.

Regarding data formats, we categorize the training data into *Fragmented Memory* and *Continuous Memory* types: While ∗c1t formats utilize only partial contextual information, both Doc and FS4 provide access to the full document contexts. Regarding training methods, while chunks typically serve as a translation target once, the UNI model may target a single chunk up to three times. Moreover, while STAIR and UNI models undergo joint training on mixed formats, SEP models focus on "expert" training specialized for individual formats.

4 GlobVDoc Dataset

Existing test sets for DocMT in the news and social domains typically consist of relatively short documents (e.g., WMT2022 (Kocmi et al., 2022) and News Commentary). To address the scarcity of long-form document pairs within this domain, we introduce GLOBVDOC, a 10-language document-level test set sourced from Global Voices, a multilingual news site previously included in OPUS (Tiedemann, 2012). This dataset is designed to be independent of mainstream DocMT training sources, including ***IWSLT*** (Cettolo et al., 2012, 2017), ***BWB*** (Jiang et al., 2022), ***GuoFeng*** (Xu et al., 2022), ***News Commentary***, ***OpenSubtitles*** (Lison and Tiedemann, 2016; Lison et al., 2018), and ***Europarl*** (Koehn, 2005), thereby enabling the evaluation of out-of-distribution translation performance.

GLOBVDOC was constructed by a team of eight individuals, comprising several authors and additional volunteers. We manually selected 10 document pairs for each en → {de, es, fr, it, ko, nl, pt, ru, zh}. Considering document translation methods derived from a sentence-level translation system, the test data was required to be strictly sentence-aligned. To ensure consistency and quality, the construction process adhered to standardized guidelines, including Source Selection and Diversity, Temporal Relevance, Alignment Quality, Structural Consistency, Manual Segmentation, Minimal Adjustments, and Mapping Priority (see Appendix C for details). The documents comprise various formats, including interviews, biographies, real-time news reports, and podcasts. Appendix C provides detailed statistics for all language pairs in the GlobVDoc dataset (Table 7), together with a comparison against other commonly used datasets in Table 8.

In addition, we introduce a quality estimation function in Appendix D to evaluate the autocorrelation within each test dataset.

5 Experiment

5.1 Datasets

We utilize the ***DocBlocks*** dataset[3] (Ramos et al., 2025) as our training data, which is composed

[3] https://huggingface.co/datasets/sardinelab/DocBlocks

of five distinct corpora: IWSLT, BWB, GuoFeng, News Commentary, and Europarl.

For evaluation[4], we use IWSLT2017 (Cettolo et al., 2017) and BWB test data, as well as our self-established GlobVDoc dataset. Initially, we utilize the IWSLT2017 test data (TED talk documents), consistent with Ramos et al. (2025), to conduct experiments on the six language pairs en ↔ {de, fr, it, ko, nl, zh}. The BWB test data was derived from continuous chapters extracted from six web novels. Due to the relatively short length of individual chapters, we concatenate two adjacent chapters from the same book for zh → en evaluation.

5.1.1 Experimental Setups

We employ QWEN2.5-7B-INSTRUCT (Qwen Team, 2024) and TOWERINSTRUCT-MISTRAL-7B (Alves et al., 2024) (hereinafter referred to as TOWER-7B-INSTRUCT) as our backbone models to facilitate a comparison of fine-tuning performance on the shared DocBlocks dataset with Ramos et al. (2025). Both models support 32k-token contexts, with TOWER-7B-INSTRUCT model having been post-trained for sentence-level MT. Due to the significant discrepancy between their tokenizers, we construct two distinct sets of fixed-range $[256, 512]$ chunk pairs by utilizing the respective tokenizer of each model for length measurement, while the Stanza tokenizer (Qi et al., 2020) is jointly used to delineate the basic units, and cosine similarity based on LaBSE (Feng et al., 2022) embeddings is used for chunk alignment.

For reference, we also include results of some widely used commercial LLMs, GPT-4.1 (OpenAI Team, 2024), DeepSeek-v3.2 (DeepSeek-AI et al., 2025), and Gemini-2.5-Pro (Gemini Team, 2025).

We provide a comparison of the train-test length distributions across all the training formats in Appendix E.2. The training data curation procedure, sample statistics, and setups for both training and decoding are detailed in Appendix E.

5.1.2 Evaluation Methods

d-BLEU, ds-BLEU Classical document-level BLEU (d-BLEU; Papineni et al., 2002; Liu et al., 2020) and ds-BLEU defined by Peng et al. (2025a), calculating the equivalent of sentence-level BLEU scores (Lin and Och, 2004).

[4]In the Limitations Section, we present the reasons for not incorporating the GuoFeng test data into our study.

d-COMET We compute d-COMET[5] at the chunk-level using SLIDE (Raunak et al., 2024) in the same manner as Ramos et al. (2025). However, we introduce variations to the chunking procedure (see Appendix A for details).

LTCR The Lexical Translation Consistency Ratio (LTCR; Lyu et al., 2021) measures terminology consistency in DocMT. The conventional implementation relies on cross-lingual word alignment between the source and hypothesis, making the metric sensitive to alignment errors. We propose a LTCR variant that performs monolingual alignment between the hypothesis and reference using the Needleman–Wunsch algorithm (Needleman and Wunsch, 1970).

Specifically, for each interest word w in the reference (see Appendix B), we collect its aligned translations in the hypothesis across all sentence pairs. The word-level consistency is:

$$C(w) = \frac{\max_v \text{count}_w(v)}{K_w}, \quad K_w \geq 2 \quad (1)$$

where K_w is the number of aligned translations and $\text{count}_w(v)$ denotes the frequency of the translation form v. The document-level score is the occurrence-weighted average:

$$\text{LTCR} = \frac{\sum_{w \in W'} C(w) \cdot K_w}{\sum_{w \in W'} K_w} \quad (2)$$

where $W' = \{w \mid K_w \geq 2\}$.

GEMBA-DA While GEMBA-MQM (Kocmi and Federmann, 2023a) has demonstrated efficacy in detecting error spans at the sentence or segment level, its applicability to document-level evaluation remains under-explored. Therefore, following Mrozinski et al. (2025) and Wu et al. (2024b), we utilize GEMBA-DA[6] for our evaluation.

5.2 Results on IWSLT2017 and BWB

Overview The performance of models trained with two different backbone architectures is reported in Table 1. First, chunking generally outperforms the d2d strategy for both original models, a benefit especially pronounced for TOWER due to its prior sentence-level fine-tuning. While all fine-tuned models naturally exceed the baselines on in-domain data, TOWER-based models show more substantial gains than QWEN2.5.

[5]https://huggingface.co/Unbabel/wmt22-comet-da

[6]https://github.com/MicrosoftTranslator/GEMBA

Models	Decoding Format	IWSLT2017 en-xx				IWSLT2017 xx-en				BWB zh-en			
		d-BLEU	ds-BLEU	d-COM.	GEM.-DA	d-BLEU	ds-BLEU	d-COM.	GEM.-DA	d-BLEU	ds-BLEU	d-COM.	GEM.-DA
Large-Scale LLMs													
GPT-4.1	d2d	36.53	34.91	86.23	94.72	39.19	38.71	85.79	91.93	22.85	23.09	81.06	94.26
Deepseek-v3.2	d2d	32.44	31.60	84.86	92.16	45.73	44.15	85.60	92.64	21.73	21.71	81.25	93.92
Gemini-2.5-Pro	d2d	40.05	38.00	86.38	94.79	49.67	48.38	86.22	94.38	21.53	21.23	80.72	94.74
Qwen2.5-7B-Instruct													
Orig Model	d2d	25.88	27.52	80.00	61.16	35.54	35.07	84.54	84.46	18.57	18.47	80.20	81.21
	0c1t	30.00	29.23	81.83	63.73	35.42	34.79	85.03	84.34	19.38	19.18	80.72	79.00
	1c1t	29.80	28.79	82.05	63.72	35.29	34.69	84.94	84.64	19.07	18.85	80.47	79.00
	2c1t	30.07	29.00	82.42	64.33	35.37	34.72	84.88	84.60	18.75	18.52	80.59	79.03
	FS4	28.29	28.42	81.06	63.14	35.56	34.79	83.33	84.57	18.81	18.66	80.27	79.56
w/ D2DFT	**d2d**	29.17	34.50	82.38	74.31	22.46	38.18	82.78	80.07	26.41	26.27	80.63	84.03
	0c1t	32.07	33.09	82.76	74.03	35.78	36.16	83.99	76.24	25.09	25.31	79.92	82.36
STAIR-FS4	d2d	31.42	35.77	83.52	78.57	39.20	41.22	85.36	84.56	26.07	25.83	80.80	85.46
	0c1t	30.43	29.52	78.51	66.29	42.00	40.81	85.06	82.96	26.20	26.05	80.71	83.95
	1c1t	32.70	32.07	80.82	71.79	43.58	42.26	85.65	84.65	26.40	26.24	80.71	83.46
	2c1t	33.05	32.46	81.16	73.02	43.75	42.43	**85.73**	84.82	26.42	26.27	80.70	84.13
	FS4	30.68	33.05	81.57	75.62	37.30	42.18	83.94	84.49	26.40	26.22	80.49	84.18
STAIR-1c1t	0c1t	32.44	32.14	80.61	71.67	44.40	42.95	85.25	81.98	**27.14**	27.06	80.94	85.23
	1c1t	36.72	35.59	83.25	79.03	46.87	45.15	85.62	84.06	**27.14**	**27.09**	**81.10**	**86.38**
STAIR-2c1t	0c1t	32.06	31.14	79.79	69.74	43.51	41.88	85.27	80.95	26.84	26.74	80.76	84.33
	1c1t	36.74	35.34	83.00	78.38	45.65	43.91	85.62	84.01	26.77	26.66	80.89	84.97
	2c1t	36.97	35.48	83.08	78.01	46.24	44.37	85.70	84.19	26.71	26.61	80.88	85.28
SEP-2c1t	**0c1t**	38.21	36.82	84.13	81.67	45.21	43.79	84.83	82.89	26.28	25.76	80.47	84.15
	1c1t	**38.51**	**37.00**	**84.70**	**82.42**	**47.07**	45.19	85.57	**84.90**	26.58	26.41	80.72	84.74
	2c1t	38.32	36.54	84.59	82.40	46.80	**45.22**	85.47	84.86	26.44	26.34	80.64	84.97
Tower-7B-Instruct-Mistral													
Orig Model	d2d	6.54	11.77	38.58	-	10.36	15.40	56.49	-	8.79	8.45	69.15	-
	0c1t	32.24	30.91	83.61	77.21	33.56	33.03	84.09	79.12	17.35	17.18	79.51	50.95
	1c1t	31.55	29.96	82.68	74.56	36.83	34.67	82.84	79.19	17.48	16.94	79.26	50.21
	2c1t	31.66	30.00	82.73	74.96	36.45	35.05	83.27	80.24	17.63	17.17	79.16	49.77
	FS4	20.16	24.34	77.83	65.20	19.80	25.75	75.53	62.37	15.56	15.73	77.21	48.92
w/ D2DFT	**d2d**	34.36	35.13	82.51	73.38	17.20	37.91	75.52	69.91	25.14	25.10	80.24	70.26
	0c1t	34.51	33.67	82.15	73.05	31.59	32.74	78.18	68.99	21.99	21.31	80.20	53.15
STAIR-FS4	d2d	21.54	32.12	80.49	73.76	31.88	44.81	80.96	71.28	26.52	26.64	80.90	80.13
	0c1t	31.35	30.04	78.48	66.12	52.34	51.21	85.33	79.57	27.43	27.38	80.96	82.77
	1c1t	34.55	33.37	81.56	75.09	56.35	55.40	86.01	83.06	27.50	27.36	81.02	83.44
	2c1t	34.96	33.90	81.84	76.10	56.79	56.15	85.87	83.32	27.47	27.38	81.03	83.38
	FS4	35.81	35.32	82.29	81.59	59.36	57.98	84.07	83.86	**27.61**	**27.62**	80.56	83.28
STAIR-1c1t	0c1t	34.34	32.98	80.58	71.97	61.81	59.87	85.46	81.53	27.04	26.97	81.03	81.92
	1c1t	38.91	37.10	84.09	83.02	66.14	64.61	86.29	85.07	27.53	27.46	**81.16**	83.33
STAIR-2c1t	0c1t	33.45	31.97	79.39	70.36	60.46	58.88	85.37	80.89	27.32	27.33	80.96	81.21
	1c1t	38.26	36.38	84.13	81.55	65.32	63.90	86.17	84.54	27.55	27.57	81.06	83.26
	2c1t	38.29	36.40	84.14	81.70	66.51	64.86	86.26	84.93	27.53	27.50	81.10	83.72
UNI-2c1t	**0c1t**	36.32	34.51	84.00	76.36	63.71	61.21	85.85	77.86	25.43	25.26	80.79	78.15
	1c1t	36.65	34.75	84.40	77.15	65.82	63.32	86.15	79.44	25.68	25.51	80.97	78.85
	2c1t	36.53	34.72	84.33	77.32	66.06	63.55	86.15	80.68	25.62	25.38	80.98	79.41
SEP-2c1t	**0c1t**	**39.61**	**37.87**	84.69	86.29	63.05	61.75	86.09	83.56	27.38	27.42	81.11	**83.95**
	1c1t	39.51	37.86	**84.83**	**86.34**	66.84	**65.29**	86.49	85.44	27.16	27.14	81.03	83.31
	2c1t	38.78	37.54	84.80	85.80	**67.20**	65.12	**86.53**	**85.73**	26.82	26.77	81.03	83.77

Table 1: Translation results of different models under various decoding formats. Decoding formats that are consistent with the model training setup are highlighted in **bold**. For the four evaluation metrics, the best performance achieved under the same backbone model is also shown in **bold**. For each individual model, comparisons across different decoding formats are indicated by underlining the better-performing results.

Since our evaluated d-BLEU scores for the original models differ from those reported by Ramos et al. (2025), we compare the relative gains over the original model in the d2d mode. On IWSLT2017, their best results via chunking reported improvements of approximately 6/26 d-BLEU points on en-xx and 13.5/47 on xx-en for QWEN2.5 / TOWER, whereas our SEP-2c1t model achieves gains of 12.44 (+6.44), 32.24 (+6.24), 11.26 (-2.24), and 56.84 (+9.84), respectively. Consequently, our approach outperforms theirs in most instances, except for the QWEN2.5-based IWSLT xx–en case.

Model-Wise Analysis From here on, we will describe the model-wise analysis.

D2DFT: Although the d2d decoding strategy consistently yields higher ds-BLEU scores than 0c1t, its d-BLEU scores on IWSLT2017 are markedly lower. Given that d-BLEU assigns greater weight to longer documents than ds-BLEU, this indicates that while the overall translation quality of the D2DFT model under d2d decoding is higher than 0c1t, chunking remains more effective for translating longer documents.

STAIR-FS4: Instead of rigidly dividing into four equal-sized chunks, this approach applies FRC followed by merging, which results in a highly discretized chunk-length distribution, increasing the coverage of test-time sequence lengths during training. Across all evaluation metrics, STAIR-FS4 outperforms D2DFT on IWSLT2017 and achieves comparable results on BWB. However, a latent issue remains regarding the mismatch between the FS4 training data format and the optimal decoding format, particularly in the QWEN2.5 variant.

