SubDocTrans: Enhancing Document-level Machine Translation with Plug-and-play Multi-granularity Knowledge Augmentation

Hanghai Hong, Yibo Xie, Jiawei Zheng, Xiaoli Wang[†]
School of Informatics, Xiamen University
Research Center of Star-XLAB, Xiamen University
hanghaih@stu.xmu.edu.cn, xlwang@xmu.edu.cn

Abstract

Large language models (LLMs) have recently achieved remarkable progress in sentence-level machine translation, but scaling to document-level machine translation (DocMT) remains challenging, particularly in modeling contexts and discourse phenomena across sentences and paragraphs. Document translations generated by LLMs often suffer from poor consistency, weak coherence, and omission errors. To address the issues, we propose a novel DocMT framework named SubDocTrans, that enables LLMs to produce high-quality translations via plug-and-play multi-granularity knowledge augmentation. SubDocTrans first performs topic segmentation to divide a document into coherent sub-documents, enabling efficient document-level translation while preserving contextual coherence. For each sub-document, both global and local knowledge are extracted including bilingual summary, theme, proper nouns, topics, and transition hint. We then incorporate the multi-granularity knowledge into the prompting strategy, to guide LLMs in producing consistent, coherent, and accurate translations. We also conduct extensive experiments across various DocMT tasks, and the results demonstrate the effectiveness of our framework, particularly in improving consistency and coherence, reducing omission errors, and mitigating hallucinations.

1 Introduction

LLMs have demonstrated remarkable performance in natural language processing tasks (Adams et al., 2023; Garcia and Firat, 2022; Hendy et al., 2023; Zhang et al., 2023a; Jiao et al., 2023). Researchers have explored the application of LLMs to DocMT, to address discourse issues, such as pronoun translation and terminological inconsistencies. Liu et al. (2025) demonstrated that large reasoning models (LRMs) (OpenAI et al., 2024; Zhang et al., 2023b; DeepSeek-AI et al., 2025) with chain-of-thought (CoT) capabilities (Wei et al., 2022) can redefine translation as a dynamic reasoning task for improving contextual coherence, cultural adaptation, and robustness through explicit reasoning.

Recent studies primarily focus on developing DocMT agents (Guo et al., 2025; Wang et al., 2025; Cui et al., 2024; Wu et al., 2024). They can be categorized into two groups: Doc2Doc and Doc2Sent. The first group adopts Doc2Doc decoding by leveraging multi-agent collaboration for translating extremely long texts (Wu and Hu, 2023). However, they may result in sentence omissions (Karpinska and Iyyer, 2023) and hallucinations (Dale et al., 2023). To mitigate sentence omissions, the second group adopts Doc2Sent decoding by decomposing documents into individual sentences (Wang et al., 2025). However, they often impair discourse phenomena, such as document coherence and pronoun translation accuracy. Multi-agent interactions can also substantially increase the computational cost.

To address these challenges, we propose a novel DocMT framework denoted as SubDocTrans, to help LLMs produce high-quality translations via plug-and-play multi-granularity knowledge augmentation. Different from Doc2Doc and Doc2Sent, we propose a new translation strategy to segment documents into sub-documents. SubDocTrans arguments translation with global and local knowledge extracted from sub-documents: bilingual summary, theme, proper nouns, topics, and transition hint. Incorporating such multi-granularity knowledge into the prompting strategy can guide LLMs in producing consistent, coherent, and accurate translations. SubDocTrans further adopts a sentence alignment strategy to avoid omission errors. Our main contributions are summarized as follows:

- We propose a novel DocMT framework denoted as SubDocTrans, which enables LLMs to gen-

† Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 14490–14506
November 4-9, 2025

erate high-quality translations via plug-and-play multi-granularity knowledge augmentation.

- We introduce a sentence alignment strategy to address key limitations of existing Doc2Doc approaches by avoiding omission errors and mitigating hallucinations.
- We conduct extensive experiments and the results demonstrate the effectiveness of our framework, particularly in improving consistency and coherence, reducing omission errors, and mitigating hallucinations.
- Compared to advanced baselines, SubDocTrans shows competitive efficiency by leveraging sub-document translation. It makes sense to take slight cost for achieving high-quality translations by incorporating multi-granularity knowledge.

2 Related Work

LLMs for DocMT LLMs are leveraged to address discourse phenomena and coherence in DocMT, and recent studies primarily focus on developing DocMT agents (Guo et al., 2025; Wang et al., 2025; Cui et al., 2024; Wu et al., 2024). They can be categorized into two groups: Doc2Doc and Doc2Sent. TransAgent (Wu and Hu, 2023) adopts Doc2Doc decoding by leveraging multi-agent collaboration to tackle the translation of extremely long texts. However, it may result in sentence omissions (Karpinska and Iyyer, 2023) and hallucinations (Dale et al., 2022; Guerreiro et al., 2023; Dale et al., 2023). To mitigate sentence omissions, IncreD (Lyu et al., 2021) and DELTA (Wang et al., 2025) focus on Doc2Sent decoding by decomposing documents into individual sentences. However, they often impair discourse phenomena, such as document coherence and the accuracy of pronoun translation. Moreover, multiple agent interactions can substantially increase the computational cost.

Knowledge Augmentation for MT Recent work leverages LLMs with external knowledge to improve translation quality, consistency, and informativeness (Merx et al., 2024; Conia et al., 2024). Wang et al. (2023a) introduce a prompting-based framework that prepends multiple knowledge types including examples, templates, and terminologies, without modifying the NMT model. Qian et al. (2023) utilize GPT-4 with style-aware instructions and human-written references to capture nuanced authorial intent. Li et al. (2025) propose KAT, which retrieves structured knowledge from Wikidata and incorporates it via entity-aware prompting to enhance low-resource translation. Existing work mainly focused on single-source or retrieval-based knowledge. Differently, our SubDocTrans leverages LLMs as agents to extract and integrate multi-granularity knowledge into a unified prompting framework, which guides LLMs to produce coherent, consistent, and accurate outputs by effectively mitigating omission errors and hallucinations.

3 Method

We propose SubDocTrans, a novel DocMT framework that leverages the CoT capabilities of LRMs. By segmenting long documents into sub-documents, SubDocTrans ensures inter-paragraph cohesion while enhancing translation through five knowledge augmentations: bilingual summary for modeling global context, theme for domain guidance, proper nouns for terminology consistency, topics for capturing local semantics, and transition hint for preserving seamless paragraph transitions. The framework of our SubDocTrans is illustrated in Figure 1, the algorithm of SubDocTrans is detailed in Algorithm Appendix A.1, and the prompts used for each module are given in Appendix A.2.

3.1 Document Segmentation

SubDocTrans first performs topic segmentation named by $\mathcal{L}_{\text{Segment}}$, to divide a source document into coherent topic sub-documents. Given a source document of $\boldsymbol{D}_s = \{s_1, s_2, \ldots, s_N\}$, where s_i is a sentence, $\mathcal{L}_{\text{Segment}}$ divides $\boldsymbol{D}_s$ into overlapping sub-documents $\{\boldsymbol{C}_1, \ldots, \boldsymbol{C}_n\}$, each with c sentences and an overlap of o sentences to ensure smooth transitions.

We first partition $\boldsymbol{D}_s$ into M segments (e.g., paragraphs), represented as $\boldsymbol{D}_s = \{p_1, \ldots, p_M\}$. Each partition is converted to a vector using TF-IDF weighted by Word2Vec embeddings to capture semantic similarity:

$$\text{SIM}(p_i, p_j) \propto \sum_{w_m \in p_i} \sum_{w_n \in p_j} \text{tfidf}(w_m, p_i) \cdot \text{SIM}(w_m, w_n) \cdot \text{tfidf}(w_n, p_j), \tag{1}$$

where $\text{tfidf}(w_m, p_i)$ is the TF-IDF score of word w_m in partition p_i, and $\text{SIM}(w_m, w_n)$ is the cosine similarity of Word2Vec embeddings of words w_m and w_n, capturing semantic closeness.