STAIR-1c1t / 2c1t: As 1c1t and 2c1t samples dominate training, using their corresponding formats during decoding naturally yields optimal performance. Moreover, in terms of d(s)-BLEU, both models consistently exceed STAIR-FS4 on IWSLT2017, despite marginal differences

Models	GlobVDoc en-xx (d-BLEU / d-COMET)									
	de	es	fr	it	ko	nl	pt	ru	zh	all
Large-Scale LLMs										
GPT-4.1	35.58 / 88.63	58.32 / 89.43	47.38 / 88.82	44.95 / 89.45	30.88 / 91.33	43.13 / 89.75	60.39 / 89.92	35.07 / 92.00	54.82 / 90.97	45.61 / 90.03
GPT-4.1-mini	34.80 / 88.53	57.29 / 89.17	46.63 / 88.98	43.65 / 89.02	30.93 / 90.88	41.76 / 89.63	60.19 / 89.91	34.57 / 91.71	52.12 / 90.34	44.65 / 89.80
Deepseek-v3.2	33.32 / 88.19	53.48 / 88.86	49.38 / 88.52	44.84 / 89.20	32.14 / 91.34	41.13 / 89.35	57.08 / 89.81	31.89 / 91.66	45.95 / 89.60	43.24 / 89.61
Gemini-2.5-Pro	37.28 / 88.70	54.08 / 89.23	53.03 / 89.06	47.57 / 89.69	35.39 / 91.67	43.98 / 90.12	56.37 / 89.08	34.32 / 91.99	56.80 / 91.18	46.54 / 90.08
Agent-based Methods										
GRAFT-Q	29.20 / 87.54	49.83 / 88.21	43.93 / 87.96	39.11 / 88.72	28.68 / 90.22	32.59 / 87.66	53.53 / 89.48	26.55 / 90.46	51.74 / 90.29	39.46 / 88.96
GRAFT-G	34.65 / 88.36	54.31 / 88.73	47.81 / 88.57	44.81 / 88.71	33.73 / 91.42	40.06 / 89.44	59.79 / 89.73	31.98 / 91.21	55.60 / 90.92	44.75 / 89.68
DELTA-Q	29.37 / 87.24	50.69 / 88.17	44.87 / 87.57	40.24 / 88.33	29.18 / 90.29	32.94 / 87.60	53.54 / 89.42	25.78 / 89.90	54.27 / 90.62	40.10 / 88.79
DELTA-G	34.76 / 88.70	55.34 / 88.99	47.28 / 88.94	42.77 / 89.41	35.22 / 91.48	40.62 / 89.73	59.41 / 90.11	30.86 / 91.46	54.13 / 90.52	44.49 / 89.93
Qwen2.5-7B-Instruct										
Orig Model	23.50 / 84.92	47.35 / 88.02	40.50 / 85.80	35.01 / 86.70	15.67 / 81.03	27.28 / 84.44	48.76 / 88.50	21.92 / 84.76	52.85 / **90.65**	34.76 / 86.09
w/ D2DFT	31.56 / 87.97	**52.91** / 88.77	46.74 / 88.01	**44.64** / 88.35	18.48 / 82.86	35.90 / 89.06	**50.33** / 88.28	28.14 / 89.30	**54.84** / 90.17	40.39 / 88.08
STAIR-FS4	31.76 / 87.24	36.57 / 84.66	47.02 / 87.47	43.53 / 87.97	11.59 / 87.91	36.28 / 88.29	50.26 / 87.22	28.11 / 89.98	54.22 / 88.36	37.70 / 87.68
STAIR-2c1t	**31.84** / 87.99	51.95 / **88.79**	**47.91** / 88.29	43.24 / 88.41	**27.52** / **89.82**	36.66 / **89.22**	49.20 / **88.46**	27.53 / **90.05**	53.91 / 90.26	**41.08** / **89.05**
SEP-2c1t	31.39 / **88.09**	52.62 / 88.78	46.64 / **88.32**	43.20 / **88.46**	16.89 / 81.16	**36.71** / 89.09	48.54 / 88.26	**28.47** / 89.66	46.96 / 83.19	39.05 / 87.49
Tower-7B-Instruct-Mistral										
Orig Model	31.79 / 87.53	50.13 / 87.69	41.84 / 87.30	36.64 / 86.50	28.33 / 87.84	37.01 / 89.01	44.19 / 85.74	25.55 / 88.96	42.76 / 86.85	37.58 / 87.49
w/ D2DFT	30.76 / 84.60	50.20 / 87.81	46.31 / 87.35	43.72 / 86.16	25.10 / 89.62	15.05 / 79.57	48.06 / 86.96	26.17 / 82.98	53.95 / 90.11	37.70 / 85.79
STAIR-FS4	33.74 / 87.51	**51.75** / 87.49	47.75 / 87.66	43.93 / 87.87	**30.25** / 89.81	39.03 / 88.49	**49.83** / 87.22	28.40 / 90.33	52.93 / 88.40	41.96 / 88.31
STAIR-2c1t	**34.06** / **88.30**	51.31 / **88.76**	48.29 / **88.15**	**44.69** / **88.99**	27.94 / **90.45**	**39.59** / 89.56	49.36 / **88.54**	**29.28** / **91.18**	**54.01** / **90.23**	**42.06** / **89.35**
SEP-2c1t	33.08 / 88.16	51.64 / 88.75	**48.62** / 88.03	44.66 / 88.74	29.37 / 90.37	38.97 / **89.61**	49.29 / 88.53	29.20 / 91.09	52.97 / 90.04	41.98 / 89.26

Table 2: Translation Benchmark on GlobVDoc. For the two evaluation metrics, the best performance achieved under the same backbone model is shown in **bold**. Large-scale LLMs utilize d2d decoding, while original model follows a 0c1t format. Other fine-tuned models utilize decoding format corresponds to the training configuration. {GRAFT, DELTA}-{Q,G} denote using QWEN3-30B-A3B-INSTRUCT and GPT-4.1-mini as backbone models, respectively.

on BWB. A direct comparison between STAIR-1c1t and STAIR-2c1t reveals a narrow performance gap, with the former maintaining a slight advantage.

SEP / UNI-2c1t: Leveraging three independently trained sub-models, the SEP strategy achieves superior performance across all $*$c1t decoding modes and yields the best overall results on IWSLT2017, attaining d(s)-BLEU and d-COMET parity with large-scale LLMs. Conversely, joint training with mixed formats (UNI-2c1t) significantly degrades accuracy. This decline likely stems from training confusion, as the same target chunk is mapped to three source texts with varying contextual depth (see Appendix F for further discussion).

5.3 Benchmark on GlobVDoc

Based on the proposed GlobVDoc test data, we establish a translation benchmark in Table 2. In the benchmark, we include two agent-based methods, GRAFT (Dutta et al., 2025) and DELTA (Wang et al., 2025b), with QWEN3-30B-A3B-INSTRUCT and GPT-4.1-mini as the underlying models.

For the QWEN2.5-based models, performance differences are mainly seen in en $\rightarrow$ {es, ko, zh}. Notably, STAIR-FS4 underperforms on es and ko, while SEP-2c1t lags on ko and zh. Therefore, STAIR-2c1t achieves the highest overall accuracy. While most TOWER-based models demonstrate consistent performance across various language pairs, the D2DFT model exhibits an obvious degradation in quality for en $\rightarrow$ {de, nl, ru} translations.

Unlike results on IWSLT2017 and BWB, performance on the out-of-distribution GlobVDoc dataset exhibits a gap compared to large LLMs, despite achieving parity in certain language pairs. This deficiency is most pronounced for pt, nl, and ko, likely due to limited samples in the DocBlocks dataset. Furthermore, while most of our models outperform {GRAFT,DELTA}-Q, they still trail GPT-4.1-mini.

In addition, Figure 2 reports the LTCR results for different models. The D2DFT and STAIR-FS4 models, leveraged by *Continuous Memory* training (Section 3.3), yield superior LTCR results, with the latter matching or surpassing large-scale LLMs. In contrast, while other models may achieve high accuracy scores, their limited contextual scope results in diminished terminology consistency.

6 Analysis

Beyond the experiments discussed in this section, Appendix G compares Source-Only and Source + MT context strategies, and Appendix H reports results of length-centric analysis and FRC unit selection during test time.

6.1 Extrapolating Position Embedding

Prior work has extensively explored Position Embedding (PE) to mitigate the discrepancy between sequence length distributions during training and inference. In this paper, we evaluate two primary strategies for comparison: (1) Position Offset Augmentation, which enhances model generalization across diverse length distributions by introducing an offset to position indices. Specifically, we implement two variants: SHAPE (Kiyono et al., 2021; Ruoss et al., 2023), which samples offsets uni-

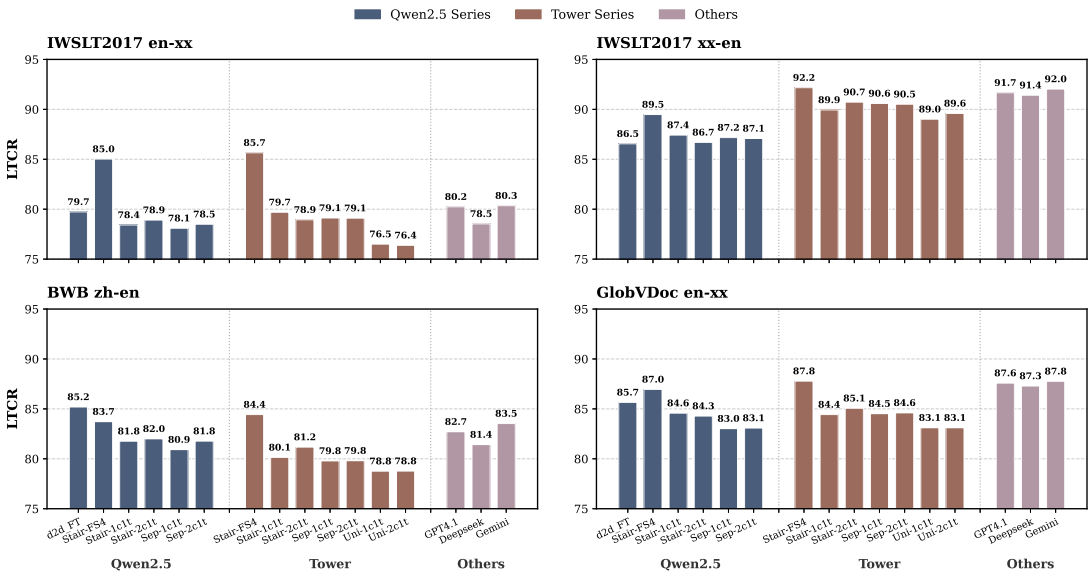


Figure 2: LTCR scores of various models on each dataset.

formly from the gap between the sample length and the model's maximum window size M; and Uniform SHAPE (UnifPE; Peng et al., 2025a), which makes each position within $[0, M-1]$ equally likely during training. (2) Position Interpolation (PI; Chen et al., 2023), which compresses or interpolates the position indices of long sequences into a pre-defined range. We implement Linear PI, which downscales position indices in both training and test data by a factor of 4 into $[0, 8192]$.

The results[7] are summarized in Table 3. All evaluated PE-based strategies successfully enhance the D2DFT baseline. This improvement is especially evident on the longer sequences of the IWSLT test set, where d-BLEU scores increased by 5.22, 9.25, and 10.58 across the backbone models. Because standard d2d decoding is prone to instability and n-gram repetition (Jin et al., 2024; Hiraoka and Inui, 2025), PE-based interventions effectively alleviate these hallucinations and stabilize text generation (Peng et al., 2025a). SEP-1c1t model achieves the best performance compared with these baselines, showing clear improvements in both accuracy and n-gram repetition rate. This advantage is particularly pronounced on the IWSLT dataset.

6.2 Chunking Strategy

In addition to our proposed FRC strategy and the FS4 strategy followed Ramos et al. (2025), we further compare our method with two conventional chunking strategies in this section:

[7]Since TOWER uses Sliding Window Attention, applying Linear PI would not remove the local-attention constraint. We therefore evaluate Linear PI on QWEN2.5 with full attention.

Models	PE	FT	d-BLEU↑ (NRR↓)		
			IWSLT	BWB	GlobVDoc
QWEN	R.	D2D	25.82 (5.75%)	26.27 (0.00%)	**40.39 (0.00%)**
	R.+L.	D2D	36.40 (1.34%)	26.12 (0.00%)	39.76 (1.11%)
	R.	SEP	**42.79 (1.07%)**	**26.58** (0.00%)	39.16 (2.22%)
TOWER	R.	D2D	25.78 (13.64%)	25.14 (0.00%)	37.70 (3.33%)
	R.+S.	D2D	31.00 (8.83%)	26.51 (0.00%)	41.11 (2.22%)
	R.+U.	D2D	35.03 (7.36%)	26.31 (0.00%)	41.52 (**0.00%**)
	R.	SEP	**53.18 (1.07%)**	**27.16** (0.00%)	**41.87 (0.00%)**

Table 3: SEP-1c1t v.s. PE-based methods. IWSLT values denote the average of en-xx and xx-en. Abbreviations: R. (RoPE), S. (SHAPE), U. (UnifPE) and L. (Linear PI). NRR denotes the n-gram repetition rate.

Fixed Window Chunking (FW). This method greedily traverses the document with a fixed-length sliding window, concatenating consecutive sentences into a chunk (aka., a *blob*; Finkelstein et al., 2024; Wang et al., 2025a) as long as their combined length does not exceed the window size. Since most chunks are close to the window size, FW is the baseline most similar to FRC. However, it suffers from the short-tail problem, where the final chunk of each document can have an arbitrary length ranging from one sentence up to the window size. FRC can be viewed as an enhanced version of FW by additionally enforcing a lower bound on chunk length, resulting in a narrower chunk-length distribution. We set the window size to 512 tokens.

Fixed Number Chunking (FN). This method greedily concatenates a fixed number of consecutive sentences into each chunk. Similar to FW, the final chunk of a document may contain fewer sentences than the predefined number. Moreover, unlike FW, the resulting chunk length is not directly controlled, as it depends on the lengths of

the sentences. Following Alabi et al. (2025), we set the fixed number to 10 for our experiments.

For both chunking strategies, we employed the same training data generation pipeline as used for FRC. All models were trained on TOWER-7B under the SEP-2c1t format.

As shown in Table 4, the proposed FRC strategy consistently outperforms the conventional FW and FN chunking strategies across almost evaluation settings. The exception is the OOD GlobVDoc dataset, where all chunking strategies achieve similar performance. In terms of d-BLEU, FW and FN remain competitive with FRC on IWSLT2017 en-xx and BWB, although a noticeable gap persists on the IWSLT2017 xx-en. In contrast, GEMBA-DA reveals a clearer advantage for FRC, which surpasses other chunking methods by approximately 1.0 point or more on IWSLT2017 and BWB.

Models	Decode	d-BLEU / Gemba-DA			
		IWS. en-xx	IWS. xx-en	BWB	GlobV.
FS: STAIR	FS4	35.81 / 81.59	59.36 / 83.86	**27.61** / 83.28	41.96 / **86.73**
FRC: SEP	0c1t	**39.61** / 86.29	63.05 / 83.56	27.38 / **83.95**	41.92 / 86.58
	1c1t	39.51 / **86.34**	66.84 / 85.44	27.16 / 83.31	41.87 / 86.53
	2c1t	38.78 / 85.80	**67.20** / **85.73**	26.82 / 83.77	41.98 / 86.42
FW: SEP	0c1t	39.02 / 84.85	62.03 / 82.59	27.15 / 83.21	**42.15** / 86.29
	1c1t	38.67 / 85.41	65.04 / 83.97	26.67 / 83.10	41.95 / 86.38
	2c1t	38.45 / 85.39	65.39 / 84.32	26.51 / 82.64	41.91 / 86.41
FN: SEP	0c1t	39.19 / 84.77	61.10 / 83.02	27.12 / 82.26	42.11 / 86.19
	1c1t	39.09 / 84.61	65.64 / 84.45	27.17 / 82.69	41.92 / 86.10
	2c1t	38.98 / 84.45	66.17 / 84.82	26.79 / 82.38	41.90 / 86.00

Table 4: Comparison of various chunking strategies. **Bold** denotes the highest score achieved on each dataset, while underline indicates the best performance for each SEP model on each dataset.

6.3 Fixed Range

Following the same data curation pipeline, we constructed two additional datasets based on fixed ranges of $[0, 256]$ and $[512, 768]$ (see Appendix E.2 for data statistics). We fine-tuned the TOWER model using the SEP-1c1t strategy, with the results shown in Table 5. Across the three fixed ranges, the performance discrepancies in the en-xx direction are relatively marginal, with the $[0, 256]$ interval performing slightly worse than the other two. Conversely, in the xx-en direction, the $[256, 512]$ range shows a clear advantage. Furthermore, the results consistently substantiate a strong dependency on contextual information for IWSLT xx-en.

6.4 Model Scale

We further examined the SEP-1c1t training strategy, which achieved the best performance on most test datasets, across different model scales using the QWEN2.5 series. The results are presented in Figure 3. As the model size increases from 3B to 14B, both d-BLEU and d-COMET exhibit a strictly increasing trend, indicating that our training strategy is broadly applicable to models of various sizes.

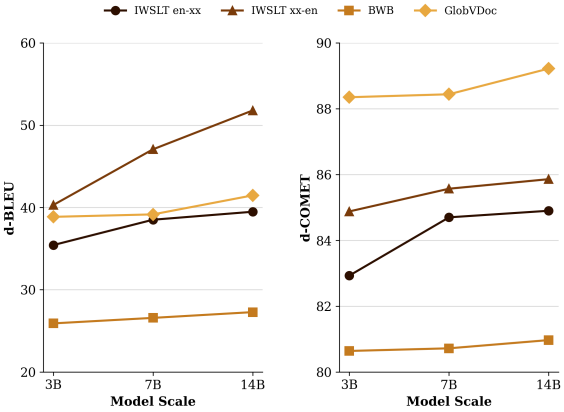


Figure 3: Performance of the SEP-1c1t training strategy under the $[256, 512]$ FRC setting across different model scales in the QWEN2.5 series.

Models	en-xx		xx-en	
	IWSLT	GlobVDoc	IWSLT	BWB
S-0	39.37 / 78.6	41.79 / 84.2	59.12 / 88.3	26.83 / 80.1
S-1	39.50 / 78.7	41.66 / 83.9	62.83 / 89.4	26.76 / 78.7
M-0	39.61 / 79.2	41.92 / **85.5**	61.75 / 89.9	**27.38** / 80.8
M-1	39.51 / 79.1	41.87 / 84.5	**66.84** / **90.6**	27.16 / 79.8
L-0	**39.66** / **79.4**	**42.15** / 85.1	61.70 / 89.0	26.61 / **81.7**
L-1	39.61 / 79.1	41.67 / 84.8	63.62 / 89.5	27.03 / 80.9

Table 5: Results of TOWER-based SEP models trained on fixed ranges S, M, and L, corresponding to $[0, 256]$, $[256, 512]$, and $[512, 768]$, respectively. Results are reported as d-BLEU / LTCR. $*$-0 / $*$-1 denote 0c1t / 1c1t.

6.5 Trade-off: Accuracy, Training Cost and Deployment Efficiency

Although SEP-2c1t achieves the best overall accuracy (Section 5), it requires three times of full-parameter fine-tuning and the deployment of three separate models, resulting in substantial training and deployment overhead. To investigate a more efficient alternative, we replace full-parameter fine-tuning with LoRA (Hu et al., 2022) and analyze the trade-offs among translation quality, training efficiency, and deployment cost.

We adopt LoRA[8] with $r = 16$, $\alpha = 32$, and a dropout rate of 0.05 on the TOWER-7B model. The results are shown in Table 6. While LoRA performs slightly worse than full-parameter fine-

[8]We additionally evaluate a higher-rank configuration $r = 64$. Although a larger rank slightly improves in-distribution performance, it leads to degradation on the OOD GlobVDoc benchmark. Detailed results are provided in Appendix I.