Sub-documents are formed by grouping partitions to maximize coherence using dynamic pro-

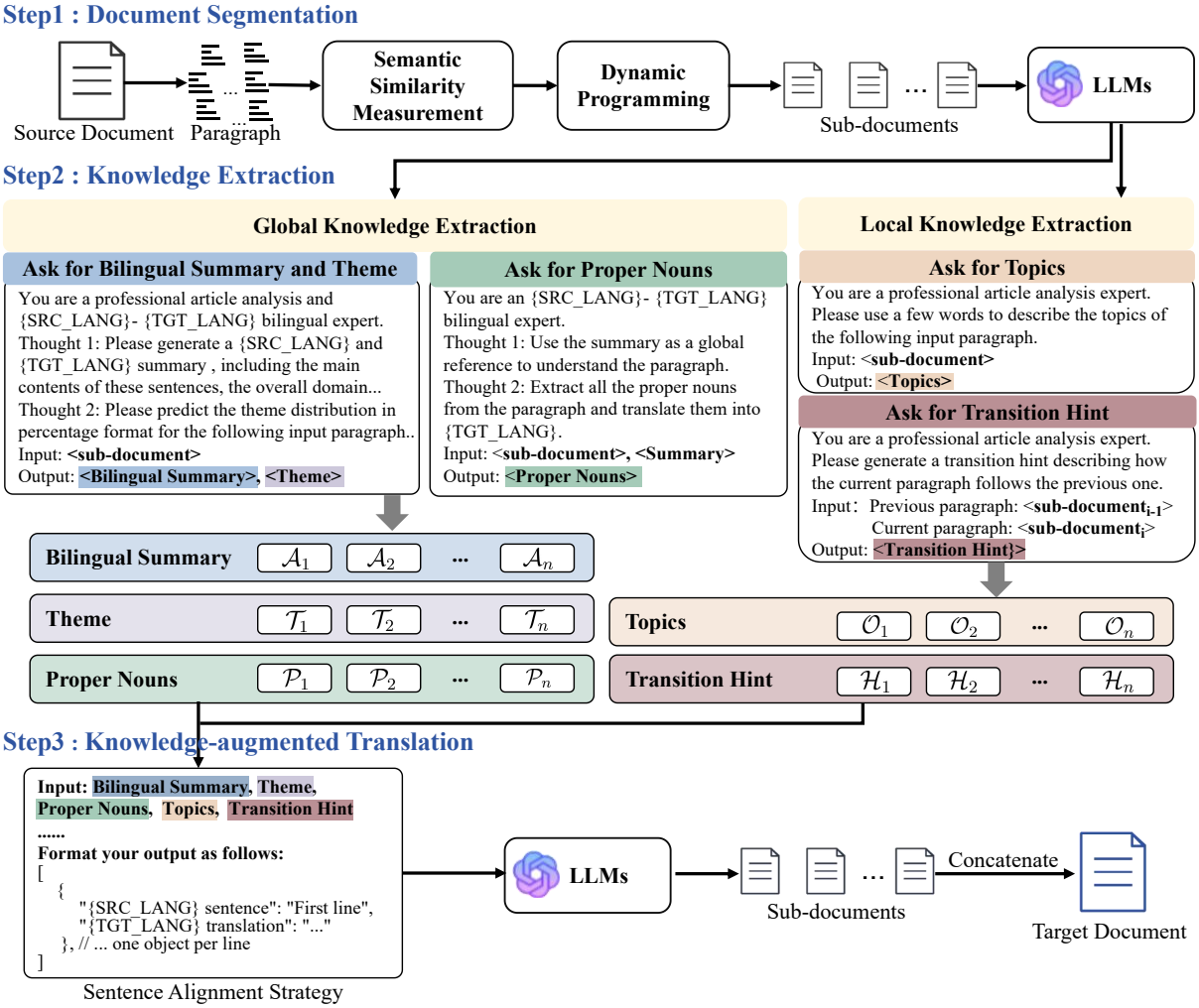

Figure 1: Framework of SubDocTrans. **Step 1** divides the source document into coherent and overlapping sub-documents to maintain contextual coherence. **Step 2** extracts global and local knowledge from sub-documents, including bilingual summary, theme, proper nouns, topics, and transition hint. **Step 3** leverages the multi-granularity knowledge into the prompt construction and adopts a sentence alignment strategy to produce high-quality translations of target document.

gramming. The quality of a sub-document $\boldsymbol{C}_i$ is:

$$q(\boldsymbol{C}_i) = \sum_{p_j \in \boldsymbol{C}_i} \text{SIM}(p_j, \boldsymbol{Z}_i), \tag{2}$$

where $\boldsymbol{C}_i = \{p_l, \ldots, p_m\}$ denotes the sub-document containing a contiguous sequence of partitions from the l-th to the m-th, and $\boldsymbol{Z}_i$ is the centroid vector of $\boldsymbol{C}_i$, averaging its partition vectors. The optimal segmentation maximizes $\sum_{i=1}^{n} q(\boldsymbol{C}_i)$, is solved by

$$f^*(b,k) = \max_{c \le l \le b} \{f^*(l-1, k-1) + q(l,b)\}, \tag{3}$$

where $f^*(b,k)$ is the maximum quality of grouping the first b partitions into k sub-documents, and $q(l,b)$ is the quality of a sub-document from partitions p_l to p_b.

The resulting sub-documents are generated as:

$$\{\boldsymbol{C}_1, \ldots, \boldsymbol{C}_n\} \leftarrow \mathcal{L}_{\text{Segment}}(\boldsymbol{D}_s, o), \tag{4}$$

where o is the overlap size.

On average, each sub-document contains 20 sentences, as determined dynamically by our segmentation strategy.

3.2 Knowledge Extraction

3.2.1 Global Knowledge Extraction

Global Knowledge Augmentation provides document-level context to guide translation, ensuring consistency and coherence across sub-documents. It includes bilingual summary, theme, and proper nouns, which are shared across all sub-documents to maintain global information consistency throughout the translation process.

Bilingual Summary and Proper Nouns Bilingual Summary ($\mathcal{A}_s, \mathcal{A}_t$) and Proper Nouns ($\mathcal{P}$) are critical to capture global context and maintain terminology consistency, respectively (Wang et al., 2025).

For each sub-document $\boldsymbol{C}_i$, a bilingual summary is first generated by the LLM-based Summary Generator $\mathcal{L}_{\text{Summary}}$:

$$\tilde{\mathcal{A}}_{s,i}, \tilde{\mathcal{A}}_{t,i} \leftarrow \mathcal{L}_{\text{Summary}}(\boldsymbol{C}_i) \quad (5)$$

The summaries are then aggregated:

$$\mathcal{A}_s \leftarrow \mathcal{A}_s \cup \{\tilde{\mathcal{A}}_{s,i}\}, \quad \mathcal{A}_t \leftarrow \mathcal{A}_t \cup \{\tilde{\mathcal{A}}_{t,i}\} \quad (6)$$

To form cohesive global summaries, they are merged by the LLM-based Summary Merger:

$$\mathcal{A}_s \leftarrow \mathcal{L}_{\text{MergeS}}(\mathcal{A}_s), \quad \mathcal{A}_t \leftarrow \mathcal{L}_{\text{MergeT}}(\mathcal{A}_t) \quad (7)$$

Proper nouns are generated at the sub-document level, using the source summary $\mathcal{A}_s$ to enhance contextual accuracy. For each sub-document $\boldsymbol{C}_i$, they are extracted by an LLM-based Proper Nouns Extractor $\mathcal{L}_{\text{ProperNouns}}$:

$$\mathcal{P}_i \leftarrow \mathcal{L}_{\text{ProperNouns}}(\boldsymbol{C}_i, \mathcal{A}_s) \quad (8)$$

Proper nouns from each sub-document $\mathcal{P}i$ are first aggregated and then merged by the LLM-based Proper Noun Merger as follows:

$$\mathcal{P} \leftarrow \mathcal{P}_n \cup \{\mathcal{P}_i\}, \mathcal{P} \leftarrow \mathcal{L}_{\text{MergeProperNouns}}(\mathcal{P}) \quad (9)$$

During merging, we adopt the first translation of each proper noun, which is generated with access to the most complete global context, as the canonical form to ensure consistency across all sub-documents.