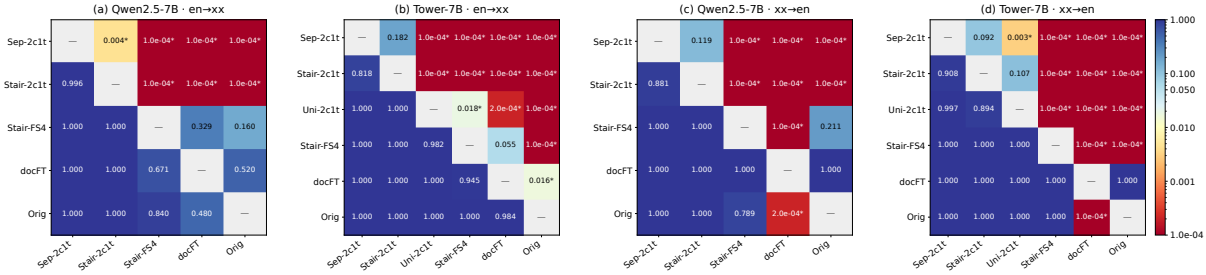


Figure 4: Pairwise one-sided significance tests on IWSLT2017 for QWEN2.5-7B and TOWER-7B. Each cell reports the paired document-bootstrap p-value for the hypothesis that the row system outperforms the column system in d-BLEU. Asterisks indicate $p < 0.05$.

Models	SEP-2c1t (LoRA)			STAIR-2c1t (Full)			SEP-2c1t (Full)		
	0c1t	1c1t	2c1t	0c1t	1c1t	2c1t	0c1t	1c1t	2c1t
Training	3 LoRA			1 Full			3 Full		
Deployment	1 model + 3 adapters			1 model			3 models		
IWSLT en-xx	38.61 / 84.04	38.65 / 84.25	37.85 / 84.26	33.45 / 79.39	38.26 / 84.13	38.29 / 84.14	**39.61** /84.69	39.51 / **84.83**	38.78 / 84.80
IWSLT xx-en	43.04 / 85.06	43.56 / 86.15	43.88 / 86.30	60.46 / 85.37	65.32 / 86.17	66.51 / 86.26	63.05 / 86.09	66.84 / 86.49	**67.20 /86.53**
BWB	25.99 / 80.42	25.96 / 80.52	25.74 / 80.51	27.32 / 80.96	**27.55** / 81.06	27.53 / 81.10	27.38 / **81.11**	27.16 / 81.03	26.82 / 81.03
GlobVDoc	42.78 / 89.02	**43.28 / 89.52**	42.76 / 89.39	41.57 / 88.98	42.16 / 89.34	42.06 / 89.35	41.92 / 89.11	41.87 / 89.25	41.98 / 89.26

Table 6: Comparison of different training and deployment configurations. The values $*$ / $*$ denote d-BLEU and d-COMET, respectively. Within each dataset, the highest score is highlighted in **bold**.

tuning overall, the performance gap remains small. In particular, it yields substantially lower d-BLEU scores on IWSLT xx-en while maintaining competitive d-COMET, suggesting comparable translation quality despite reduced lexical overlap. Furthermore, LoRA surpasses full-parameter fine-tuning on the OOD GlobVDoc benchmark, indicating improved robustness under distribution shifts.

Overall, the three configurations represent different points on the efficiency–performance trade-off. SEP-2c1t (LoRA) offers the best balance between effectiveness and efficiency, STAIR-2c1t (Full) minimizes deployment complexity while maintaining strong performance, and SEP-2c1t (Full) achieves the highest accuracy at the cost of increased training and deployment overhead.

6.6 Statistical Significance Analysis

We assess the reliability of the d-BLEU differences using one-sided paired bootstrap resampling (Koehn, 2004), with documents as the sampling unit and 10,000 bootstrap samples. For each system pair, the alternative hypothesis is that the row system outperforms the column system. We adopt a significance threshold of $p < 0.05$.

The results on IWSLT2017 are shown in Figure 4. SEP-2c1t and STAIR-2c1t significantly outperform STAIR-FS4, DOCFT, and original model across both backbones and translation directions ($p \leq 10^{-4}$) on IWSLT2017. The difference between SEP-2c1t and STAIR-2c1t is significant only for QWEN2.5-7B on en-xx ($p = 0.004$), indicating that their relative advantage is not consistent across settings. For TOWER-7B, UNI-2c1t also significantly outperforms STAIR-FS4 in both directions. Overall, the results suggest the advantage of the 2c1t training strategies over the document-level and fixed-sized systems. The results of the statistical significance tests on BWB and GlobVDoc are provided in Figure 14 in the Appendix.

7 Conclusion

In this study, to handle training-inference length discrepancies caused by document length variability, we introduced FRC, a DP-based strategy that unifies length intervals across training and testing. Building upon FRC, we conducted a systematic analysis of various training paradigms across diverse data formats, finding that separate training exhibited superior stability on in-domain test data, while offering a trade-off analysis among translation quality, training efficiency, and deployment cost. Furthermore, we benchmarked our fine-tuned models against commercial LLMs and agent-based methods on our self-established GlobVDoc dataset. The results showed that our models surpassed QWEN3-30B-A3B-INSTRUCT based agent methods in out-of-distribution performance.

Limitations

Although the GuoFeng dataset is widely used in the current DocMT task, we discovered only upon completing all training tasks that the DocBlocks dataset included GuoFeng's Test 1, Test 2, Valid 1, and Valid 2 data (Xu et al., 2022). However, some chapters of them overlap entirely with Test 3, which was used in the WMT23 and WMT24 Discourse-Level Literary Translation shared tasks (Wang et al., 2023b, 2024b). While this overlap constitutes only a small portion, we elected to exclude the GuoFeng test data from our evaluation to ensure the integrity and stringency of our results.

Since Ramos et al. (2025) did not release the parameters of their trained model, faithfully reproducing their method is challenging. Consequently, in Section 5.2, we compare our approach with theirs only in terms of the relative performance improvement over the corresponding baseline. However, we cannot guarantee that this comparison is fully equivalent.

In Section 6.2, we compare FRC with both FW and FN under the SEP-2c1t setting. Although we controlled the experimental variables as much as possible by using an identical data construction pipeline that differed only in the chunking method, and by training all models with the same hyperparameters and random seed, the resulting training information may still differ because intermediate processing steps, such as chunk alignment and filtering. Furthermore, the SEP-2c1t setting may encourage homogeneous representations, thereby reducing the observable performance differences among the compared methods.

Prior research (Kim et al., 2019; Peng et al., 2025b; O'Brien et al., 2025; Li et al., 2026) suggests that context selection can significantly influence model precision. Consequently, we leave the analysis of how to leverage FRC for optimized context selection as a subject for future work.

To ensure a fair comparison with existing methods, all the models undergo full-parameter fine-tuning, which incurs substantial computational costs. Consequently, most of our analysis and comparative experiments are conducted on a single backbone model under a specific training setting (e.g., TOWER under D2DFT, SEP-1c1t).

Ethical Statement

All models, datasets, and tools utilized in this study are publicly accessible and intended for research purposes.

We ensure that all source material for the GlobVDoc Dataset, sourced from Global Voices, carries a Creative Commons (CC) license. Comprehensive metadata, including titles, authors, and translators, is recorded in Table 20 in the Appendix.

Regarding agent-based methods, we employed GRAFT[9] (Dutta et al., 2025) and DELTA[10] (Wang et al., 2025b) in their original states without modification to the source code.

Furthermore, while the formulation of our proposed reference-based LTCR metric is informed by Lyu et al. (2021), the software implementation was developed independently by the authors.

Acknowledgements

We would like to thank all of our co-authors and the team members who voluntarily contributed to the construction of the GlobVDoc dataset.

We are grateful to all the anonymous reviewers for their constructive feedback, which greatly improved this paper. We also sincerely thank Adam Nohejl, Daiki Matsuoka, Koki Ryu, Ryoma Kumon, Rongzhi Li, Taisei Yamamoto, and Tomoki Doi for their valuable discussions, suggestions, and insightful comments.

This work was supported by JST CREST Grant Number JPMJCR2565, Japan.

References

Jesujoba Oluwadara Alabi, Israel Abebe Azime, Miaoran Zhang, Cristina España-Bonet, Rachel Bawden, Dawei Zhu, David Ifeoluwa Adelani, Clement Oyeleke Odoje, Idris Akinade, Iffat Maab, Davis David, Shamsuddeen Hassan Muhammad, Neo Putini, David O. Ademuyiwa, Andrew Caines, and Dietrich Klakow. 2025. AFRIDOC-MT: Document-level MT corpus for African languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27758–27794, Suzhou, China. Association for Computational Linguistics.

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*.

[9] https://anonymous.4open.science/r/graft-docmt/README.md

[10] https://github.com/YutongWang1216/DocMTAgent

Kushal Arora, Layla El Asri, Hareesh Bahuleyan, and Jackie Cheung. 2022. Why exposure bias matters: An imitation learning perspective of error accumulation in language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 700–710, Dublin, Ireland. Association for Computational Linguistics.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 2–14, Tokyo, Japan. International Workshop on Spoken Language Translation.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy. European Association for Machine Translation.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. Extending context window of large language models via positional interpolation.

Ritvik Choudhary, Rem Hida, Masaki Hamada, Hayato Futami, and Toshiyuki Sekiya. 2025. Exploring context strategies in LLMs for discourse-aware machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24382–24391, Suzhou, China. Association for Computational Linguistics.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2025. Deepseek-v3 technical report.

Himanshu Dutta, Sunny Manchanda, Prakhar Bapat, Meva Ram Gurjar, and Pushpak Bhattacharyya. 2025. GRAFT: A graph-based flow-aware agentic framework for document-level machine translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 2405–2428, Suzhou (China). Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Mara Finkelstein, David Vilar, and Markus Freitag. 2024. Introducing the NewsPaLM MBR and QE dataset: LLM-generated high-quality parallel data outperforms traditional web-crawled data. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1355–1372, Miami, Florida, USA. Association for Computational Linguistics.

Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge.

Jiaxin Guo, Yuanchang Luo, Daimeng Wei, Ling Zhang, Zongyao Li, Hengchao Shang, Zhiqiang Rao, Shaojun Li, Jinlong Yang, Zhanglin Wu, and Hao Yang. 2025. Doc-guided sent2sent++: A sent2sent++ agent with doc-guided memory for document-level machine translation.

Jiaxin Guo, Hao Yang, Zongyao Li, Daimeng Wei, Hengchao Shang, and Xiaoyu Chen. 2024. A novel paradigm boosting translation capabilities of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 639–649, Mexico City, Mexico. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation.

Tatsuya Hiraoka and Kentaro Inui. 2025. Repetition neurons: How do language models produce repetitions? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 483–495, Albuquerque, New Mexico. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Hanxu Hu, Jannis Vamvas, and Rico Sennrich. 2025. Source-primed multi-turn conversation helps large language models translate documents. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23702–23712, Suzhou, China. Association for Computational Linguistics.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565, Seattle, United States. Association for Computational Linguistics.

Linghao Jin, Li An, and Xuezhe Ma. 2024. Towards chapter-to-chapter context-aware literary translation via large language models.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. Association for Computational Linguistics.

Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 24–34, Hong Kong, China. Association for Computational Linguistics.

Shun Kiyono, Sosuke Kobayashi, Jun Suzuki, and Kentaro Inui. 2021. SHAPE: Shifted absolute position embedding for transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3309–3321, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Chen Li, Meishan Zhang, Xuebo Liu, Zhaocong Li, Derek Wong, and Min Zhang. 2024. Towards demonstration-aware large language models for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13868–13881, Bangkok, Thailand. Association for Computational Linguistics.

Ying Li, Xinglin Lyu, Junhui Li, Jinlong Yang, Hengchao Shang, Min Zhang, Shimin Tao, and Daimeng Wei. 2026. Cross-preference learning for sentence-level and context-aware machine translation.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Bin Liu, Xinglin Lyu, Junhui Li, Daimeng Wei, Min Zhang, Shimin Tao, and Hao Yang. 2025. Improving llm-based document-level machine translation with multi-knowledge fusion.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pretraining for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Xinglin Lyu, Junhui Li, Zhengxian Gong, and Min Zhang. 2021. Encouraging lexical translation consistency for document-level neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3265–3277, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sameen Maruf and Gholamreza Haffari. 2018. Document context neural machine translation with memory networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1275–1284,

Melbourne, Australia. Association for Computational Linguistics.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Nikolay Mikhaylovskiy and Ilya Churilov. 2023. Autocorrelations decay in texts and applicability limits of language models.

Hideya Mino, Hitoshi Ito, Isao Goto, Ichiro Yamada, and Takenobu Tokunaga. 2020. Effective use of target-side context for neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4483–4494, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Krzysztof Mrozinski, Minji Kang, Ahmed Khota, Vincent Michael Sutanto, and Giovanni Gatti De Giacomo. 2025. Quality estimation reranking for document-level translation.

Saul B. Needleman and Christian D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48 3:443–53.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Dayyán O'Brien, Bhavitvya Malik, Ona de Gibert, Pinzhen Chen, Barry Haddow, and Jörg Tiedemann. 2025. DocHPLT: A massively multilingual document-level translation dataset. In *Proceedings of the Tenth Conference on Machine Translation*, pages 286–300, Suzhou, China. Association for Computational Linguistics.

OpenAI Team. 2024. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ziqian Peng, Rachel Bawden, and François Yvon. 2025a. Investigating length issues in document-level machine translation. In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 4–23, Geneva, Switzerland. European Association for Machine Translation.

Ziqian Peng, Rachel Bawden, and François Yvon. 2025b. Self-retrieval from distant contexts for document-level machine translation. In *Proceedings of the Tenth Conference on Machine Translation*, pages 220–240, Suzhou, China. Association for Computational Linguistics.

Viet-Thanh Pham, Minghan Wang, Hao-Han Liao, and Thuy-Trang Vu. 2025. Discourse graph guided document translation with large language models.

Hugo Pitorro, Pavlo Vasylenko, Marcos Treviso, and André Martins. 2024. How effective are state space models for machine translation? In *Proceedings of the Ninth Conference on Machine Translation*, pages 1107–1124, Miami, Florida, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz Rojas. 2020. Bifixer and bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal. European Association for Machine Translation.

Miguel Moura Ramos, Patrick Fernandes, Sweta Agrawal, and Andre Martins. 2025. Multilingual contextualization of large language models for document-level machine translation. In *Second Conference on Language Modeling*.

Vikas Raunak, Tom Kocmi, and Matt Post. 2024. SLIDE: Reference-free evaluation for machine translation using a sliding document window. In *Proceedings of the 2024 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 205–211, Mexico City, Mexico. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Anian Ruoss, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Róbert Csordás, Mehdi Bennani, Shane Legg, and Joel Veness. 2023. Randomized positional encodings boost length generalization of transformers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1889–1903, Toronto, Canada. Association for Computational Linguistics.

Rico Sennrich and Martin Volk. 2011. Iterative, MT-based sentence alignment of parallel texts. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 175–182, Riga, Latvia. Northern European Association for Language Technology (NEALT).

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual embeddings with task lora.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. Rethinking document-level neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3537–3548, Dublin, Ireland. Association for Computational Linguistics.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. 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 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Dusan Varis and Ondřej Bojar. 2021. Sequence length is a domain: Length-based overfitting in transformer models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8246–8257, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.

Yu Wan, Baosong Yang, Derek Fai Wong, Lidia Sam Chao, Liang Yao, Haibo Zhang, and Boxing Chen. 2022. Challenges of neural machine translation for short texts. *Computational Linguistics*, 48(2):321–342.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Multilingual e5 text embeddings: A technical report.

Longyue Wang, Siyou Liu, Chenyang Lyu, Wenxiang Jiao, Xing Wang, Jiahao Xu, Zhaopeng Tu, Yan Gu, Weiyu Chen, Minghao Wu, Liting Zhou, Philipp Koehn, Andy Way, and Yulin Yuan. 2024b. Findings of the WMT 2024 shared task on discourse-level literary translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 699–700, Miami, Florida, USA. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023a. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

Longyue Wang, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, Chao-Hong Liu, Yufeng Ma, Weiyu Chen, Yvette Graham, Bonnie Webber, Philipp Koehn, Andy Way, Yulin Yuan, and Shuming Shi. 2023b. Findings of the WMT 2023 shared task on discourse-level literary translation: A fresh orb in the cosmos of LLMs. In *Proceedings of the Eighth Conference on Machine*

Translation, pages 55–67, Singapore. Association for Computational Linguistics.

Xiaotian Wang, Takehito Utsuro, and Masaaki Nagata. 2025a. BiMax: Bidirectional MaxSim score for document-level alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13095–13116, Suzhou, China. Association for Computational Linguistics.

Yutong Wang, Jiali Zeng, Xuebo Liu, Derek F. Wong, Fandong Meng, Jie Zhou, and Min Zhang. 2025b. DelTA: An online document-level translation agent based on multi-level memory. In *The Thirteenth International Conference on Learning Representations*.

Rachel Wicks, Matt Post, and Philipp Koehn. 2024. Recovering document annotations for sentence-level bitext. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9876–9890, Bangkok, Thailand. Association for Computational Linguistics.

Lijun Wu, Xu Tan, Di He, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2018. Beyond error propagation in neural machine translation: Characteristics of language also matter. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3602–3611, Brussels, Belgium. Association for Computational Linguistics.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024a. Adapting large language models for document-level machine translation.

Minghao Wu, Yulin Yuan, Gholamreza Haffari, and Longyue Wang. 2024b. (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *Trans. Assoc. Comput. Linguistics*, 13:901–922.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

Mingzhou Xu, Longyue Wang, Derek F. Wong, Hongye Liu, Linfeng Song, Lidia S. Chao, Shuming Shi, and Zhaopeng Tu. 2022. GuoFeng: A benchmark for zero pronoun recovery and translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11266–11278, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Joonho Yang, Seunghyun Yoon, Hwan Chang, Byeongjeong Kim, and Hwanhee Lee. 2025. Hallucinate at the last in long response generation: A case study on long document summarization.

Long Zhang, Tong Zhang, Haibo Zhang, Baosong Yang, Wei Ye, and Shikun Zhang. 2021. Multi-hop transformer for document-level machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3953–3963, Online. Association for Computational Linguistics.

Jihao Zhao, Zhiyuan Ji, Yuchen Feng, Pengnian Qi, Simin Niu, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2025. Meta-chunking: Learning text segmentation and semantic completion via logical perception.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2024. JudgeLM : Fine-tuned large language models are scalable judges.

A Evaluation Procedures

d-BLEU The computation of the d-BLEU score is contingent upon the translation direction. For the "xx-en" direction, the sacreBLEU (Post, 2018) "13a" tokenization scheme is uniformly applied across the entire corpus. In contrast, for the "en-xx" direction, d-BLEU scores are first computed individually for each target language and subsequently averaged. The tokenization method for the target language is specified as "zh" for Chinese, "ko-mecab" for Korean, and "13a" for all other languages.

d-COMET Although many studies on DocMT use the contextual-embedding COMET (Rei et al., 2022; Vernikos et al., 2022) as their evaluation method, the metric essentially remains sentence-level and incorporates additional contextual information. Therefore, we follow Ramos et al. (2025) and compute d-COMET[11] at the chunk-level using SLIDE (Raunak et al., 2024). Since Ramos et al. (2025) relied solely on overlapping 512-token chunks to compute chunk-level scores independently, which may result in misalignment among the source, hypothesis, and reference, we construct three types of alignment-oriented chunk triplets (src, hyp, ref), and across all strategies, we maintain a 50% overlap between consecutive chunks:

- **Src–Ref Aligned Stacking**: Source segments and their corresponding reference segments are concatenated until one sequence approaches the 512-token limit. Subsequently, the hypothesis is segmented and stacked to approximate the reference length, ensuring that the total count remains under 512 tokens.
- **Segment Independent Stacking**: Source, hypothesis, and reference segments are concatenated independently until each sequence

[11] https://huggingface.co/Unbabel/wmt22-comet-da

Domain	Source	Language Pair	|D|	|S|	|W|	|W|/|D|
Global Voices	GlobVDoc	En → De	10	458	10.4K / 10.8K	1,038 / 1,080
		En → Es	10	596	11.6K / 12.9K	1,162 / 1,292
		En → Fr	10	489	11.2K / 13.4K	1,118 / 1,342
		En → It	10	584	11.9K / 13.2K	1,195 / 1,316
		En → Ko	10	487	10.8K / 21.7K	1,082 / 2,172
		En → Nl	10	492	9.6K / 9.9K	963 / 989
		En → Pt	10	529	10.9K / 11.6K	1,088 / 1,164
		En → Ru	10	492	9.4K / 7.8K	942 / 780
		En → Zh	10	522	12.3K / 19.4K	1,233 / 1,944
		En → Xx	90	4,649	98.2K / 120.8K	1,091 / 1,342

Table 7: Statistics of GlobVDoc Dataset. |D| represents the document count, |S| represents the total sentence count, and |W| indicates the total number of words in both the source language (English) and the corresponding target language (the second language listed in the Language Pair).

reaches its maximum capacity of 512 tokens, disregarding cross-sequence alignment.

- **Token Independent Stacking**: A 512-token sliding window is applied separately to the source, hypothesis, and reference sequences, ignoring segment boundaries and alignment.

The final d-COMET score is computed as the average of the scores derived from these three distinct segmentation methods.

GEMBA-DA GEMBA-DA (Kocmi and Federmann, 2023b) is an LLM-as-a-judge method (Zhu et al., 2024; Gu et al., 2025). We use GPT-4.1 as the judge model and set the temperature to 0 for deterministic evaluation. By conditioning on the source, hypothesis, and reference, the model conducts a direct assessment of translation quality at the document level, and the final score is obtained by averaging scores over the test set. The prompt used for GEMBA-DA evaluation is provided in Appendix K.

B Interest Word Selection for LTCR

Interest words are defined as named entities extracted from the reference text using spaCy[12]. We retain entities corresponding to the following labels: PERSON, PER, PS, ORG, OG, GPE, NORP, FAC, LOC, LC, MISC, PRODUCT, EVENT, WORK_OF_ART, LAW, and LANGUAGE. Only entities appearing at least twice in the reference document are considered for consistency evaluation.

C Details of GlobVDoc

All contributors hold at least a Master's degree and are engaged in research related to natural language processing, specializing in areas such as machine translation, linguistics, and interpretability. The team consisted of native speakers of English, Chinese, Spanish, and Dutch. For sentence alignment involving non-native languages, we conducted cross-verification using multiple high-precision translation tools and multilingual dictionaries. The team members are instructed to follow a standardized formatting protocol: the article title is placed in the first row, the introduction in the second, and the main body from the third row. Moreover, Global Voices articles that relied heavily on multimedia elements, such as videos, images, or audio, are excluded during the document compilation process to ensure textual consistency. In addition, while Global Voices features interview-based articles that may contain personal details, such information is publicly available and does not constitute sensitive or privacy-infringing data. Each document pair is checked by at least two volunteers, and disagreements are discussed and resolved collectively.

The detailed guidelines for data construction are provided below:

1. **Source Selection and Diversity**: English was designated as the source language. To ensure the diversity of the test data and avoid redundancy, each unique English document was utilized only once.

2. **Temporal Relevance**: Priority was given to the most recent documents to ensure the timeliness and contemporary relevance of the data.

3. **Alignment Quality**: We selected documents where sentence-level alignment was as complete as possible. Texts exhibiting significant alignment gaps or structural discrepan-

[12] https://spacy.io/

Domain	Source	Language Pair	|D|	|S|	|W|	|W|/|D|
News	WMT2022	Zh → En	38	505	18.5K	487
Social		Zh → En	25	478	13.3K	532
TED	IWSLT2017	En ↔ Xx	374	37,082	615.8K	1,646
News	News Commentary v11	En → De	155	2,999	56.8K	366
Europarl	Europarl v7	En → De	360	5,134	130.1K	361
Web Novel	GuoFeng Valid1	Zh → En	22	755	18.3K	832
	GuoFeng Test1	Zh → En	22	697	19.5K	884
	GuoFeng Valid2	Zh → En	10	853	16.0K	1.6K
	GuoFeng Test2	Zh → En	12	917	16.7K	1.4K
	BWB	Zh → En	80	2,633	58.5K	731
Global Voices	GlobVDoc (Our)	En → Xx	90	4,649	98.2K	1,092

Table 8: Statistics of commonly used document-level test datasets. |D| denotes the number of documents, |S| denotes the number of sentences, |W| denotes the total number of words in English, and |W|/|D| denotes the average number of words per document. For IWSLT2017, the statistics reported in the table only include en ↔ {de, fr, it, ko, nl, zh} language pairs used in this work.

cies were excluded in favor of more consistent alternatives.

4. **Structural Consistency**: Preference was given to translations that maintained a structural sequence consistent with the source text. This avoided the alignment complexities often introduced by translations that radically reorder sentences within a paragraph.

5. **Manual Segmentation**: Sentence segmentation was performed manually. Rather than strictly decomposing texts into the smallest possible fragments, we merged related short segments to preserve discourse integrity.

6. **Minimal Adjustments**: In cases of minor content misalignment (e.g., where a translator added brief explanatory notes), slight adjustments were made while strictly respecting the original intent. No other modifications to the content were permitted.

7. **Mapping Priority**: In instances of many-to-many or many-to-one sentence mappings, priority was given to establishing complete semantic correspondence.

Detailed counts of documents, sentences, and words in the GlobVDoc dataset are summarized across different language pairs in Table 7. The specifications of the data utilized to construct the dataset, including document IDs, source and translated titles, source and target languages, and the names of authors and translators, are presented in Table 20.

Table 8 compares the statistics of GlobVDoc with those of other commonly used document-level machine translation test datasets. We compute the statistics for the IWSLT2017 and BWB datasets used in this work, while the statistics for the remaining datasets refer to Wang et al. (2023a, 2024b). We do not include GuoFeng Test 3, which was used in the WMT23 and WMT24 Discourse-Level Literary Translation shared tasks, because its reference translations are not publicly available at the time of this work. As shown in the table, the scale of GlobVDoc is comparable to existing commonly used document-level evaluation datasets.

D Quality Estimation of Test Data

The intrinsic quality of the test data warrants careful examination. Following the word-level definition of autocorrelations using distributional semantics proposed by Mikhaylovskiy and Churilov (2023), we define a document-level non-centered autocorrelation function and further extend it to the corpus level.

Given a document $d \in \mathcal{D}$ consisting of units $\{u_1, \ldots, u_{T_d}\}$, the corpus autocorrelation function at lag k, denoted as $R[k]$, is defined as:

$$R[k] = \frac{\sum_{\mathrm{d}} \sum_{t=k}^{T_d} \mathrm{sim}(u_t, u_{t-k}) \cdot \left(1 - \frac{k}{T_d}\right)}{\sum_{\mathrm{d}} (T_d - k) \cdot \left(1 - \frac{k}{T_d}\right)} \tag{3}$$

where the term $(1 - k/T_d)$ serves as a length-penalty weight to mitigate the impact of large variations in document length within the corpus (e.g., in some documents, u_1 and u_4 are still located near

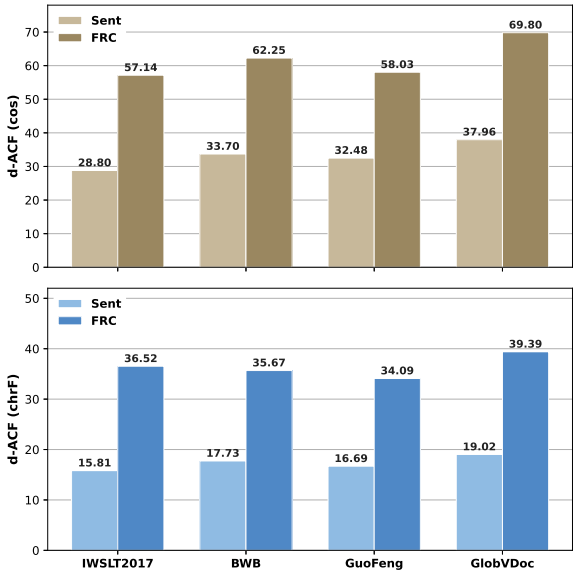


Figure 5: The results of d-ACF ($\times 100$) on four datasets.

the beginning of the document, whereas in others u_4 already appears at the end). We then take a weighted average of the first four lags as the autocorrelation coefficient of the data, referred to as d-ACF:

$$\text{d-ACF} = \sum_{k=1}^{K} w_k R[k] \tag{4}$$

Here, we simply adopt $K = 4$ and uniform weights. The similarity $\text{sim}(u_t, u_{t-k})$ is computed using two approaches: cosine similarity based on sentence embeddings and sentence-level chrF (Popović, 2015). For the former, we employ the jina-embeddings-v3 model (Sturua et al., 2024) for encoding. Our analysis is conducted on the English side[13] of IWSLT2017, BWB, GuoFeng, and GlobVDoc, considering two types of units: sentences and FRC pairs.

As shown in Figure 5, on the cosine-based d-ACF metric, GlobVDoc outperforms all other datasets at both the sentence level and the chunk level, indicating stronger semantic autocorrelation and greater document-level topical coherence. From a lexical perspective, the d-ACF value gap between GlobVDoc and the other datasets is smaller than that observed under the cosine-based setting, but GlobVDoc still maintains a slight advantage.

Due to the above discussion, which considers only autocorrelation values at relatively close intra-document positions (with the lag k limited to 4), we further examine the overall document-level behavior by analyzing the autocorrelation distributions of all datasets across all lag values k. As presented in Figure 6, these distributions are computed at the sentence level using cosine similarity.

[13]Although the original languages of some datasets are not English, both the chrF metric and the performance of the jina-embedding-v3 model exhibit obvious variation across different languages. Therefore, to maintain relative fairness, we conduct the evaluation on the common English side.

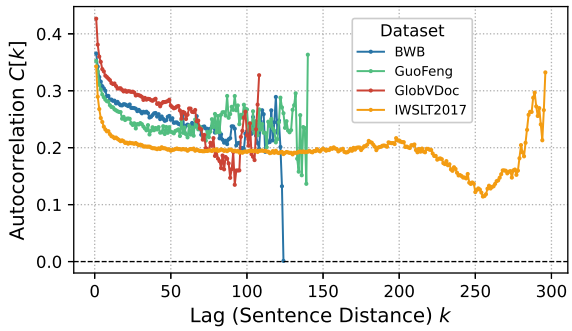


Figure 6: Autocorrelation distributions of all datasets across all lag values k.

We observe that, at the initial stage of the curves, all datasets exhibit a sharp decline followed by a more gradual decrease, indicating that strong semantic relatedness is primarily concentrated among sentences that are close to each other. Furthermore, both GlobVDoc and IWSLT2017 show another significant drop followed by an increase toward the end of the curves, resulting in an overall U-shaped pattern. In particular, for IWSLT2017, the notable rise after around lag $k = 290$ suggests that certain documents in the corpus may exhibit strong correspondences between their beginnings and endings. In contrast, BWB and GuoFeng do not display a clear U-shaped trend in the later stages. Instead, after the steady decline, their curves gradually increase. This indicates that although semantic correlation at medium distances continues to weaken, these datasets still demonstrate meaningful long-range coherence at larger sentence separations.

E Experimental Setups

E.1 Hyperparameters for FRC and Chunk Alignment

For Fixed-Range Chunking, we set the maximum number of chunks allowed to fall outside the predefined range during the relaxed optimization stage to $K = 8$. This setting is intended to improve the robustness of the algorithm. Under this configuration, no invalid samples were encountered during either training data construction or test-time chunk generation.

Lang Pair	Median Ratio [L → B]	Outlier Count [L → B]	Outlier%↓ (L)	Outlier%↓ (B)	Delta↓ [L → B]	Median Ratio [Q → T]	Outlier Count [Q → T]	Outlier%↓ (Q)	Outlier%↓ (T)	Delta↓ [Q → T]
en↔de	1.56 → 1.57	3,847 → 22,038	1.40%	8.02%	+6.62%	1.56 → 1.62	3,847 → 4,058	1.40%	1.35%	-0.05%
en↔es	1.41 → 1.41	2,007 → 5,923	0.71%	2.10%	+1.39%	1.41 → 1.50	2,007 → 2,217	0.71%	0.70%	-0.01%
en↔fr	1.50 → 1.50	3,061 → 9,609	1.14%	3.57%	+2.43%	1.50 → 1.56	3,061 → 3,342	1.14%	1.13%	-0.01%
en↔it	1.53 → 1.53	2,235 → 10,135	1.05%	4.75%	+3.70%	1.53 → 1.56	2,235 → 2,463	1.05%	1.07%	+0.02%
en↔ko	1.51 → 1.51	557 → 2,339	3.31%	13.91%	+10.60%	**1.51 → 2.51**	557 → 1,087	3.31%	4.34%	+1.03%
en↔nl	1.61 → 1.61	5,587 → 25,912	2.58%	11.99%	+9.41%	1.61 → 1.66	5,587 → 5,991	2.58%	2.54%	-0.04%
en↔pt	1.45 → 1.45	2,921 → 9,643	1.45%	4.79%	+3.34%	1.45 → 1.59	2,921 → 3,269	1.45%	1.45%	0.00%
en↔ru	1.75 → 1.75	746 → 1,793	0.64%	1.55%	+0.91%	1.75 → 1.89	746 → 632	0.64%	0.48%	-0.16%
en↔zh	0.88 → 0.81	**20,396 → 232,791**	**4.48%**	**51.13%**	**+46.65%**	**0.88 → 1.34**	20,396 → 40,353	4.48%	6.37%	+1.89%

Table 9: Outlier sample statistics across different similarity approaches and tokenizers. **L** denotes chunk alignment methods using the QWEN2.5-7B-INSTRUCT tokenizer paired with either LaBSE embedding cosine similarity. And **B** denotes nllb-200-distilled-600M translation based sentence-BLEU as the similarity score, respectively. **Q** and **T** represent the use of QWEN2.5-7B-INSTRUCT and TOWER-7B-INSTRUCT tokenizers while maintaining LaBSE-based cosine similarity as the alignment criterion. "en↔xx" encompasses bidirectional data, the length ratio is uniformly calculated for all pairs as the length of the non-English language relative to the English text.

For the Dual-Boundary Matching-based Chunk Alignment algorithm, we set $\lambda = 0.5$ and $\sigma = 0.05$. In addition, the number k of target-unit candidates considered for each source boundary is set to 256.

E.2 Training Data Filtering and Statistics

Given that the DocBlocks dataset comprises high-quality data filtered via quality-aware methods (Rei et al., 2023; Ramírez-Sánchez et al., 2020), we adopt a simple strategy to process chunk pairs:

1. Exclude chunk pairs where the source chunk length exceeds the upper threshold of the predefined range.

2. Determine the median length ratio for each language pair, and then filter the parallel chunk pairs by retaining only those within the interval $[\text{median}/1.5, \text{median} \times 1.5]$.