Theme Unlike summary, theme directly provides the paragraph's theme using a few words, such as *Social issues 50%, Economy 30%, International relations 20%*.

For each sub-document $\boldsymbol{C}i$, a theme is generated by the LLM-based Theme Predictor $\mathcal{L}_{\text{Theme}}$:

$$\mathcal{T}_i \leftarrow \mathcal{L}_{\text{Theme}}(\boldsymbol{C}_i) \quad (10)$$

Themes from each sub-document $\mathcal{T}i$ are aggregated and then merged by the LLM-based Theme Merger as follows:

$$\mathcal{T} \leftarrow \mathcal{T} \cup \{\mathcal{T}_i\}, \mathcal{T} \leftarrow \mathcal{L}_{\text{MergeTheme}}(\mathcal{T}) \quad (11)$$

This global theme informs the translation process, helping ensure domain-appropriate lexical choices and stylistic consistency—particularly in literary and technical documents.

3.2.2 Local Knowledge Extraction

Local knowledge augmentations focus on sub-document-specific semantics and cohesion, enhancing translation quality within and between sub-documents. These include topics and transition hint.

Topics Topics ($\mathcal{O}$) capture local semantics, ensuring translations reflect the core ideas of each sub-document. Unlike global knowledge, topics are not merged, as they are specific to each sub-document.

For each sub-document $\boldsymbol{C}_i$, topics are generated by an LLM-based component known as the Topic Extractor $\mathcal{L}_{\text{Topics}}$:

$$\mathcal{O}_i \leftarrow \mathcal{L}_{\text{Topics}}(\boldsymbol{C}_i) \quad (12)$$

Transition Hint Transition Hint ($\mathcal{H}$) ensures inter-paragraph cohesion by guiding translations to maintain discourse flow between sub-documents.

For each sub-document $\boldsymbol{C}_i$, a transition hint is generated by an LLM-based component known as the Hint Generator $\mathcal{L}_{\text{Hint}}$:

$$\mathcal{H}_i \leftarrow \mathcal{L}_{\text{Hint}}(\boldsymbol{C}_{i-1}, \boldsymbol{C}_i) \quad (13)$$

These hints are used during the translation of all sub-documents except the first ($i > 1$) to ensure smooth transitions across segments.

3.3 Knowledge-Augmented Translation

The final translation is performed by an LLM-based component, Document Translator, denoted as $\mathcal{L}_{\text{Translate}}$. For each sub-document $\boldsymbol{C}_i$, the translation is generated as:

$$\boldsymbol{T}_i \leftarrow \mathcal{L}_{\text{Translate}}(\boldsymbol{C}_i, \mathcal{P}, \mathcal{A}_s, \mathcal{A}_t, \mathcal{T}, \mathcal{O}_i, \mathcal{H}_i) \quad (14)$$

All knowledge augmentations are integrated into the prompt to support the translator in producing high-quality and consistent translations.

SubDocTrans adopts a **Sentence Alignment Strategy** to guide the LLM in generating one-to-one aligned translations for each source sentence, reducing sentence omissions and ensuring output validity. To enhance robustness, the translation allows up to r_{max} retries to ensure valid output and full sentence coverage.

Post-translation, overlapping content is removed, producing target sub-documents:

$$\boldsymbol{T}_i \leftarrow \text{RemoveOverlap}(\boldsymbol{T}_i, o) \quad (15)$$

These target sub-documents are concatenated to form the Target Document $\boldsymbol{D}_t$.

4 Experiments

We conducted extensive experiments to answer the following research questions:

- **RQ1: Translation Quality and Consistency.** Does SubDocTrans lead to improvements in overall translation quality and consistency compared to advanced baselines?
- **RQ2: Effectiveness of Knowledge Augmentation.** How do different types of knowledge contribute to DocMT task?
- **RQ3: Handling of Discourse Phenomena.** Does SubDocTrans give improvement in the handling of discourse phenomena?
- **RQ4: Omission Errors and Hallucination Mitigation.** Can SubDocTrans reduce sentence omissions and mitigate hallucinations more effectively than Doc2Doc systems?
- **RQ5: Inference Efficiency.** Does SubDocTrans achieve better inference efficiency compared to existing DocMT systems?

4.1 Settings

Datasets The tst2017 is designed for the IWSLT2017 translation task (Akiba et al., 2004), comprising parallel documents from TED talks. we focus on eight bidirectional pairs: En ⇔ Zh, De, Fr, and Ja. Each language pair includes 10 to 12 sentence-aligned parallel documents, with approximately 1,500 sentences per language pair. The Guofeng Web-novel (Wang et al., 2023c, 2024) is a high-quality, discourse-level web novel corpus. We conduct our experiments on the Guofeng V1 TEST_2 set in the Zh ⇒ En direction.

Backbone Models We employ two GPT models of `GPT-3.5-Turbo-0125` and `GPT-4o-mini` as our backbone models, which are obtained from the official OpenAI API . We set the temperature to 0.8 and the top-p sampling threshold to 0.8, while keeping other hyperparameters at their default values. We also employ two open-source LRMs of `DeepSeek-R1:14B` and `DeepSeek-R1:32B` as our backbone models. To accommodate document-level translation outputs, we set `num_predict` to -1, allowing unrestricted generation lengths. The maximum retry count $r_{\max}$ is set to 20, and the o-sentence overlap is set to 3.

Baselines We compare SubDocTrans against four baselines, all using the same LLMs and decoding settings. **Sentence** treats each sentence independently without any document-level context. **Context** (Wu et al., 2024) adds three preceding source-target sentence pairs to the prompt to improved local coherence. **Doc2Doc** (Wang et al., 2023b) prompts the LLM to translate 10 sentences in a single conversation turn. prompts the LLM to translate ten sentences at once without intermediate alignment or guidance. **DeLTA** (Wang et al., 2025) performs sentence-by-sentence translation using multi-level memory prompts to guide consistency and fluency. We also include results from `NLLB-3.3B` (Costa-jussà et al., 2022) and `GoogleTrans`, for further evaluation.

Metrics We adopt both quality and consistency metrics to evaluate the performance. For quality, we utilize two neural-based COMET metrics. One is the sentence-level COMET score (sCOMET), for which we use the model `Unbabel/wmt22-comet-da`. Another is the document-level COMET score (dCOMET), proposed by Vernikos et al. (2022), using `wmt21-comet-qe-mqm` to compute reference-free scores. To assess terminological translation consistency, we use the consistency metrics of LTCR-1 and LTCR-1_{f} (Wang et al., 2025).

4.2 Main Results

We report the main results of SubDocTrans on the IWSLT2017 and Guofeng datasets, comparing its performance against strong baselines across multiple translation directions.