We utilize these outlier samples as breakpoints to further segment documents into subdocuments. Table 9 presents the statistics for outlier samples, which can also characterize the data attrition rate of our proposed Dual-Boundary Matching-based Chunk Alignment Algorithm.

In addition to the LaBSE-based cosine similarity utilized for chunk alignment in Section 5.1.1, we introduce a comparative strategy using sentence-BLEU scores derived from the nllb-200-distilled-600M model[14] (NLLB Team et al., 2022) translations. Both methods employ the QWEN2.5-7B-INSTRUCT tokenizer for length measurement during the initial FRC phase. Our results indicate that while both approaches yield nearly identical median ratios due to the shared tokenizer, the translation-based method leads to a substantial increase in the outlier rate, particularly for zh. This discrepancy likely stems from the high proportion of GuoFeng and BWB data within the DocBlocks dataset. As these materials belong to the web novel domain, they pose relative translation challenges.

To quantitatively assess the difference in alignment quality between chunk alignment based on SentenceBLEU scores and that based on LaBSE-based cosine similarity, we also trained QWEN2.5-7B-INSTRUCT model using the SEP-1c1t strategy on samples obtained with the former method. A simple comparison of the results is shown in Figure 7. After data filtering, the LaBSE-based method, which retained a larger number of samples, unsurprisingly achieves consistently better performance on the test datasets than the SentenceBLEU-based approach. Notably, although the SentenceBLEU-based approach discarded a substantial number of outlier Chinese data points, its performance on the BWB dataset is not far behind that of the LaBSE-based approach. This suggests that the performance gains brought by large-scale Chinese training data exhibit a clear pattern of diminishing marginal returns.

Provided that LaBSE cosine similarity is used consistently, the sample attrition rates for FRC using either the QWEN2.5-7B-INSTRUCT or TOWER-7B-INSTRUCT tokenizers do not differ significantly. However, inherent discrepancies between these tokenizers lead to substantial variations in cross-lingual length ratios, a phenomenon particularly prominent in ko and zh.

Subsequently, based on these subdocuments and their corresponding FRCs, we construct training samples tailored to the formatting requirements of the D2DFT, STAIR, SEP, and UNI models. Ta-

[14] https://huggingface.co/facebook/nllb-200-distilled-600M

Lang Pair	Total Samples	Median Ratio	Outlier Count	Outlier%↓
		$[0,256] \rightarrow [256,512] \rightarrow [512,768]$		
en↔de	894,029 → 301,176 → 179,509	1.62 → 1.62 → 1.62	32,799 → 4,058 → 2,124	3.67% → 1.35% → 1.18%
en↔es	934,705 → 314,982 → 187,526	1.50 → 1.50 → 1.50	14,699 → 2,217 → 1,304	1.57% → 0.70% → 0.70%
en↔fr	877,691→ 295,207 → 176,007	1.56 → 1.56 → 1.56	25,941 → 3,342 → 1,859	2.96% → 1.13% → 1.06%
en↔it	683,886 → 230,570 → 137,893	1.56 → 1.56 → 1.56	17,819 → 2,463 → 1,376	2.61% → 1.07% → 1.00%
en↔ko	73,353→ 25,073 → 15,034	2.51 → 2.51 → 2.51	9,061 → 1,087 → 466	12.4% → 4.34% → 3.10%
en↔nl	700,761 → 235,593 → 140,879	1.66 → 1.66 → 1.67	45,464 → 5,991 → 3,226	6.49% → 2.54% → 2.29%
en↔pt	669,903→ 225,892 → 135,099	1.59 → 1.59 → 1.59	18,179 → 3,269 → 1,859	2.71% → 1.45% → 1.38%
en↔ru	393,606 → 132,690 → 79,083	1.88 → 1.89 → 1.89	9,271 → 632 → 197	2.36% → 0.48% → 0.25%
en↔zh	1,862,572→ 633,248 → 381,331	1.35 → 1.34 → 1.33	289,068 → 40,353 → 13,312	15.51% → 6.37% → 3.49%

Table 10: Sample statistics across different fixed ranges. Data samples were curated using LaBSE-based cosine similarity for chunk alignment and processed with the TOWER-7B-INSTRUCT tokenizer. The sequence $* \rightarrow * \rightarrow *$ denotes the data corresponding to fixed-range intervals of $[0,256]$, $[256,512]$, and $[512,768]$, respectively.

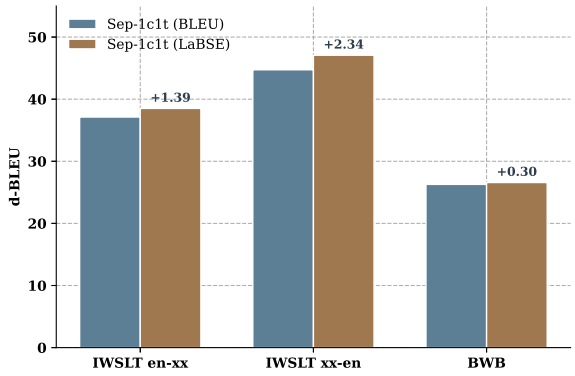


Figure 7: Comparison of QWEN2.5-7B-INSTRUCT trained with SEP-1c1t on samples derived from chunk alignment using SentenceBLEU scores v.s. LaBSE-based cosine similarity.

Model	Tokenizer	
	QWEN2.5	TOWER
D2DFT	241,828	241,828
STAIR-FS4	953,631	1,016,814
STAIR-1c1t	2,004,084	2,331,028
STAIR-2c1t	2,004,084	2,331,028
(SEP:) 0c1t	2,004,084	2,331,028
(SEP:) 1c1t	1,743,169	2,056,597
(SEP:) 2c1t	1,494,053	1,797,915
UNI-2c1t	–	6,185,540

Table 11: The number of training data utilized for each model configuration.

ble 11 reports the number of final training samples under the fixed range of $[256,512]$, while Table 10 presents the statistics for the three fixed-range datasets. In addition, Table 10 reports only the number of samples for en↔xx language pairs. However, there are also a very small number of samples corresponding to language pairs that do not include English. For example, under the fixed-range setting of $[256,512]$, there are 63,403 such samples. These samples were also included in the final training dataset.

To further illustrate the effect of FRC on length control, Figure 8 visualizes the training- and test-time source-side input length distributions of different training formats under the TOWER backbone with the fixed range $[256,512]$. Overall, D2DFT and STAIR-FS4 exhibit broader or less-aligned train-test length distributions, whereas SEP-2c1t, STAIR-2c1t, and UNI-2c1t, show more compact and better-aligned distributions. Notably, the three sub-models of SEP-2c1t exhibit particularly sharp and narrowly bounded length distributions.

E.3 Training and Decoding Setups

We employ supervised fine-tuning (SFT) to train our models, where the loss is computed only on the target-side tokens. Given a document pair $d = (d^s, d^t) \in \mathcal{D}$ after FRC-based chunk alignment, we denote the aligned chunk pairs as $\{(c^s_{d,i}, c^t_{d,i})\}_{i=1}^{|d|}$. $C^*_{d,i}$ is the set of contextual source chunks used for predicting the i-th target chunk.

D2DFT

$$\mathcal{L}_{\text{D2DFT}} = -\sum_{d\in\mathcal{D}} \log p_\theta\big(d^t \mid d^s\big) \tag{5}$$

$$\theta_{\text{D2DFT}} = \arg\min_\theta \mathcal{L}_{\text{D2DFT}}(\theta) \tag{6}$$

STAIR-2c1t

$$C^{\text{STAIR-2c1t}}_{d,i} = \begin{cases} \varnothing, & i = 1, \\ \{c^s_{d,1}\}, & i = 2, \\ \{c^s_{d,i-2}, c^s_{d,i-1}\}, & i \geq 3. \end{cases} \tag{7}$$

$$\mathcal{L}_{\text{STAIR-2c1t}} = -\sum_{d\in\mathcal{D}} \sum_{i=1}^{|d|} \log p_\theta\big(c^t_{d,i} \mid c^s_{d,i}, C^{\text{STAIR-2c1t}}_{d,i}\big) \tag{8}$$

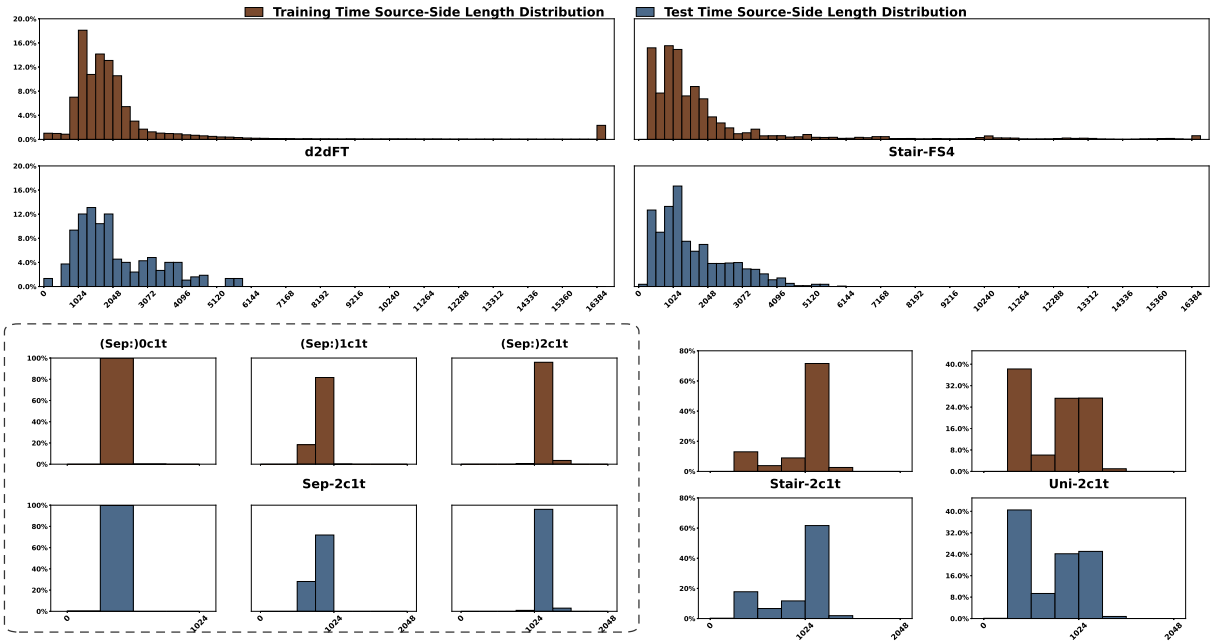


Figure 8: Training- and test-time (source-side) input length distributions. The x-axis denotes input length, and the y-axis denotes the proportion of documents or chunks in each length bin. Lengths are measured using the TOWER's tokenizer. Training statistics are computed on DocBlocks, and test statistics are computed on IWSLT2017 en-xx.

$$\theta_{\text{STAIR-2c1t}} = \arg\min_{\theta} \mathcal{L}_{\text{STAIR-2c1t}}(\theta) \quad (9)$$

STAIR-FS4 We merge the aligned chunk pairs into four consecutive chunks within one document, defined as $\{(z^s_{d,i}, z^t_{d,i})\}_{i=1}^4$.

$$C^{\text{STAIR-FS4}}_{d,i} = \begin{cases} \varnothing, & i = 1, \\ \{z^s_{d,1}\}, & i = 2, \\ \{z^s_{d,1}, z^s_{d,2}\}, & i = 3 \\ \{z^s_{d,1}, z^s_{d,2}, z^s_{d,3}\}, & i = 4. \end{cases} \quad (10)$$

$$\mathcal{L}_{\text{STAIR-FS4}} = -\sum_{d \in \mathcal{D}} \sum_{i=1}^{4} \log p_\theta\left(z^t_{d,i} \mid z^s_{d,i}, C^{\text{STAIR-FS4}}_{d,i}\right) \quad (11)$$

$$\theta_{\text{STAIR-FS4}} = \arg\min_{\theta} \mathcal{L}_{\text{STAIR-FS4}}(\theta) \quad (12)$$

SEP / UNI-2c1t

$$C^{(k)}_{d,i} = \begin{cases} \varnothing, & k = \text{0c1t}, \\ \{c^s_{d,i-1}\}, & k = \text{1c1t}, \\ \{c^s_{d,i-2}, c^s_{d,i-1}\}, & k = \text{2c1t}. \end{cases} \quad (13)$$

$$\mathcal{L}^{(k)} = -\sum_{d \in \mathcal{D}} \sum_{i=1}^{|d|} \log p_\theta\left(c^t_{d,i} \mid c^s_{d,i}, C^{(k)}_{d,i}\right) \quad (14)$$

$$\theta^{(k)}_{\text{SEP-2c1t}} = \arg\min_{\theta} \mathcal{L}^{(k)}(\theta), k \in \{\text{0c1t}, \text{1c1t}, \text{2c1t}\} \quad (15)$$

$$\theta_{\text{UNI-2c1t}} = \arg\min_{\theta} \sum_{k} \mathcal{L}^{(k)}(\theta) \quad (16)$$

The training and inference prompts are presented in the Appendix K.

The training hyperparameters almost follow the configurations established by Ramos et al. (2025). However, we adjust the batch size to 16 for *Doc2Doc* fine-tuning while setting it to 64 for all other models. All training processes were conducted using 8 NVIDIA H100 GPUs, optimized by AdamW (Loshchilov and Hutter, 2019). Table 12 contains the detailed hyperparameter configuration for the training step.

Batch size	16 (D2DFT) / 64 (others)
Number of Epochs	2
Learning rate	7×10^{-6}
LR Scheduler	cosine
Warmup Steps	125
Weight Decay	0.01
Optimizer	AdamW
Adam β_1	0.9
Adam β_2	0.999
Adam ϵ	1×10^{-8}
Maximum Sequence Length	32,768

Table 12: Hyperparameter configuration to fine-tune models.

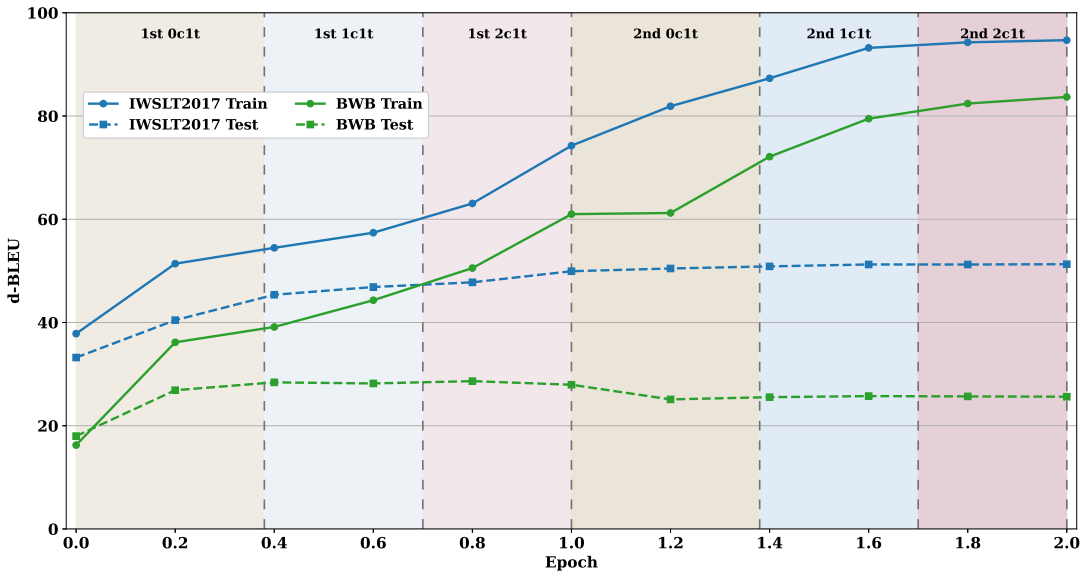


Figure 9: Training Process of TOWER-based UNI-2c1t model.

Both inference and evaluation were conducted on a single H100 GPU. For the decoding process, we utilized vLLM (Kwon et al., 2023) and employed a greedy decoding strategy.

During inference, d2d refers to direct document-to-document translation. The ∗c1t models utilize a staged decoding approach: 0c1t translates all fixed-range chunks without context; 1c1t initiates the first chunk of each document in 0c1t mode, while subsequent chunks follow the 1c1t format; 2c1t decoding begins with 0c1t for the first chunk and 1c1t for the second, with all remaining chunks processed using 2c1t context. FS4 follows a similar staged procedure. All inference prompts are kept consistent with the training prompts shown in Appendix K.

F Analysis of Joint Training

F.1 Training Process of UNI-2c1t

Given that the training data for UNI-2c1t reaches 6M samples (see Table 11) and that the model repeatedly encounters each source chunk three times through the 0c1t, 1c1t, and 2c1t formats, we sought to investigate the potential risk of overfitting. We monitored the performance of the UNI-2c1t model on both the test set and a subset of the training data throughout the training process. The training subset was constructed via random sampling, consisting of 12 document pairs selected from each of the IWSLT2017 and BWB datasets.

The results are illustrated in Figure 9. Since the IWSLT2017 training data is essentially bi-directional (xx-xx), we report the averaged performance across en-xx and xx-en directions on the test set. During the training process, model checkpoints were saved every 0.2 epochs. Since the training data is in a fixed sequence of 0c1t, 1c1t, and 2c1t formats within each epoch, we have explicitly marked the specific time points at which the learning phase for each data format concludes.