Results for RQ1: Translation Quality and Consistency. The main experiment results on the IWSLT2017 test set are demonstrated in Table 1. For more detailed scores, please refer to Appendix A.3. SubDocTrans achieves the highest scores

https://wit3.fbk.eu/2017-01-d/

https://github.com/longyuewangdcu/GuoFeng-Webnovel/

https://platform.openai.com/docs/guides/text-generation/

https://ollama.com/library/deepseek-r1:14b/

https://ollama.com/library/deepseek-r1:32b/

https://py-googletrans.readthedocs.io/

https://github.com/Unbabel/COMET/

https://github.com/amazon-science/doc-mt-metrics/

https://unbabel-experimental-models.s3.amazonaws.com/comet/wmt21/wmt21-comet-qe-mqm.tar.gz

System	En ⇒ Xx				Xx ⇒ En			
	sCOMET	dCOMET	LTCR-1	LTCR-1_f	sCOMET	dCOMET	LTCR-1	LTCR-1_f
NLLB	82.11	6.36	74.56	81.87	84.10	6.98	79.03	90.76
GOOGLE	80.41	5.83	81.38	84.72	80.17	5.96	81.43	90.81
`GPT-3.5-Turbo`								
LLM + Sentence	84.80	6.58	77.06	82.81	84.47	7.05	81.98	91.86
LLM + Context	85.40	6.70	77.34	83.12	84.97	7.15	85.03	95.27
LLM + Doc2Doc	–	6.62	79.12	86.39	–	6.96	85.17	92.98
DELTA	**85.58**	**6.73**	82.96	88.83	84.95	7.15	86.53	96.26
SubDocTrans	85.46	6.72	**83.03**	**88.84**	**85.11**	**7.19**	**86.88**	**96.39**
`GPT-4o-mini`								
LLM + Sentence	81.51	6.35	78.59	85.07	84.01	6.99	81.42	91.34
LLM + Context	84.78	6.65	80.01	86.99	84.95	7.15	84.40	94.34
LLM + Doc2Doc	–	6.75	80.54	85.39	–	7.01	83.50	93.39
DELTA	**85.85**	**6.80**	81.80	86.33	85.26	**7.24**	85.25	95.89
SubDocTrans	85.79	6.76	**83.14**	**87.43**	**85.27**	**7.24**	**86.33**	**96.32**
`DeepSeek-R1:14b`								
LLM + Sentence	81.63	6.17	73.92	80.51	83.15	6.93	80.01	87.53
LLM + Context	82.01	6.27	74.12	81.41	83.55	7.01	81.39	91.17
LLM + Doc2Doc	–	6.54	80.03	86.41	–	**7.16**	84.48	93.83
DELTA	82.49	6.40	81.74	86.49	83.37	7.02	83.90	93.94
SubDocTrans	**83.41**	**6.57**	**81.92**	**87.58**	**84.06**	7.14	**85.05**	**94.47**
`DeepSeek-R1:32b`								
LLM + Sentence	83.16	6.42	74.17	82.75	83.47	6.99	77.71	86.98
LLM + Context	82.89	6.45	79.10	85.77	83.65	7.07	83.61	92.37
LLM + Doc2Doc	–	6.70	78.53	85.46	–	7.25	82.88	92.24
DELTA	83.29	6.51	**81.85**	86.76	83.39	7.07	84.85	93.54
SubDocTrans	**84.50**	**6.71**	80.78	**88.44**	**84.88**	**7.28**	**86.05**	**95.86**

Table 1: Test results on the IWSLT2017 dataset. Since the translations produced by the LLM + Doc2Doc method are not aligned at the sentence level with the source text, we do not report the sCOMET scores for this method. The best performance in each block is highlighted in **bold**, while the second-best is underlined. Part of the data is sourced from Wang et al. (2025).

across nearly all translation directions and models. In the Xx ⇒ En direction, SubDocTrans surpasses DELTA with significant gains, such as a 1.49 point improvement in sCOMET (84.88 vs. 83.39) and a 1.2 point increase in LTCR-1 (86.05 vs. 84.85) with `DeepSeek-R1:32b`. Compared to Sentence and Context, SubDocTrans boosts sCOMET by approximately 2 points and LTCR-1 by 6 points in En ⇒ Xx, while outperforming Doc2Doc by nearly 2 points in LTCR-1.

Table 2 reports results on the more challenging Guofeng dataset. SubDocTrans consistently outperforms all baselines. While DELTA slightly outperforms in LTCR-1 under `GPT-4o-mini`, SubDocTrans offers a better trade-off between quality and consistency, with clearer gains under stronger models. These findings also confirm the effectiveness of our multi-granularity knowledge and CoT reasoning.

4.3 Ablation Study

Results for RQ2: Effectiveness of Knowledge Augmentation. Table 3 presents an ablation study for the Zh ⇒ En direction in the Guofeng dataset, using `Deepseek-R1:32B` as the backbone model. For more detailed scores, please refer to Appendix A.4. When provided with Theme, the model exhibits improved translation quality scores, but no significant enhancement in translation consistency is observed. The introduction of Topics contributes to more consistent translations, as evidenced by an increase of 4.73 percentage points in LTCR-1_f across the baselines, reinforcing consistency by ensuring that translations reflect the core topics. Transition Hint enhances discourse fluency by improving coherence between sub-documents, leading to an increase of 0.10 points in sCOMET and 0.08 points in dCOMET. The combination of Theme, Topics, and Transition in the Doc2Doc + Summary + Proper Nouns baseline achieves the best performance.

System	sCOMET	dCOMET	LTCR-1	LTCR-1_f	sCOMET	dCOMET	LTCR-1	LTCR-1_f
		GPT-3.5-Turbo				GPT-4o-mini		
LLM + Sentence	77.62	3.07	61.58	78.82	77.87	3.10	58.82	70.59
LLM + Context	**78.57**	3.19	70.10	81.37	78.56	3.19	64.32	74.37
LLM + Doc2Doc	–	2.82	77.46	89.02	–	2.96	82.04	91.62
DELTA	78.45	3.17	**85.57**	96.52	78.77	3.34	**88.94**	**96.48**
SubDocTrans	**78.57**	**3.20**	81.64	**97.58**	**78.89**	**3.35**	81.55	94.66
		DeepSeek-R1:14b				DeepSeek-R1:32b		
LLM + Sentence	74.96	2.95	45.00	60.00	75.28	3.10	57.77	63.11
LLM + Context	74.89	2.98	42.57	65.84	75.76	3.17	62.32	74.88
LLM + Doc2Doc	–	3.02	59.31	77.45	–	3.26	60.49	81.46
DELTA	**76.78**	3.18	79.81	92.79	76.55	3.22	84.54	93.72
SubDocTrans	76.52	**3.21**	**86.87**	**95.45**	**78.06**	**3.42**	**89.05**	**96.52**

Table 2: Test results on the Guofeng dataset. Part of the data is sourced from (Wang et al., 2025).

Setting	sCOMET	dCOMET	LTCR-1	LTCR-1_f
Doc2Doc + Summary + Proper Nouns	77.73	3.34	79.23	92.27
+ Theme	77.83	3.38	74.38	89.66
+ Topics	77.95	3.37	80.49	95.12
+ Transition Hint	77.88	3.38	82.67	93.56
+ Theme + Topics	77.96	3.34	84.39	94.39
+ Theme + Transition Hint	77.92	3.38	75.12	90.24
+ Topics + Transition Hint	77.66	3.34	88.24	**96.57**
+ Theme + Topics + Transition Hint	**78.06**	**3.42**	**89.05**	96.52

Table 3: Ablation Study.

4.4 Effectiveness of SubDocTrans

We conducted additional experiments to further evaluate the robustness and practicality of SubDocTrans. Specifically, we evaluate its effectiveness in handling discourse phenomena and pronoun consistency (RQ3), its ability to mitigate sentence omissions and hallucinations (RQ4), and its inference efficiency compared to existing methods (RQ5).

Results for RQ3: Handling of Discourse Phenomena. To evaluate whether our method effectively uses document-level context to enhance discourse coherence, we utilized the Zh-En test set from Sun et al. (2022), covering tense consistency (TC), conjunction presence (CP) and pronoun translation (PT). The overall TCP score, defined as the geometric mean of these metrics, aligns well with human judgments. As shown in Table 4, SubDocTrans achieves the best performance in discourse phenomena compared to all baselines under DeepSeek-R1:32B.

For pronoun translation, we adopt the accuracy of pronoun translation (APT) metric from Miculicich Werlen and Popescu-Belis (2017), refined with alignment heuristics, to evaluate the correctness of pronoun triplets (source, reference, candidate). Table 5 shows that SubDocTrans achieves the highest APT score in En ⇒ Zh, confirming its strength in preserving pronoun consistency across document contexts.

System	TC	CP	PT	TCP
LLM + Sentence	56.2	31.6	65.2	48.8
LLM + Context	55.1	31.9	63.5	48.1
LLM + Doc2Doc	58.7	**32.7**	67.9	50.7
DELTA	57.3	29.9	62.2	47.4
SubDocTrans	**60.3**	32.1	**68.9**	**51.1**

Table 4: Evaluation results of discourse phenomena.

Metric	Sentence	Context	Doc2Doc	DELTA	SubDocTrans
APT	58.58	58.65	58.30	57.90	**58.67**

Table 5: Evaluation results of pronoun translation accuracy (APT).

Results for RQ4: Omission Errors and Hallucination Mitigation. Table 6 shows that SubDocTrans significantly reduces omission errors compared to Doc2Doc, with an omission rate of 0.75 vs. 8.75. This improvement stems from the use of a sentence alignment strategy that enforces one-to-one mapping between source and target sentences. We also reports Avg. Retries, measuring the average number of attempts needed by our method

System	Missing Sents	Avg. Retries	Time (hour)
LLM + Doc2Doc	8.75	–	**0.21**
SubDocTrans	**0.75**	0.40	0.48

Table 6: Effect of the sentence alignment strategy on omission errors. Since the LLM + Doc2Doc method does not require retries, Avg. Retries is not reported.

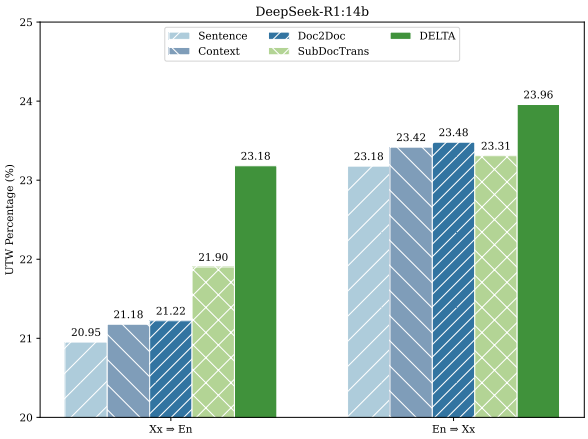

Figure 2: Comparison of UTW percentages between our method and other baselines.

to produce a valid output, and Time, indicating total translation time. While Doc2Doc achieves the fastest runtime by translating multiple sentences without alignment and retry mechanisms, SubDocTrans completes translation in 0.48 hours with only 0.40 retries, demonstrating that it ensures high-quality translations with minimal cost.

To further evaluate hallucinations, we compute Unaligned Translation Words (UTW) using the alignment tool from Dou and Neubig (2021). This measurement is also used by Hendy et al. (2023) to investigate the presence of words that do not support the source sentences. Figure 2 reports UTW scores on IWSLT2017 across four language pairs using `DeepSeek-R1:14B`. SubDocTrans achieves an UTW of 23.31 in En ⇒ Xx, lower than that of Context and Doc2Doc, respectively, and consistently below DELTA. This suggests that SubDocTrans helps reduce hallucinations, thereby improving the translation quality.

Results for RQ5: Inference Efficiency. In Figure 3, we utilized DeepSeek-R1:14B and DeepSeek-R1:32B to translate documents in the En ⇒ Zh, De, Fr, and Ja directions. Since the Doc2Doc method translates multiple sentences at once without any additional input, it achieves the highest efficiency. Apart from Doc2Doc, our method outperforms the other baselines.

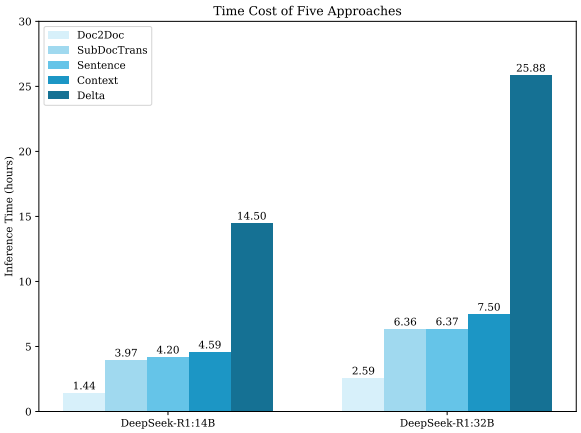

Figure 3: Comparison of time cost between our method and other baselines.

Other extended analysis and implementation details are reported in Appendix A.5.

5 Conclusion

This paper proposed SubDocTrans, a novel DocMT framework that leverages sub-document segmentation, five knowledge augmentations, and chain-of-thought reasoning to address the challenges of document-level translation. By segmenting documents into overlapping sub-documents of similar topics, SubDocTrans ensures seamless inter-paragraph cohesion and robust contextual modeling. The sentence alignment strategy further mitigates omission errors. Experimental results show that our framework outperforms all advanced baselines, and effectively improves the accuracy of pronoun translation and mitigate hallucinations.

6 Limitations

Our SubDocTrans demonstrates significant improvements in DocMT, while some concerns can be further investigated. First, the computational overhead of the framework is driven by the knowledge extraction. This may limit its applicability in real-time or resource-constrained settings. Second, the use of LLMs for extracting proper nouns and their translations, though flexible, may be less efficient than lightweight, rule-based extraction techniques. Employing more precise and task-specific techniques for this purpose could further improve the overall efficiency of our framework.

7 Ethics Statement

This work builds upon publicly available datasets that are widely used in the community. These

datasets do not contain personally identifying information or offensive content. No new user-generated or sensitive data was collected or processed in this study. Our use of these datasets complies with their original licenses and intended use.

8 Acknowledgements

This work was supported by the fund from the research center of Star-XLAB in Xiamen University.

References

Griffin Adams, Alexander Fabbri, Faisal Ladhak, Eric Lehman, and Noémie Elhadad. 2023. From sparse to dense: Gpt-4 summarization with chain of density prompting. *Preprint*, arXiv:2309.04269.

Yasuhiro Akiba, Marcello Federico, Noriko Kando, Hiromi Nakaiwa, Michael Paul, and Jun'ichi Tsujii. 2004. Overview of the IWSLT evaluation campaign. In *Proceedings of the First International Workshop on Spoken Language Translation: Evaluation Campaign*, Kyoto, Japan.

Simone Conia, Daniel Lee, Min Li, Umar Farooq Minhas, Saloni Potdar, and Yunyao Li. 2024. Towards cross-cultural machine translation with retrieval-augmented generation from multilingual knowledge graphs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16343–16360.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *ArXiv preprint*, abs/2207.04672.

Menglong Cui, Jiangcun Du, Shaolin Zhu, and Deyi Xiong. 2024. Efficiently exploring large language models for document-level machine translation with in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10885–10897, Bangkok, Thailand. Association for Computational Linguistics.

David Dale, Elena Voita, Loïc Barrault, and Marta R. Costa-jussà. 2022. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity even better. *Preprint*, arXiv:2212.08597.

David Dale, Elena Voita, Janice Lam, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Loïc Barrault, and Marta R. Costa-jussà. 2023. Halomi: A manually annotated benchmark for multilingual hallucination and omission detection in machine translation. *Preprint*, arXiv:2305.11746.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Xavier Garcia and Orhan Firat. 2022. Using natural language prompts for machine translation. *Preprint*, arXiv:2202.11822.