Regarding the training data, both the IWSLT2017 and BWB subsets exhibit a natural and consistent improvement in fitting as training progresses. For the test data, the performance on IWSLT2017 demonstrates a gradual increase before eventually plateauing, with no discernible signs of overfitting. However, a distinct trend emerges for the BWB data during the transition from epoch 1.0 to 1.2: we observe a marginal gain on the training set accompanied by a performance dip on the test set. This suggests that the transition in data format, reverting from 2c1t back to 0c1t at the start of a new epoch, may introduce temporary optimization confusion for the BWB domain. In the subsequent phased learning stages, while the model continues to fit the training data rapidly, its generalization performance on the test data eventually stabilizes. However, given the model's exceptionally high degree of fitting to the training data, its generalization performance may already have been limited.

F.2 Analysis of Context Sensitivity

To better understand the behavior of UNI-2c1t under joint training, we conduct additional analyses

Model	Decoding	IWSLT en-xx		IWSLT xx-en		BWB xx-en	
		Orig-C	RDM-C	Orig-C	RDM-C	Orig-C	RDM-C
SEP-2c1t	1c1t	39.26	37.86 (-1.40)	67.04	59.85 (-7.19)	26.55	23.37 (-3.18)
	2c1t	38.40	37.02 (-1.38)	67.26	57.70 (-9.56)	26.15	21.96 (-4.19)
STAIR-2c1t	1c1t	39.12	37.43 (-1.69)	65.79	57.72 (-8.07)	27.18	23.12 (-4.06)
	2c1t	39.16	37.60 (-1.56)	66.97	56.33 (-10.64)	27.18	22.16 (-5.02)
UNI-2c1t	1c1t	36.36	36.13 (**-0.23**)	66.28	64.25 (**-2.03**)	25.22	24.72 (**-0.50**)
	2c1t	36.55	35.84 (**-0.71**)	66.43	63.66 (**-2.77**)	25.28	24.28 (**-1.00**)

Table 13: d-BLEU scores under original context (Orig-C) and randomly replaced context (RDM-C). Numbers in parentheses denote the absolute performance drop after context replacement. Results are averaged over five runs.

Model	Original			RDM1		RDM2		RDM3	
	0c1t	1c1t	2c1t	1c1t	2c1t	1c1t	2c1t	1c1t	2c1t
STAIR-2c1t	49.29	52.36	56.99	47.95	46.63	47.08	45.29	46.52	44.56
	–	–	–	-8.42%	-18.18%	-10.08%	-20.53%	-11.15%	-21.81%
UNI-2c1t	97.35	97.50	97.35	95.40	95.27	96.62	95.78	96.14	95.22
	–	–	–	**-2.15%**	**-2.14%**	**-0.90%**	**-1.61%**	**-1.39%**	**-2.19%**

Table 14: d-BLEU on sampled training instances under different decoding formats and after random context replacement. Percentages denote the relative performance drop compared with the corresponding original score.

to investigate whether the model effectively utilizes contextual information or instead primarily relies on the source chunk itself.

Context replacement: Test. We randomly replace the context associated with each source chunk in the in-distribution test set while keeping the source chunk and target chunk unchanged. To maximize the perturbation, the replacement context is sampled from a different document written in a different language. During evaluation, we remove the first chunk of each document since it has no preceding context. We denote the original context as Orig-C and the replaced context as RDM-C. All results are averaged over five independent runs.

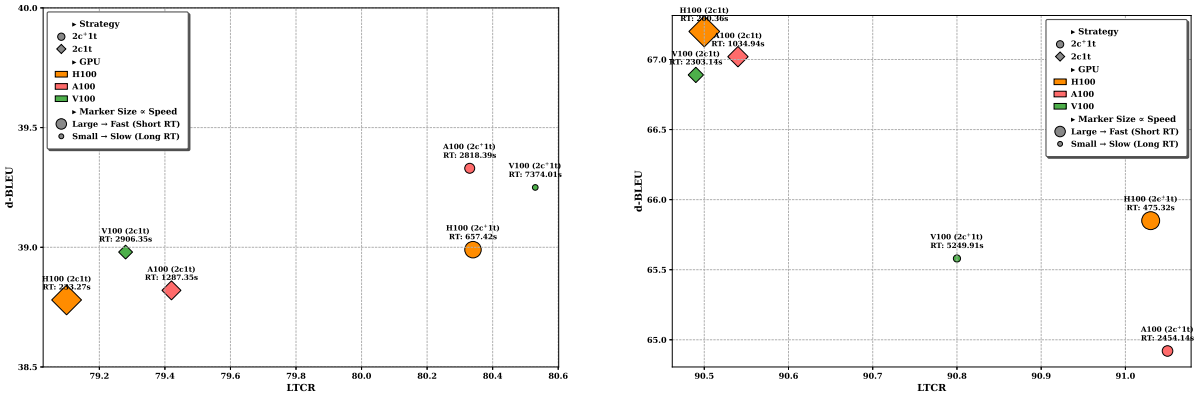
As shown in Table 13, replacing the context causes substantial performance degradation for both SEP-2c1t and STAIR-2c1t across all test sets. In contrast, UNI-2c1t exhibits only marginal performance drops under the same perturbation. This observation suggests that UNI-2c1t is considerably less sensitive to contextual information, even when the original context is replaced with unrelated content.

Context replacement: Train. To further investigate this phenomenon, we sample 320 source chunks from the training set. These chunks appear only in the 2c1t format for STAIR-2c1t, while they appear in all three formats (0c1t, 1c1t, and 2c1t) for UNI-2c1t. We first evaluate how well each model fits these training instances under different decoding formats and then perform the same context replacement experiment.

Table 14 reveals two notable observations. First, STAIR-2c1t achieves its highest fitting accuracy under the 2c1t decoding format, whereas UNI-2c1t attains nearly perfect d-BLEU scores across all three decoding formats, indicating that it fits the training data well regardless of the decoding format. Second, replacing the context results in approximately a 20% performance drop for STAIR-2c1t, while UNI-2c1t shows only negligible degradation. These results suggest that, although UNI-2c1t is trained with contextual information, the learned representations become largely insensitive to context. Instead, the model appears to rely primarily on the source chunk itself when generating translations. This behavior may partially explain why jointly training all three data formats does not perform as well as other models.

F.3 Training Cost

Regarding computational efficiency, the (SEP:) 0c1t, 1c1t, and 2c1t sub-models required 39h 48m, 37h 21m, and 36h 42m respectively on 8× H100 GPUs, totaling 113h 51m. In contrast, the UNI-2c1t model required 119h 35m under the same hardware configuration. Given its shorter total training time and the capacity for simultaneous parallel training of its sub-models, the SEP approach



(a) Results on IWSLT2017 en-xx.

(b) Results on IWSLT2017 xx-en.

Figure 10: Results of Source-Only Context v.s. Source-MT Context on the IWSLT2017 dataset. "RT" denotes runtime. The marker size is directly proportional to the decoding speed and inversely proportional to the runtime.

can offer a higher degree of training efficiency.

G Source-Only v.s. Source-MT Context

G.1 Translation Quality–Efficiency Trade-off

We employ the same FRC dataset that was constructed previously. However, we introduce reference translations into the context parts of the 1c1t and 2c1t data, naming these augmented formats as $1c^{+}1t$ and $2c^{+}1t$. Notably, due to cross-lingual differences in lexical representation, appending target-language text to the context precludes a unified input-side length distribution. As a result, the length distribution transitions to a language-wise paradigm.

The results are presented in Figure 10. Utilizing the TOWER-based model, we evaluated the two strategies in terms of accuracy, term consistency, and efficiency across various hardware configurations on the IWSLT2017 dataset. However, it is worth noting that, although the hyperparameter configurations remained strictly consistent for the comparative experiments conducted across different GPUs, we cannot guarantee completely identical execution environments. Specifically, the H100 trials were deployed on an AArch64 architecture, whereas the A100 and V100 experiments were executed on an x86-64 architecture. We utilize the total runtime (RT) required to execute the complete decoding script as the primary metric for evaluating decoding efficiency.

As a consistent trend across all three hardware environments, under the $2c^{(+)}$ configuration, incorporating previous translations naturally yields superior terminological consistency (Choudhary et al., 2025), alongside marginally higher d-BLEU scores in the IWSLT2017 en-xx direction compared to the Source-Only method. Conversely, the Source-Only approach exhibits a highly significant edge in decoding efficiency, paired with a distinct accuracy advantage in the xx-en direction. Specifically, the decoding latency under the Source-MT context is, on average, approximately 2.5 that of the Source-Only baseline. This efficiency gap is readily apparent even on high-speed GPUs such as the H100, and becomes especially pronounced on comparatively slower hardware like the V100. This is because leveraging previous translations as context inherently requires strictly sequential decoding, in contrast to the Source-Only approach, which permits all chunks to be preprocessed and decoded in parallel. Given that document-level machine translation is fundamentally a long-sequence generation task, decoding efficiency remains an indispensable consideration for practical, real-world deployment.

To further evaluate the translation quality of both approaches, we conducted comprehensive experiments across all datasets using a single H100 GPU for inference. The results are summarized in Table 15. In terms of LTCR metric, the $1c^{+}1t$ configuration demonstrates comprehensive superiority. This advantage is particularly pronounced on the BWB dataset, where it approaches the peak performance of the STAIR-FS4 method (achieved under the Source-Only setting) and surpasses large-scale LLMs (see Figure 2). Given the characteristics of the BWB dataset's web novel domain, domain-specific terminology (e.g., character names) likely appears at a higher frequency than in other evaluated domains. Consequently, incorporating the previous translation as context yields the most sub-

Dataset	Lang	d-BLEU ↑						LTCR ↑						NRR (%) ↓					
		1c1t			2c1t			1c1t			2c1t			1c1t			2c1t		
		Base	c^+	Δ_1	Base	c^+	Δ_2	Base	c^+	Δ_1	Base	c^+	Δ_2	Base	c^+	Δ_1	Base	c^+	Δ_2
IWSLT2017	en-xx	**39.51**	39.35	−0.16	38.78	**38.99**	+0.21	79.12	**80.06**	+0.94	79.10	**80.34**	+1.24	**1.34**	1.87	+0.53	1.60	1.60	0.00
	xx-en	**66.84**	65.88	−0.96	**67.20**	65.85	−1.35	90.59	**90.95**	+0.36	90.52	**91.03**	+0.51	**0.80**	1.87	+1.07	**0.53**	1.07	+0.54
BWB	xx-en	**27.16**	26.91	−0.25	**26.82**	26.26	−0.56	79.81	**83.73**	+3.92	79.79	**83.08**	+3.29	0.00	0.00	0.00	0.00	0.00	0.00
GlobVDoc	en-xx	41.87	**42.24**	+0.37	**41.98**	41.54	−0.44	84.52	**85.70**	+1.18	84.60	**85.59**	+0.99	0.00	0.00	0.00	**0.00**	1.11	+1.11

Table 15: Comparison among 1c1t, $1c^+1t$, 2c1t, and $2c^+1t$ across datasets. "Base" and "c^+" represent 1c1t and $1c^+1t$, respectively. Δ_1 denotes $1c^+1t - 1c1t$, and Δ_2 denotes $2c^+1t - 2c1t$. The better result for each metric on each dataset in the comparison is highlighted in bold.

stantial improvement in terminological consistency. However, regarding translation accuracy measured by d-BLEU, the Source-Only approach maintains the advantage in the majority of cases. This performance is strongly negatively correlated with the NRR; a lower NRR indicates a more stable generation output.

G.2 Exposure Bias: A Case Study

Furthermore, for the c^+ strategy, conditioning on previous translations is vulnerable to exposure bias and error propagation (Wu et al., 2018; Arora et al., 2022): a suboptimal translation used as context can lead to cascading deterioration in the quality of subsequent chunk translations within the same document (Mino et al., 2020; Zhang et al., 2021). To illustrate this point, we provide a concise analytical overview alongside a qualitative case study. We define "broken chunks" as outputs with log probabilities (lp) below -5, and "repetition chunks" that contain n-gram repetitions. The distribution of these broken chunks across various languages under the $1c^+1t$ setting on the IWSLT2017 en-xx data is presented in Table 16. Notably, the number of repetition chunks exceeds that of broken chunks. This discrepancy arises because, in some instances, the repetition initiates near the tail end of the chunk. Consequently, the overall log probabilities for the chunk may not drop below the -5 threshold (typically hovering between -3 and -5). Importantly, these specific chunks frequently act as the initial trigger for cascading deterioration, as illustrated in Figure 11.

We categorize the model's capacity to return to standard output following a generation collapse into three distinct groups: Full Recovery indicates complete restoration after a brief degradation; Partial Recovery denotes a return to normal generation but with a sustained drop in output quality; and Failed to Recover signifies a continuous collapse until the end. In our analysis, we identified a total of 22 repetition chunks. Excluding those

Lang	Chunks	Broken	Repetition	Mean Log-Prob
ko	519	8	13	-3.496
zh	520	9	7	-3.354
de	460	1	2	-0.349
fr	521	0	0	-0.234
it	92	0	0	-0.238
nl	94	0	0	-0.285

Table 16: Distribution of broken and repetition chunks across languages on the IWSLT en-xx dataset (TOWER-based SEP-$1c^+1t$ model.

situated at the end of a document, 19 chunks remained. Strikingly, only two of them (the specific instances detailed in our case study) managed to recover when conditioned on a degraded translation context. This yields an error propagation rate of 89% (17/19), demonstrating that a mid-translation quality collapse is highly likely to trigger a subsequent cascading failure. In contrast to the independent decoding mechanism of the Source-Only context approach, this susceptibility to error propagation represents a critical latent issue that must be addressed when utilizing previous target translations as context.

H Test-Time Chunking

H.1 Length Analysis at Test Time

As shown in Figure 12, we perform a two-type length-centric analysis. First, we investigate test-time performance across varying FRC length intervals. Second, we increase test data lengths by concatenating adjacent chapters from the same book in the BWB dataset until full books are formed. Evaluation is conducted at the book level by concatenating all translated chapters for each of the six books.

For the STAIR-FS4 and D2DFT, d-BLEU scales positively as the FRC increases. However, when the length of the test data itself exceeds a certain threshold, performance begins to degrade. This indicates that while both models generalize effectively within their operational range, their capacity

Error Propagation and Recovery Across Chunks

En → Ko ($n = 6$ chunks) *Status: Partial Recovery*

Chunk No.	Log-Prob	State	Description
0, 1	≈ -0.71	Normal	Standard Korean output.
2	-3.43	Trigger	**[REP]** Initial repetition trigger.
3	-368.97	Broken	Severely collapsed: "의 의 의 의 ..."
4, 5	-1.08 ± 0.3	Recovered	Returns to normal flow.

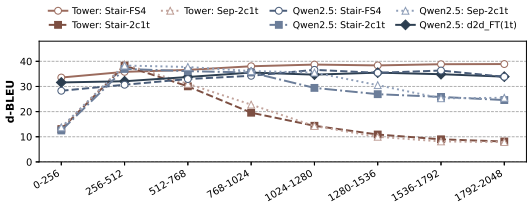
En → Zh ($n = 11$ chunks) *Status: Failed to Recover*

Chunk No.	Log-Prob	State	Description
0, 1	≈ -1.00	Normal	Standard Chinese output.
2	-6.63	Trigger	**[REP]** Initial repetition trigger.
3	-518.00	Broken	Deep failure: "计 计 计 计 ..."
$4 \sim 10$	$-32 \sim -212$	Persistent	Never recovers; remains in error state.

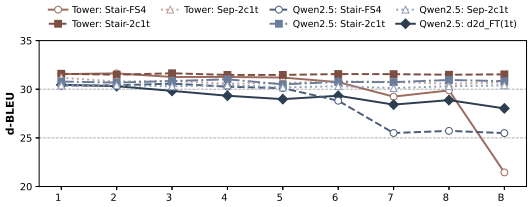
En → De ($n = 7$ chunks) *Status: Full Recovery*

Chunk No.	Log-Prob	State	Description
0	-1.48	Trigger	**[REP]** Early trigger.
1	-36.54	Broken	Broken; inherits errors from No.0 Chunk.
$2 \sim 6$	≈ -0.25	Full Rec.	Rapid and stable recovery.

Figure 11: A Case Study: Potential Exposure issues in the Source-MT Context Setting.



(a) d-BLEU results on IWSLT2017 en–xx with increasing FRC length intervals.



(b) d-BLEU results on BWB zh–en under progressive connection of adjacent chapters.

Figure 12: Test-time analysis of length variation.

is strictly bounded by the training data distribution.

Both the STAIR-2c1t and SEP-2c1t models achieve peak accuracy only when the test FRC aligns with the training range ($[256, 512]$). Meanwhile, TOWER-based models exhibit higher sensitivity to FRC than QWEN2.5-based models. In summary, as long as the FRC is set to match the training configuration, variations in test data length have little impact on their performance.

H.2 Impact of Unit Selection on FRC

Beyond the baseline approach based on line breaks, we also adopt embedding-based semantic chunking and Meta-chunking (Zhao et al., 2025). For the former, we use the mE5-large model (Wang et al., 2024a) to compute embeddings for all sentences, measure the similarity between each pair of adjacent sentences, and retain only the top 20% of connection points with the highest similarity. For the latter, we directly employ the Meta-chunking method, which identifies appropriate breakpoints based on perplexity, using the original TOWER-7B-INSTRUCT-MISTRAL model for perplexity (PPL) computation. These generated units are subsequently used to construct fixed-range chunks. The experimental results are summarized in Table 17.

Compared to the BWB and GlobVDoc datasets, unit selection induces significantly greater perturbations on IWSLT en-xx. This effect is predominantly detrimental to both STAIR models, whereas SEP-2c1t exhibits a slight improvement in d-BLEU. To elucidate the underlying causes, we further calculated the ds-BLEU scores, n-gram repetition rate (NRR), and the proportion of problematic document lengths within the inference results for these three models on IWSLT2017 en-xx, as shown in Table 18.