Nuno M. Guerreiro, Elena Voita, and André F. T. Martins. 2023. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. *Preprint*, arXiv:2208.05309.

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. *Preprint*, arXiv:2501.08523.

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. *Preprint*, arXiv:2302.09210.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *Preprint*, arXiv:2301.08745.

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.

Bryan Li, Jiaming Luo, Eleftheria Briakou, and Colin Cherry. 2025. Leveraging domain knowledge at inference time for llm translation: Retrieval versus generation. In *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pages 91–106.

Sinuo Liu, Chenyang Lyu, Minghao Wu, Longyue Wang, Weihua Luo, Kaifu Zhang, and Zifu Shang. 2025. New trends for modern machine translation with large reasoning models. *Preprint*, arXiv:2503.10351.

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.

Raphaël Merx, Aso Mahmudi, Katrina Langford, Leo Alberto de Araujo, and Ekaterina Vylomova. 2024. Low-resource machine translation through retrieval-augmented llm prompting: A study on the mambai language. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI)@ LREC-COLING 2024*, pages 1–11.

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. Validation of an automatic metric for the accuracy of pronoun translation (APT). In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 17–25, Copenhagen, Denmark. Association for Computational Linguistics.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, and 244 others. 2024. Openai o1 system card. *Preprint*, arXiv:2412.16720.

Ming Qian, HQ Wu, Lenny Yang, and Arthur Wan. 2023. Augmented machine translation enabled by gpt4: Performance evaluation on human-machine teaming approaches. In *Proceedings of the First Workshop on NLP Tools and Resources for Translation and Interpreting Applications*, pages 20–31.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. Rethinking document-level neural machine translation. *Preprint*, arXiv:2010.08961.

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.

Ke Wang, Jun Xie, Yuqi Zhang, and Yu Zhao. 2023a. Improving neural machine translation by multi-knowledge integration with prompting. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023b. 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. 2023c. 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.

Yutong Wang, Jiali Zeng, Xuebo Liu, Fandong Meng, Jie Zhou, and Min Zhang. 2024. TasTe: Teaching large language models to translate through self-reflection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6144–6158, Bangkok, Thailand. Association for Computational Linguistics.

Yutong Wang, Jiali Zeng, Xuebo Liu, Derek F. Wong, Fandong Meng, Jie Zhou, and Min Zhang. 2025. Delta: An online document-level translation agent based on multi-level memory. *Preprint*, arXiv:2410.08143.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *Preprint*, arXiv:2109.01652.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024. Adapting large language models for document-level machine translation. *Preprint*, arXiv:2401.06468.

Yangjian Wu and Gang Hu. 2023. Exploring prompt engineering with GPT language models for document-level machine translation: Insights and findings. In *Proceedings of the Eighth Conference on Machine Translation*, pages 166–169, Singapore. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR.

Zhuocheng Zhang, Shuhao Gu, Min Zhang, and Yang Feng. 2023b. Addressing the length bias problem in document-level neural machine translation. *Preprint*, arXiv:2311.11601.

A Appendix

A.1 The algorithm of SubDocTrans

The complete procedure of SubDocTrans is outlined in Algorithm 1, which illustrates how document segmentation, knowledge extraction, and translation are integrated into a unified framework.

A.2 Prompts

This part presents the prompt templates used in each module of our framework. The prompt for the Bilingual Summary and Theme Writer in Figure 4, followed by the Proper Nouns Extractor in Figure 5, the Topic Extractor in Figure 6, and the Transition Hint Writer in Figure 7. Figure 8 illustrates the prompt used by the Knowledge-augmented Translation Module to incorporate multi-level contextual knowledge. All prompt designs are instruction-style and follow a structured format to elicit reliable and consistent responses from the LLM.

A.3 Detailed Results of the Main Experiment

The scores for the En ⇒ Zh, De, Fr, Ja translation directions are presented in Table 7, while the scores for the Zh, De, Fr, Ja ⇒ En are shown in Table 8.

A.4 Detailed Results of the Ablation Study

The Detailed Ablation Study are shown in Table 9. Based on Doc2Doc with Bilingual Summary, Proper Nouns, and their combination as baselines. When provided with Theme, the model exhibits improved translation quality scores across all three baselines, but no significant enhancement in translation consistency is observed. The introduction of Topics contributes to more consistent translations, as evidenced by an average increase of 1.93 percentage points in LTCR-1 and 3.00 percentage points in LTCR-1_f across the baselines. Transition Hint enhances discourse fluency by improving coherence between sub-documents, leading to an average increase of 0.16 points in sCOMET and 0.05 points in dCOMET, with significant gains in the Summary baseline, where sCOMET reaches 78.31 and dCOMET reaches 3.41. The combination of Theme, Topics, and Transition Hint in the Doc2Doc + Summary + Proper Nouns baseline achieves the best performance.

A.5 Additional Details on Inference Efficiency

SubDocTrans involves three stages: *Document Segmentation*, *Knowledge Extraction*, and *Knowledge-Augmented Translation*. The first stage is lightweight, while most time cost arises from LLM calls in the latter two. Specifically, the number of LLM calls equals the number of sub-documents multiplied by $(1 + r_{\text{avg}})$, where r_{avg} is the average retry count per sub-document.

Algorithm 1: The Overall Framework of SubDocTrans

input : Source document $\boldsymbol{D}_s = \{s_1, \ldots, s_N\}$, Large Reasoning model $\mathcal{L}$, Overlap Sentences o, Maximum Retry Attempts r_{max}, Proper Nouns $\mathcal{P} = \emptyset$, Source and Target Language Summaries $\mathcal{A}_s = \emptyset, \mathcal{A}_t = \emptyset$, Theme $\mathcal{T} = \emptyset$, Topics $\mathcal{O} = \emptyset$, Transition Hint $\mathcal{H} = \emptyset$

output : Target document $\boldsymbol{D}_t = \{t_1, \ldots, t_N\}$

$\boldsymbol{D}_t \leftarrow \emptyset$

$\boldsymbol{C} \leftarrow \mathcal{L}_{\text{Segment}}(\boldsymbol{D}_s, o)$

for $i = 1$ *to* $|\boldsymbol{C}|$ **do**

  $\boldsymbol{C}_i \leftarrow \{s_l, \ldots, s_m\}$

  $\mathcal{H}_i \leftarrow \mathcal{L}_{\text{Hint}}(\boldsymbol{C}_{i-1}, \boldsymbol{C}_i)$

  $\mathcal{O}_i \leftarrow \mathcal{L}_{\text{Topics}}(\boldsymbol{C}_i)$

  $\mathcal{T}_i \leftarrow \mathcal{L}_{\text{Theme}}(\boldsymbol{C}_i)$

  $\mathcal{T} \leftarrow \mathcal{T} \cup \{\mathcal{T}_i\}$

  $\tilde{\mathcal{A}}_{s,i}, \tilde{\mathcal{A}}_{t,i} \leftarrow \mathcal{L}_{\text{Summary}}(\boldsymbol{C}_i)$

  $\mathcal{A}_s \leftarrow \mathcal{A}_s \cup \{\tilde{\mathcal{A}}_{s,i}\}$

  $\mathcal{A}_t \leftarrow \mathcal{A}_t \cup \{\tilde{\mathcal{A}}_{t,i}\}$

  $\mathcal{P}_i \leftarrow \mathcal{L}_{\text{ProperNouns}}(\boldsymbol{C}_i, \mathcal{A}_s)$

  $\mathcal{P} \leftarrow \mathcal{P} \cup \{\mathcal{P}_i\}$

end

$\mathcal{P} \leftarrow \mathcal{L}_{\text{MergeProperNouns}}(\mathcal{P})$

$\mathcal{A}_s \leftarrow \mathcal{L}_{\text{MergeS}}(\mathcal{A}_s)$

$\mathcal{A}_t \leftarrow \mathcal{L}_{\text{MergeT}}(\mathcal{A}_t)$

$\mathcal{T} \leftarrow \mathcal{L}_{\text{MergeTheme}}(\mathcal{T})$

for $i = 1$ *to* $|\boldsymbol{C}|$ **do**

  $\boldsymbol{T}_i \leftarrow \mathcal{L}_{\text{Translate}}(\boldsymbol{C}_i, \mathcal{P}, \mathcal{A}_s, \mathcal{A}_t, \mathcal{T}, \mathcal{O}_i, \mathcal{H}_i)$

  $\boldsymbol{D}_t \leftarrow \boldsymbol{D}_t \cup \boldsymbol{T}_i$

end

Prompt for Bilingual Summary and Theme Writer

You are a professional article analysis agent. You are provided with a context (delimited by “‘) and you need to summarize the input text into an abstract and a theme distribution.
Thought 1: Generate a {SOURCE_LANG} language summary (30–60 words), including the main contents of these sentences, the overall domain, style, and tone, while preserving key information as much as possible.
Thought 2: Translate the {SOURCE_LANG} summary to {TARGET_LANG}.
Thought 3: Generate a theme: Output the article’s theme in {TARGET_LANG} (e.g., politics 60%, economy 30%, technology 10%).

Paragraph: {PARAGRAPH}

Format your output as a list of JSON. Like the following:

```
[
   {
       "{SOURCE_LANG} summary": "",
       "{TARGET_LANG} summary": "",
       "Theme": ""
   }
]
```

Figure 4: Prompt template for Bilingual Summary and Theme Writer.

Prompt for Proper Nouns Extractor

You are a {SOURCE_LANG}-{TARGET_LANG} bilingual expert. You are provided with a context chunk (delimited by “‘) containing a summary and a {SOURCE_LANG} paragraph.
Thought 0: Use the summary as a global reference to understand the broader context of the entire paragraph. The summary helps identify key entities (e.g., persons, locations, organizations) or concepts that might be ambiguous or unrecognized in the context.
Thought 1: Extract all the proper nouns from the {SOURCE_LANG} paragraph and translate them to {TARGET_LANG}. Ensure that the {TARGET_LANG} translations are accurate, consistent, and culturally appropriate.

Summary: {SUMMARY}
Paragraph: {PARAGRAPH}

Format your output as a list of JSON. Like the following:

```
[
   {
       "{SOURCE_LANG} proper nouns": "",
       "{TARGET_LANG} proper nouns": ""
   }
]
```

Figure 5: Prompt template for Proper Nouns extraction

Prompt for Topics Extraction

You are a professional article analysis agent. You are provided with a context (delimited by “‘). Use a few words to describe the topics of the following input {SOURCE_LANG} article.

Paragraph: {PARAGRAPH}

Format your output as a list of JSON. Like the following:

```
[
   {
       "Topics": ""
   }
]
```

Figure 6: Prompt template for Topics extraction.

Prompt for Transition Hint Writer

You are a professional article analysis agent. Please generate a hint for the input {SOURCE_LANG} paragraph, linking its logical relationship with the previous paragraph. The transition hint should be concise (20–30 words) and reflect logical connections (e.g., cause-effect, progression). You are provided with a context (delimited by “‘).
Thought 1: Analyze the logical relationship between the two paragraphs (e.g., cause-effect, progression, contrast).
Thought 2: Generate a {SOURCE_LANG} language transition hint describing how the current paragraph follows the previous one.
Thought 3: Self-evaluate whether the transition hint accurately reflects the logical relationship; if inaccurate, optimize it.
Thought 4: Translate the {SOURCE_LANG} hint into {TARGET_LANG}.

Paragraph 1: {PARAGRAPH_1}
Paragraph 2: {PARAGRAPH_2}

Format your output as a list of JSON. Like the following:

```
[
   {
       "Transition Hint": ""
   }
]
```

Figure 7: Prompt template for Transition Hint Writer.

Prompt template for the Knowledge-augmented Translation.

You are a {SOURCE_LANG}-{TARGET_LANG} bilingual expert translating a long {SOURCE_LANG} article. You are provided with a context (delimited by "') including the article's {SOURCE_LANG} and {TARGET_LANG} summaries, theme, transition hint between the previous and current paragraphs, the current paragraph's topic, and historical translations of some proper nouns.
Translate the current {SOURCE_LANG} source paragraph into {TARGET_LANG} as a cohesive unit, ensuring that proper nouns remain consistent with their historical translations and that the style is coherent across all sentences.

Summaries:
<{SRC_LANG} summary> {SRC_SUMMARY}
<{TGT_LANG} summary> {TGT_SUMMARY}

Proper nouns:
{PROPER NOUNS}

Theme:
{THEME}

Topics:
{TOPICS}

Transition hint:
{TRANSITION_HINT}

Format your output as a list of JSON. Like the following:

```
[
   {
       "{SOURCE_LANG} sentence": "First line",
       "{TARGET_LANG} translation": "First line translated to {TARGET_LANG}"
   },
   // ... one object per line
]
```

Figure 8: Prompt template for the Knowledge-augmented Translation.