A comparison of the three models reveals that

Method	Unit	IWSLT2017		BWB		GlobVDoc	
		d-B.	d-C.	d-B.	d-C.	d-B.	d-C.
STAIR-FS4	line-break	35.81	82.29	27.61	80.56	41.96	88.31
	emb-sim	-0.08	**-0.33**	-0.13	-0.05	-0.07	+0.01
	meta-ppl	**-0.65**	**-0.35**	-0.17	-0.13	**+0.31**	+0.01
STAIR-2c1t	line-break	38.29	84.14	27.53	81.10	42.06	89.35
	emb-sim	-0.09	**-0.85**	**+0.30**	+0.11	+0.07	-0.01
	meta-ppl	**+0.33**	**-0.59**	-0.12	+0.09	-0.08	+0.02
SEP-2c1t	line-break	38.78	84.80	26.82	81.03	41.98	89.26
	emb-sim	**+0.44**	-0.02	+0.14	-0.08	+0.05	+0.01
	meta-ppl	**+0.56**	+0.04	+0.01	-0.10	-0.10	+0.06

Table 17: The results obtained with different compositions of FRC units. The translation directions for the three datasets are IWSLT2017 en-xx, BWB zh-en, and GlobVDoc en-xx. Increases and decreases are calculated relative to the "line-break" within each respective model. Significant changes in performance are emphasized in **bold**. All the three models are TOWER-based.

Method	Unit	IWSLT2017		
		ds-B.	NRR↓	Prop.↓
STAIR-FS4	line-break	35.32	9.89%	28.25 %
	emb-sim	35.22	10.70%	28.62 %
	meta-ppl	34.41	14.17%	36.94 %
STAIR-2c1t	line-break	36.40	9.09%	16.37%
	emb-sim	36.33	10.43%	18.79%
	meta-ppl	36.67	8.02%	14.16%
SEP-2c1t	line-break	37.54	1.87%	4.56%
	emb-sim	37.53	1.60%	3.45%
	meta-ppl	37.57	1.87%	4.03%

Table 18: Further investigation and analysis of unit selecetion results for IWSLT en-xx. NRR denotes the n-gram repetition rate, and Prop. represents the proportion of problematic document lengths.

SEP-2c1t exhibits the most stable decoding performance, characterized by the lowest NRR. In addition, this stability remains largely invariant to the FRC unit selection strategy. In contrast, both STAIR models demonstrate clear volatility in decoding results when unit configurations are varied. According to the findings of Jin et al. (2024), n-gram repetition tends to emerge when the input sequence length exceeds 1k tokens. However, while the inputs for SEP-2c1t (including context) typically surpass this 1k threshold, they exhibit a relatively low NRR. In contrast, STAIR-2c1t shows obviously higher NRR levels. This discrepancy can likely be attributed to the training phase: (SEP)-2c1t was trained on a single length distribution, whereas STAIR-2c1t was exposed to three. In addition, STAIR-FS4 incorporated even greater distributional diversity, with input lengths substantially exceeding those of the other two models.

Comparing the results in Table 17 and Table 18

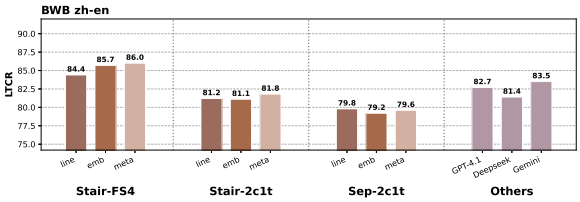


Figure 13: LTCR results of different unit selections across three models on BWB zh-en.

indicates that increases in NRR and the proportion of problematic document lengths correlate strongly with declines in d-BLEU and vise versa. While the NRR for SEP-2c1t remains relatively consistent across configurations, its d-BLEU improvements appear more salient because the other two strategies yield a smaller proportion of problematic document lengths compared to the line-break baseline. However, this trend is less pronounced in ds-BLEU, which utilizes averaged weights. The negligible variations in d-BLEU observed for the three models on the BWB and GlobVDoc datasets can be attributed to the fact that their NRR remains consistently at zero across all three unit selection settings.

In summary, utilizing various unit configurations for FRC fails to provide a stable boost in global accuracy. One explanation is that the structural significance of the units becomes generalized during the formation of fixed-range chunks. Additionally, d-BLEU and d-COMET may be insensitive to improvements in discourse coherence within each unit. Consequently, we conducted an LTCR analysis on the in-domain BWB dataset (NRR=0) to better differentiate the three methods. The results are presented in Figure 13.

Regarding STAIR-FS4 and STAIR-2c1t, the Meta-chunking approach yields marginal improvements (+1.6 and +0.6, respectively). Both unit strategies provide relatively notable gains for STAIR-FS4, even surpassing the performance of Gemini-2.5-Pro. These findings suggest that strategic unit partitioning can effectively enhance overall terminological consistency to a certain extent.

I Results of LoRA Fine-Tuning

Table 19 presents the results of the TOWER-based SEP-2c1t models under both full-parameter fine-tuning and LoRA fine-tuning. Although LoRA fine-tuning does not achieve the same performance as full-parameter fine-tuning, it still provides substantial improvements over the original model. In-

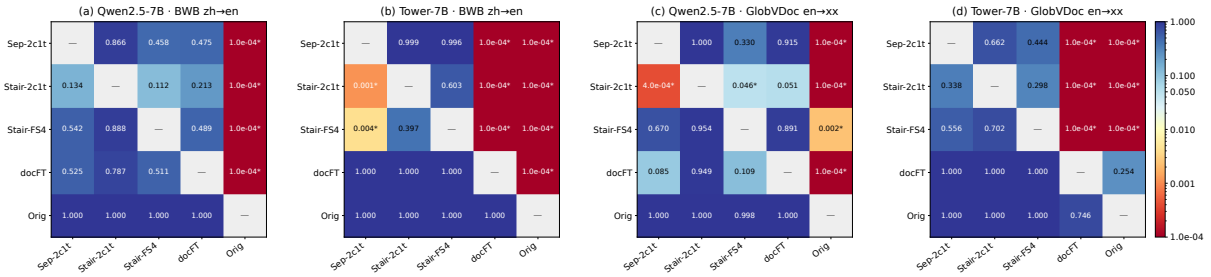


Figure 14: Pairwise one-sided significance tests on BWB zh-en and GlobVDoc en-xx for QWEN2.5-7B and TOWER-7B. Each cell reports the paired document-bootstrap p-value for the hypothesis that the row system outperforms the column system in d-BLEU. Asterisks indicate $p < 0.05$.

creasing the LoRA rank from $r = 16$ to $r = 64$ introduces a larger number of trainable parameters; however, the performance gains on the in-distribution test sets are relatively limited.

J AI Assistance Usage

We used AI-assisted tools, including ChatGPT[15], Google Gemini[16], and Claude[17], to support code development and writing refinement (in accordance with ACL Policy on AI Writing Assistance).

K Prompts

Prompt: D2D and 0c1t

<lim_start|>user
Translate the following source text from {source_lang} into {target_lang}.
{source_lang}: {source_text}.
{target_lang}: <lim_end|>
<lim_start|>assistant
{target_text}.<lim_end|>

Prompt: 1c1t

<lim_start|>user
Context:
{source_lang}: {source_context}
Translate the following source text from {source_lang} into {target_lang}.
{source_lang}: {source_text}.
{target_lang}: <lim_end|>
<lim_start|>assistant
{target_text}.<lim_end|>

Prompt: 2c1t

<lim_start|>user
Context1:
{source_lang}: {source_context1}
Context2:
{source_lang}: {source_context2}
Translate the following source text from {source_lang} into {target_lang}.
{source_lang}: {source_text}.
{target_lang}: <lim_end|>
<lim_start|>assistant
{target_text}.<lim_end|>

Prompt: FS4

<lim_start|>user
Context1:
{source_lang}: {source_context1}
Context2:
{source_lang}: {source_context2}
Context3:
{source_lang}: {source_context3}
Translate the following source text from {source_lang} into {target_lang}.
{source_lang}: {source_text}.
{target_lang}: <lim_end|>
<lim_start|>assistant
{target_text}.<lim_end|>

[15] https://chatgpt.com
[16] https://gemini.google.com
[17] https://claude.ai

Models	Decoding Format	IWSLT2017 en-xx		IWSLT2017 xx-en		BWB zh-en		GlobVDoc en-xx	
		d-BLEU	d-COM.	d-BLEU	d-COM.	d-BLEU	d-COM.	d-BLEU	d-COM.
Orig Model	0c1t	30.00	81.83	35.42	85.03	19.38	80.72	37.58	87.49
	1c1t	29.80	82.05	35.29	84.94	19.07	80.47	36.45	85.87
	2c1t	30.07	82.42	35.37	84.88	18.75	80.59	37.51	86.18
SEP-2c1t (Full)	0c1t	**39.61**	84.69	63.05	86.09	**27.38**	**81.11**	41.92	89.11
	1c1t	39.51	**84.83**	66.84	86.49	27.16	81.03	41.87	89.25
	2c1t	38.78	84.80	**67.20**	**86.53**	26.82	81.03	41.98	89.26
SEP-2c1t (LoRA $r = 16$)	0c1t	38.61	84.04	43.04	85.06	25.99	80.42	42.78	89.02
	1c1t	38.65	84.25	43.56	86.15	25.96	80.52	**43.28**	**89.52**
	2c1t	37.85	84.26	43.88	86.30	25.74	80.52	42.76	89.39
SEP-2c1t (LoRA $r = 64$)	0c1t	38.26	83.94	43.70	85.17	26.36	80.61	41.47	88.23
	1c1t	38.81	84.28	44.43	86.35	26.21	80.17	42.03	88.77
	2c1t	38.85	84.29	44.68	86.38	26.51	80.64	41.78	88.69

Table 19: Comparison among original, full-parameter fine-tuned SEP-2c1t and LoRA fine-tuned SEP-2c1t models. For each evaluation metric, the best performance is shown in **bold**.

Gemba-DA

Score the following translation from {src_lang} to {tgt_lang} with respect to human reference on a continuous scale 0 to 100 where score of zero means "no meaning preserved" and score of one hundred means "perfect meaning and grammar".

{src_lang} source: {src_text}
{tgt_lang} human reference: {ref_text}
{tgt_lang} machine translation: {tgt_text}

Without any thinking, output only JSON in the following format: {"score": <int>}

L Algorithm

Algorithm 1: Dual-Boundary Matching based Chunk Alignment

Input : Source chunks $\mathcal{C}$, Target units $\mathcal{T}$, λ, σ, Initial k, Δk

Output : Aligned lower bound indices $J = \{j_1, \ldots, j_{|\mathcal{C}|}\}$, $j \in \{1, .., |\mathcal{T}|\}$

Function $\texttt{SimwRP}(u, v)$**:**
  $s \leftarrow \text{Sim}(u, v)$;
  $w \leftarrow \exp\left(-\left(\frac{\text{rp}(u)-\text{rp}(v)}{\sigma}\right)^2\right)$;
  return $s \times ((1-\lambda) + \lambda w)$;

Initialize score matrix $\Phi \in \mathbb{R}^{|\mathcal{C}|\times|\mathcal{T}|}$;
while $J = \emptyset$ **do**
  for $i \leftarrow 1$ **to** $|\mathcal{C}|$ **do**
    for $j \leftarrow 1$ **to** $|\mathcal{T}|$ **do**
      $\Phi[i, j] \leftarrow \texttt{SimwRP}(c_i^{\text{end}}, t_j)$;
      Find top-k candidates $\{j_{i,1}, ..., j_{i,k}\}$;
      for $r \leftarrow 1$ **to** k **do**
        $\Phi[i, j] \leftarrow \Phi[i, j] + \texttt{SimwRP}(c_{i+1}^{\text{first}}, t_{j_{i,r}+1})$;
  Find $j_1, \ldots, j_{|\mathcal{C}|}$ maximizing $\sum \Phi[i, j_i]$;
    subject to $j_{i-1} \leq j_i$ (Monotonicity);
    and $j_{|\mathcal{C}|} = |\mathcal{T}|$ (Boundary Condition);
if $J = \emptyset$ **then** $k \leftarrow \min(k + \Delta k, |\mathcal{T}|)$;
return J ***or*** *Failure*

Algorithm 2: Two-Stage Dynamic Programming Algorithm for Fixed-Range Chunking

Input: Source units $S = \{s_1, s_2, \ldots, s_n\}$;
Target length $L^* = (m + M)/2$ derived from $[m, M]$;
Separator cost δ;
Max allowed violations K.
Output: Packed segments $\mathcal{P}$.

Calculate $|s_i|$ for all $s_i \in S$;
Initialize $\mathcal{P} \leftarrow \emptyset$;
`/* Pre-processing: Isolate super-long sentences */`
Divide S into blocks $\{B_1, B_2, \ldots\}$ by splitting at any s_i where $|s_i| > M$;
foreach *block* $B \in \{B_1, B_2, \ldots\}$ *with units* $u_1 \ldots u_T$ **do**
  if $|B| = 1$ ***and*** $|u_1| > M$ **then**
    Add $\{u_1\}$ to $\mathcal{P}$; **continue**;
  `/* Define length function with separator cost */`
  Let $Len(i, j) = \sum_{k=i}^{j} |u_k| + (j - i) \cdot \delta$;
  `/* Stage 1: Strict Constraint DP (Attempt to keep all segments` $\in [m, M]$`) */`
  Initialize $dp[0] \leftarrow 0$, $dp[1 \ldots T] \leftarrow \infty$;
  for $i \leftarrow 1$ **to** T **do**
    for $j \leftarrow i - 1$ **to** 0 **do**
      $l \leftarrow Len(j + 1, i)$;
      if $m \leq l \leq M$ **then**
        $dp[i] \leftarrow \min(dp[i], dp[j] + (l - L^*)^2)$;
  if $dp[T] < \infty$ **then**
    Backtrack dp to recover partition, add to $\mathcal{P}$;
  else
    `/* Stage 2: Relaxed DP with Violated Segment Constraints */`
    Let $dp_{rel}[v][i]$ be min cost for first i units with v violations;
    Initialize $dp_{rel}[0 \ldots K][0] \leftarrow 0$, others ∞;
    for $i \leftarrow 1$ **to** T **do**
      for $j \leftarrow i - 1$ **to** 0 **do**
        $l \leftarrow Len(j + 1, i)$;
        if $m \leq l \leq M$ **then**
          for $v \leftarrow 0$ **to** K **do**
            $dp_{rel}[v][i] \leftarrow \min(dp_{rel}[v][i], dp_{rel}[v][j] + (l - L^*)^2)$;
        else
          $C_{out} \leftarrow (m - l)^2$ if $l < m$ else $(l - M)^2$;
          for $v \leftarrow 1$ **to** K **do**
            $dp_{rel}[v][i] \leftarrow \min(dp_{rel}[v][i], dp_{rel}[v - 1][j] + C_{out})$;
    Find v^* minimizing $dp_{rel}[v][T]$ for $v \in [1, K]$;
    Backtrack $dp_{rel}[v^*]$ to recover partition, add to $\mathcal{P}$;
return $\mathcal{P}$ *ordered by original unit indices*;

ID	Source Title	Src	Translated Title	Tgt	Authors	Translators
0	The pros and cons of Chinese investment in Tajikistan's gold mining sector	En	发展与环保的两难：塔吉克斯坦金矿开采业的中国投资	Zh	Shahida Yakub	Global Climate Justice Fellowship
1	The complex role China plays in Africa's energy transition	En	中国在非洲能源转型中，扮演何种角色?	Zh	Ruohan Xie, Desire Nimubona	Sicong Zou
2	Boycotting Xinjiang cotton: What does it mean for environmental and labor justice in Central Asia?	En	抵制新疆棉，对中亚的环境和劳工权利影响几何?	Zh	Shahida Yakub, Global Climate Justice Fellowship	Global Climate Justice Fellowship
3	As electric vehicles gain momentum in Brazil, China's influence shines through	En	巴西电动汽车蓬勃发展背后的中国影响力	Zh	Laís Martins, Luo Jieqi	Sicong Zou
4	China increases gas imports from Turkmenistan for green energy transition. It's impact is unclear	En	中国增加土库曼斯坦天然气进口，绿色转型与甲烷泄露矛盾突显	Zh	Global Climate Justice Fellowship, Shahida Yakub	Global Climate Justice Fellowship
5	Will Ecuador lift Amazon oil block despite a historic referendum?	En	历史性公投后，厄瓜多尔会停止在亚马逊地区开采新油田吗?	Zh	Gabriela Mesones Rojo, Alicia Chen	Global Climate Justice Fellowship
6	With the reintroduction of import taxes on Chinese solar panels, Brazil hopes to develop its own industry	En	对中国光伏板重新征税，巴西希望发展本土产业	Zh	Laís Martins, Luo Jieqi	Luo Jieqi
7	Is China partly responsible for the destruction of Africa's Miombo woodlands?	En	非洲米欧波林地遭破坏，中国是否需要承担责任?	Zh	Ruohan Xie, Desire Nimubona	Ruohan
8	China helped Cameroon build drinking water infrastructure. Is it a debt crisis or developmental aid?	En	中国助喀麦隆饮建设用水基础设施：是援助发展，还是债务危机?	Zh	Desire Nimubona, Ruohan Xie	Sicong Zou
9	In Guadeloupe, 'Creole gardens' give climate lessons in a spirit of solidarity	En	Em Guadalupe, as "hortas crioulas" ensinam sobre o clima em um ambiente solidário	Pt	Olivia Losbar	Rogério de Sá
10	May Day or Detention Day? Turkey marks Labor Day	En	Primeiro de Maio ou Dia da Detenção? A Turquia celebra o Dia do Trabalho	Pt	Arzu Geybullayeva	Ronaldo Santos
11	In Russia, dozens of volunteers are working to save bat populations	En	Na Rússia, dezenas de voluntários trabalham para salvar populações de morcegos	Pt	Daria Dergacheva	Isabela Torezan
12	When the ICE agent at the airport echoes Trump's motto: A Brazilian journalist at the US border	En	Quando o agente da ICE no aeroporto ecoa o slogan de Trump: uma jornalista brasileira na fronteira dos EUA	Pt	Agencia Publica	Pública - Agência de jornalismo investigativo

Continued on next page

ID	Source Title	Src	Translated Title	Tgt	Authors	Translators
13	Saydnaya: In Syria, a legacy of pain looking for an honorable cleansing	En	Sednaya: na Síria, um legado de dor em busca de purificação honrosa	Pt	Rami Alhames	Denise Andréia Stange
14	'Let's talk about something else': China's AI chatbot DeepSeek censors sensitive topics	En	"Vamos mudar de assunto": chatbot de IA chinês DeepSeek censura temas sensíveis	Pt	Hong Kong Free Press	Joao Pedro Ribeiro
15	AI's bitter truth: It has biases, too	En	A amarga verdade da IA: ela não é neutra e pode falhar com seus vieses	Pt	Tactical Tech	Fernando Baumgarten
16	The Golem: A mythical protector inspiring US comics but also a metaphor for AI	En	O Golem: Um protetor mítico que inspira quadrinhos dos EUA, mas também uma metáfora para a IA	Pt	Fred Petrossian	Fernando Baumgarten
17	The Svrzo House: A window into the past of everyday life in Sarajevo	En	A Casa Svrzo: Uma janela ao passado da vida diária em Sarajevo	Pt	Balkan Diskurs	Gabriel Ferreira Wittaker
18	'Embracing imperfection is key to artistic evolution': An interview with Iranian artist Sadegh Adham	En	'Aceitar a imperfeição é a chave para a evolução artística': Entrevista com o artista iraniano Sadegh Adham	Pt	Omid Memarian	Isabela Torezan
19	The Conscience under attack: A story of aid, silence, and starvation	En	Das Gewissen unter Beschuss: Eine Geschichte von Hilfe, Schweigen und Hungersnot	De	Walid El Houri	GV Deutsch
20	A new aid regime for Gaza: Humanitarian facade, military core	En	Ein neues Hilfssystem für Gaza: Humanitäre Fassade, militärischer Kern	De	Saher	GV Deutsch
21	The murder of a young girl in Ethiopia reveals TikTok's content moderation failures	En	Die Ermordung eines jungen Mädchens in Äthiopien offenbart TikToks Versagen bei der Inhaltsmoderation	De	Endalkachew Chala	FTSK
22	How climate change is affecting farmers in Tobago	En	Wie der Klimawandel die Landwirt*innen in Tobago beeinflusst	De	Cari-Bois News	FTSK
23	'I don't feel safe': Reactions to Germany's suppression of pro-Palestine solidarity	En	„Ich fühle mich nicht sicher" – Reaktionen auf Deutschlands Unterdrückung der pro-Palästina Bewegung	De	Safa	Salma
24	Heatwave highlights climate vulnerabilities in Southeast Asia	En	Hitzewelle macht Klimaanfälligkeit Südostasiens deutlich	De	Sydney Allen	Mirjam Stiegel
25	43 years in Syria's prisons for refusing to bomb a city	En	43 Jahre in Syriens Gefängnissen, weil er sich weigerte, eine Stadt zu bombardieren	De	Rami Alhames	Anne Hemeda
26	Nepal's journey to electric public transport	En	Nepals Weg hin zum elektrischen Nahverkehr	De	Nepali Times	Julia K., FTSK
27	Blood and democracy: the fight for animal rights in Azerbaijan	En	Der Kampf für Tierrechte und Demokratie in Aserbaidschan	De	Arzu Geybullayeva, OC Media	Mirjam Stiegel
28	Why I might not go back to El Salvador	En	Warum ich vielleicht nie wieder nach El Salvador zurückkehre	De	Melissa Vida	Monica Seemann
29	How the Caribbean views the Trump administration's mass deportations	En	Caraïbes : le prisme des déportations massives de l'administration Trump	Fr	Janine Mendes-Franco, Candice Stewart	Corinne Molton
30	Chinese social media users call this age "The Garbage Time of History"	En	Chine : les utilisateurs des réseaux sociaux appellent cette ère « L'âge des ordures »	Fr	Oiwan Lam	Rodrigue Macao
31	World Environment Day: In Jamaica, the battle against plastics continues	En	Journée mondiale de l'environnement : en Jamaïque, la lutte contre les déchets plastiques se poursuit	Fr	Emma Lewis	Corinne Molton
32	Across war zones, targeting healthcare has become a strategy, not an accident	En	Les attaques délibérées contre les systèmes de santé deviennent une caractéristique de la guerre moderne - et un test pour le droit international.	Fr	Walid El Houri	Helene Senecot

Continued on next page

ID	Source Title	Src	Translated Title	Tgt	Authors	Translators
33	This Ghanaian software engineer is working to raise women's financial literacy	En	Un ingénieur logiciel ghanéen à pied d'œuvre pour améliorer l'éducation financière des femmes africaines	Fr	Zita Zage	Fierté Mbadzi
34	Monoliths of Nartiang: The remnants of India's Jaintia tribal kingdom through photos	En	Monolithes de Nartiang : les vestiges du royaume tribal indien de Jaintia en images	Fr	Arpita Das Choudhury	Corinne Molton
35	Power, myth, and the personal: A conversation with Iranian-American artist Shiva Ahmadi	En	Pouvoir, mythe et personnalité : entretien avec l'artiste irano-américaine Shiva Ahmadi	Fr	Omid Memarian	Pierre-Emmanuel Farret
36	Integrating Indigenous peoples and local communities (IPLCs) in Nepal's conservation efforts	En	Népal : intégrer les populations autochtones et les communautés locales (IPLC) dans les efforts de conservation du pays	Fr	Biswash Chepang	Pierre-Emmanuel Farret
37	Tobago's coral reefs brace for 'imminent threat'	En	Les récifs coralliens de Tobago se préparent à une « menace imminente »	Fr	Janine Mendes-Franco	Stanislav Kibalnyk
38	Redefining belonging: From statelessness to collective strength	En	Redéfinir l'appartenance : de l'apatridie à la force collective	Fr	Guest Contributor	Ridley Mortensen
39	China's warning against cross-border marriage scams reveals the pitfalls of human trafficking	En	Advertencia de China contra las estafas matrimoniales transfronterizas revela trampas de la trata de personas	Es	Rezwan, Oiwan Lam	Karen Lopez
40	El Salvador's soaring property costs are hurting locals	En	Elevados precios de viviendas en El Salvador perjudican a lugareños	Es	Eddie Galdamez	Valeria Malavolta
41	Podcast: Daria on carrying a Russian identity and setting her children free of it	En	Podcast: Daria sobre la carga de identidad rusa y su decisión de criar hijas sin esa carga	Es	Akwe Amosu, Daria Dergacheva	Georgina Loscalzo
42	No lambs for Eid: Drought, deforestation, and decline in Morocco	En	No hay corderos para el Eid: Sequía, deforestación y deterioro en Marruecos	Es	Mohamed Belkasen	Eva González
43	Painting as a shelter: Spanish artist Bárbara Alegre on art, grief, and emotional healing	En	La pintura como refugio: Artista española Bárbara Alegre sobre el arte, el duelo y la sanación emocional	Es	Omid Memarian	Mariela Arnst
44	Kenya outlawed the sharing of Indigenous seeds. A village choir is fighting back	En	Kenia prohibió intercambiar semillas autóctonas. Un coro comunitario contraataca	Es	Minority Africa	Mariela Arnst
45	Ahead of municipal elections, Lebanese voter data is at risk	En	Antes de elecciones municipales, datos de los votantes libaneses corren peligro	Es	SMEX	Gabriela García Calderón Orbe
46	The Congo Basin, a vital home for global biodiversity is at risk	En	Cuenca del Congo, vital para la biodiversidad global, está en peligro	Es	Guest Contributor	Mariela Arnst
47	How the California gold rush continues to shape Africa and other global majority countries	En	Cómo la fiebre del oro en California sigue dando forma a África y a otros países de la mayoría global	Es	Abdallah Abdallah	Mariela Arnst
48	Gains and gaps in gender equality in North Macedonia	En	Avances y brechas de igualdad de género en Macedonia del Norte	Es	Kristina Hadzi-Vasileva	Mariela Arnst
49	In Turkey, a controversial law on cybersecurity is widely seen as yet another censorship tool	En	In Turkije wordt een omstreden wet over cyberbeveiliging gezien als een nieuw middel tot censuur	Nl	Arzu Geybullayeva	Mijke Luttikholt
50	Bringing 'Pateh' to the world: Sara Qashghai's artistic reinterpretation of Iranian needlework	En	'Pateh' naar de wereld toe brengen: Sara Qashghai's artistieke herinterpretatie van Iraans handwerk	Nl	Omid Memarian	Els Whittle
51	Indigenous People defend traditional farming in northern Thailand	En	In Noord-Thailand zet de de bevolking zich in voor traditionele landbouwmethoden	Nl	Prachatai	Eric Crena Uiterwijk
52	From Myanmar to Thailand: Displaced journalists tell their stories	En	Van Myanmar tot Thailand: Ontheemde journalisten vertellen hun verhaal	Nl	Prachatai	Eric Crena Uiterwijk

Continued on next page

ID	Source Title	Src	Translated Title	Tgt	Authors	Translators
53	Iran sees 80% spike in executions two years after protests	En	Iran: 80% meer executies twee jaar na protesten	Nl	Iran Open Data Center	Thomas De Backer
54	Women are paying the ultimate price in Cameroon's armed conflict	En	Vrouwen betalen de hoogste prijs in het gewapende conflict van Kameroen	Nl	Minority Africa	Jasmijn Dagevos
55	Pouring concrete on rice fields in Nepal	En	In Nepal veranderen de rijstvelden in een betonnen woestenij	Nl	Nepali Times	Eric Crena Uiterwijk
56	Where are the unusual swarms of bees in Chișinău, Moldova, coming from?	En	Waar komen de ongebruikelijke bijenzwermen uit Chișinău in Moldavïe vandaan?	Nl	Daria Dergacheva	Eric Crena Uiterwijk
57	Cat lovers boost tourism in Taiwan village as feline residents revive once-flourishing mining town	En	Kattenliefhebbers stimuleren toerisme in een eens bloeiend, door kattenbewoners nieuw leven ingeblazen, Taiwanees mijnstadje	Nl	Hong Kong Free Press	Els Whittle
58	When it comes to bullying, small actions can help stop big problems	En	Kleine daden kunnen groot verschil maken tegen pesten	Nl	Guest Contributor	Mijke Luttikholt
59	Koryo-saram: The long and tragic story of Koreans in Russia	En	고려사람 : 러시아 한인의 길고 비극적인 역사	Ko	Daria Dergacheva	June Lee
60	Vaping loopholes are endangering children and the environment in Nigeria and Burkina Faso	En	나이지리아 및 부르키나파소의 어린이 그리고 환경, 전자담배의 허점으로 위험에 처해	Ko	The Colonist Report	June Lee
61	China strives to go green in South America's 'Lithium Triangle'	En	중국, 남미의 '리튬 삼각지대'에서 친환경화를 위해 분투하다	Ko	Alicia Chen, Gabriela Mesones Rojo	June Lee
62	Hong Kong secondary students may soon be schooled in 'Xi Jinping Thought'	En	홍콩 중학생들, 곧 '시진핑 사상' 교육 받을 수도	Ko	Hong Kong Free Press	June Lee
63	Coup and resistance in Myanmar: A timeline of the first month under the 2021 military junta	En	미얀마에서의 쿠데타와 저항 : 2021년 군부 독재 하의 첫 달 타임라인	Ko	Global Voices South East Asia	Seojin Lim
64	Masculinity in my genes/jeans	En	청바지(진)에 들어있는 사내다움	Ko	Amilcar Sanatan	May Cho
65	The Caribbean's case for reparations: Part I	En	카리브해 지역 배상 문제: 제 1부	Ko	Janine Mendes-Franco	May Cho
66	The Caribbean's case for reparations: Part II	En	카리브 해 지역의 배상 문제: 제 2 부	Ko	Janine Mendes-Franco	May Cho
67	The Caribbean's case for reparations: Part III	En	카리브 해 지역의 배상 문제: 제 3부	Ko	Janine Mendes-Franco	May Cho
68	Women 'don't have to fit themselves into someone else's perception,' says Turkish aerospace engineer	En	터키의 항공 우주 기술자: 여성은 타인의 인식에 맞출 필요가 없어요	Ko	Sevgi Yagmur Bulut	Keun Jeong
69	From barren land to thriving forest: The story of photographer Sebastião Salgado's Instituto Terra in Brazil	En	Dalla terra arida alla foresta florida: la storia dell'Instituto Terra del fotografo Sebastião Salgado in Brasile	It	Ana Cavalcanti	Margherita Lista
70	Global digital rights report reveals unexpected boost in transparency from Chinese tech giants	En	Rapporto globale sui diritti digitali rivela un inaspettato aumento della trasparenza da parte dei giganti tecnologici cinesi	It	Marisa Petricca	Marisa Petricca
71	Two years on, Turkey earthquake survivors continue to live in limbo	En	Due anni dopo, i sopravvissuti al terremoto in Turchia continuano a vivere nell'incertezza	It	Arzu Geybullayeva	Martina Cesarano
72	Argentine resistance hinders Milei's forest and glacier destruction	En	La resistenza argentina ostacola la distruzione di foreste e ghiacciai operata da Milei	It	Climate Home News	Vivi
73	A ballerina battles for European Georgia	En	Una ballerina che lotta per una Georgia europea	It	Giorgi Ninos	Martina Cesarano

Continued on next page

ID	Source Title	Src	Translated Title	Tgt	Authors	Translators
74	Women's rights under threat in Uganda as conservative groups push disinformation campaign	En	Diritti delle donne sotto attacco in Uganda mentre gruppi conservatori promuovono una campagna di disinformazione	It	Prudence Nyamishan	Laura Carlevero
75	How China's investment in Indonesia's nickel industry is impacting local communities	En	Gli investimenti cinesi nell'industria del nichel hanno un'impatto sulle comunità locali in Indonesia	It	Hasya Nindita, Zhaoyin Feng	Barbara Foggiato
76	Pacific communities seek to protect kava as it gains global popularity	En	Le comunità del Pacifico cercano di proteggere la kava, una bevanda la cui popolarità continua a crescere	It	Mong Palatino	Barbara Foggiato
77	On International Women's Day, a call for Jamaica to combat systemic barriers that hold women back	En	Nella Giornata internazionale della donna, la Giamaica deve combattere le barriere sistemiche che frenano le donne	It	Janine Mendes-Franco	Martina Cesarano
78	Money from trees: What of Guyana's Indigenous people and their rights — and do they benefit from the carbon trade?	En	Denaro dagli alberi: cosa rimane agli indigeni della Guyana e i loro diritti? Beneficiano del commercio del carbonio?	It	Guest Contributor	Barbara Foggiato
79	For how long? Aramaic language and its enduring legacy in Syria	En	Арамейский язык в Сирии: бесценное наследие, которое мы можем потерять?	Ru	Rami Alhames	Елена Кузминцева
80	One man is trying to save a language in Bangladesh with only six native speakers	En	Одиночка, пытающийся спасти язык, на котором говорят всего шесть человек	Ru	Pantha	Rezwan, Laura Sabit
81	Peru adopts controversial 'anti-NGO' law	En	Перу принимает спорный закон против НПО	Ru	IFEX, Laura Vidal	Dalia Tarek, Anastasia Pestova
82	A decade of digital rights: Where has Big Tech's progress gone?	En	Десятилетие цифровых прав: куда исчез прогресс Big Tech?	Ru	Giovana Fleck	Elise Patterson
83	A new aid regime for Gaza: Humanitarian facade, military core	En	Новый режим помощи Газе: гуманитарный фасад, но военная основа	Ru	Saher	Marina Bovkalo
84	Chinese social media users are outraged by the mysterious death of an internet celebrity cat	En	Гибель звезды соцсетей: в Китае гадают, что произошло с котом Укуном	Ru	Oiwan Lam	Elise Patterson
85	In the Llanos Orientales, seduction is another weapon of war	En	Колумбия: соблазнение как ещё одно оружие войны	Ru	Mi Historia	Hebe Powell, Anastasia Pestova
86	The environmental impact of Chinese cement plants in Tajikistan remains hidden	En	Что скрывают китайские цементные заводы в Таджикистане?	Ru	Nurbek Bekmurzaev, Brian Hioe	Elise Patterson
87	Nelly Gesare turns trash into treasure, highlighting the potential of sustainability-driven businesses in Africa	En	Как в Африке бывшая журналистка Нелли Джесаре создала процветающий бизнес из отходов	Ru	Bird	Marina Bovkalo
88	Chechen opposition activist fears deportation to Russia from Kazakhstan	En	Чеченский оппозиционер опасается депортации в Россию из Казахстана	Ru	Guest Contributor, Daria Dergacheva	Anastasia Pestova
89	Rebels and Rebellion in Classic Chinese Literature	En	中国古典文学里的反骨精神与叛逆分子	Zh	I-fan Lin	Yanne C

Table 20: The metadata for the documents within the GlobVDoc dataset includes the document ID, source and translated titles, source and target languages, and the names of authors and translators.