System	sCOMET	dCOMET	LTCR-1	LTCR-1_f	sCOMET	dCOMET	LTCR-1	LTCR-1_f
	En ⇒ Zh				En ⇒ De			
NLLB	76.81	6.20	71.68	84.96	84.15	6.64	91.76	99.61
GOOGLE	78.46	5.76	89.45	92.36	80.23	5.78	93.55	99.19
`GPT-3.5-Turbo`								
LLM + Sentence	83.78	6.55	80.27	88.78	84.97	6.71	92.06	98.81
LLM + Context	84.50	6.68	77.89	87.41	85.12	6.74	**93.70**	99.21
LLM + Doc2Doc	–	6.29	82.04	94.29	–	**6.81**	88.89	97.94
DELTA	84.70	6.72	86.44	95.25	**85.37**	6.78	93.46	99.23
Ours	**84.75**	**6.73**	**87.76**	**98.30**	85.10	6.75	92.05	**99.24**
`GPT-4o-mini`								
LLM + Sentence	82.13	6.43	78.04	91.89	81.41	6.39	90.70	98.84
LLM + Context	84.36	6.68	78.95	93.42	84.83	6.70	92.66	99.61
LLM + Doc2Doc	–	6.60	82.33	88.35	–	**6.90**	91.05	99.22
DELTA	84.94	**6.81**	85.52	91.72	**85.47**	6.79	92.19	**100.0**
Ours	**85.22**	6.80	**86.09**	**94.04**	85.00	6.71	**93.46**	99.23
`DeepSeek-R1:14b`								
LLM + Sentence	83.32	6.59	74.71	85.44	79.82	6.00	86.32	95.75
LLM + Context	83.81	6.72	72.98	81.40	79.57	6.04	81.69	95.31
LLM + Doc2Doc	–	6.92	78.74	88.04	–	**6.26**	84.74	95.98
DELTA	83.56	6.76	85.53	90.13	80.07	6.15	86.75	**96.79**
Ours	**84.91**	**7.00**	**87.38**	**90.94**	**80.97**	6.20	**87.06**	96.47
`DeepSeek-R1:32b`								
LLM + Sentence	84.47	6.79	77.09	86.91	81.79	6.33	83.98	94.17
LLM + Context	83.89	6.78	76.60	85.11	81.24	6.32	85.32	93.58
LLM + Doc2Doc	–	6.92	79.58	87.54	–	**6.60**	87.35	96.33
DELTA	83.92	6.87	82.67	88.33	81.78	6.38	88.67	98.05
Ours	**85.09**	**7.00**	**84.35**	**90.14**	**83.22**	6.56	**91.37**	**98.43**
	En ⇒ Fr				En ⇒ Ja			
NLLB	85.35	6.23	87.84	89.86	82.12	6.37	46.94	53.06
GOOGLE	82.21	5.53	88.61	91.46	80.74	6.23	53.92	55.88
`GPT-3.5-Turbo`								
LLM + Sentence	85.84	6.18	83.55	88.49	84.61	6.89	52.34	55.14
LLM + Context	**86.49**	6.27	83.06	89.25	85.50	7.09	54.72	56.60
LLM + Doc2Doc	–	6.28	**92.28**	**94.63**	–	7.10	53.26	58.70
DELTA	86.48	**6.30**	88.96	94.16	**85.76**	**7.13**	**62.96**	**66.67**
Ours	86.44	**6.30**	90.49	94.17	85.54	7.10	61.82	63.64
`GPT-4o-mini`								
LLM + Sentence	80.79	5.82	88.89	90.91	81.72	6.74	56.73	58.65
LLM + Context	85.10	6.14	90.52	92.81	84.84	7.09	57.89	62.11
LLM + Doc2Doc	–	6.24	**91.00**	**94.00**	–	7.25	57.78	60.00
DELTA	86.38	**6.28**	90.94	93.85	**86.61**	**7.32**	58.54	59.76
Ours	**86.50**	6.26	89.97	91.22	86.42	7.28	**63.04**	**65.22**
`DeepSeek-R1:14b`								
LLM + Sentence	82.57	5.87	84.64	85.77	80.80	6.22	50.00	55.07
LLM + Context	83.00	5.95	83.15	87.27	81.67	6.35	58.65	61.65
LLM + Doc2Doc	–	6.13	86.32	88.27	–	6.84	**70.30**	**73.33**
DELTA	83.44	6.06	**93.42**	**95.07**	82.88	6.62	61.26	63.96
Ours	**83.64**	**6.14**	87.46	90.43	**84.10**	**6.92**	65.77	72.48
`DeepSeek-R1:32b`								
LLM + Sentence	83.45	5.96	81.34	85.56	82.91	6.59	54.26	64.34
LLM + Context	83.42	6.03	84.12	89.89	82.99	6.65	**70.34**	**74.48**
LLM + Doc2Doc	–	**6.24**	88.50	92.65	–	7.03	58.67	65.33
DELTA	83.84	6.07	**94.95**	**97.16**	83.61	6.71	61.11	63.49
Ours	**84.93**	6.23	84.59	92.46	**84.77**	**7.05**	62.81	72.73

Table 7: Detailed results of our experiments in En ⇒ Xx directions.

System	sCOMET	dCOMET	LTCR-1	LTCR-1_f	sCOMET	dCOMET	LTCR-1	LTCR-1_f
	Zh ⇒ En				De ⇒ En			
NLLB	82.14	7.01	75.31	88.27	85.63	7.23	95.98	98.85
GOOGLE	78.06	5.68	72.89	86.14	82.31	6.47	96.53	98.27
`GPT-3.5-Turbo`								
LLM + Sentence	83.34	7.17	73.99	86.71	85.92	7.26	**98.88**	**100.0**
LLM + Context	83.88	7.29	76.92	90.53	86.10	**7.30**	98.30	**100.0**
LLM + Doc2Doc	–	7.08	76.77	88.39	–	7.16	98.24	98.82
DELTA	83.88	7.30	**80.00**	93.53	86.14	**7.30**	98.33	**100.0**
Ours	**84.08**	**7.34**	79.43	**94.29**	**86.40**	**7.30**	98.85	**100.0**
`GPT-4o-mini`								
LLM + Sentence	83.55	7.24	71.93	84.80	85.11	7.17	98.20	**100.0**
LLM + Context	83.96	7.35	78.24	91.18	86.12	7.27	**98.88**	**100.0**
LLM + Doc2Doc	–	7.15	79.62	92.36	–	7.19	95.24	97.62
DELTA	84.10	**7.47**	79.41	94.71	**86.61**	7.31	98.32	**100.0**
Ours	**84.33**	7.41	**83.15**	**95.51**	86.20	**7.35**	98.84	**100.00**
`DeepSeek-R1:14b`								
LLM + Sentence	82.45	7.20	69.54	79.89	84.18	7.12	97.10	97.10
LLM + Context	82.72	7.24	78.24	88.82	84.75	7.20	95.68	98.56
LLM + Doc2Doc	–	7.38	82.46	93.57	–	7.31	96.55	98.28
DELTA	82.56	7.29	80.23	90.12	82.64	7.15	95.71	96.93
Ours	**83.73**	**7.42**	**84.88**	**94.19**	**85.85**	**7.34**	**97.66**	**98.83**
`DeepSeek-R1:32b`								
LLM + Sentence	82.76	7.22	60.12	82.66	84.42	7.18	94.57	98.45
LLM + Context	83.11	7.30	68.79	83.24	84.70	7.24	95.71	96.43
LLM + Doc2Doc	–	7.42	81.14	90.86	–	**7.36**	97.01	98.80
DELTA	83.00	7.31	83.62	94.35	85.42	7.30	98.19	99.40
Ours	**84.22**	**7.50**	**85.14**	**96.00**	**85.74**	7.34	**98.83**	**99.42**
	Fr ⇒ En				Ja ⇒ En			
NLLB	87.59	6.79	93.56	97.42	81.02	6.90	51.27	78.48
GOOGLE	84.64	6.19	95.63	96.83	75.67	5.50	60.67	82.00
`GPT-3.5-Turbo`								
LLM + Sentence	87.60	6.78	94.94	97.89	81.00	6.98	60.12	82.82
LLM + Context	**88.03**	6.84	94.96	97.90	81.85	7.17	69.94	92.64
LLM + Doc2Doc	–	6.78	94.78	97.39	–	6.80	70.90	87.31
DELTA	88.02	**6.86**	96.17	98.30	81.76	7.13	71.60	**93.21**
Ours	88.02	6.84	**97.47**	**99.49**	**81.93**	**7.27**	**71.76**	91.76
`GPT-4o-mini`								
LLM + Sentence	87.32	6.77	94.42	97.85	80.04	6.76	61.11	82.72
LLM + Context	87.72	6.81	94.85	97.42	82.00	7.17	65.62	88.75
LLM + Doc2Doc	–	6.83	94.42	98.28	–	6.86	64.71	85.29
DELTA	88.13	**6.90**	96.58	98.72	82.20	7.29	**66.67**	**90.12**
Ours	**88.15**	6.88	**97.99**	**100.0**	**82.38**	**7.33**	65.34	89.77
`DeepSeek-R1:14b`								
LLM + Sentence	86.32	6.75	95.88	**98.45**	79.65	6.86	57.58	78.79
LLM + Context	86.81	6.80	93.72	96.34	79.92	6.94	58.02	81.48
LLM + Doc2Doc	–	**6.91**	93.91	96.95	–	7.26	**68.26**	**91.62**
DELTA	86.85	6.84	94.55	98.02	80.27	7.01	65.27	90.42
Ours	**87.55**	6.90	**94.85**	96.39	**81.10**	**7.31**	67.30	90.57
`DeepSeek-R1:32b`								
LLM + Sentence	86.75	6.78	94.33	96.91	80.40	6.91	52.35	71.76
LLM + Context	86.85	6.83	96.43	**99.49**	80.74	7.08	66.06	87.27
LLM + Doc2Doc	–	6.90	95.92	98.47	–	7.32	58.43	80.72
DELTA	87.07	6.84	**97.56**	99.02	80.72	7.09	65.06	86.14
Ours	**87.80**	**6.92**	92.83	98.44	**81.77**	**7.35**	**67.65**	**90.59**

Table 8: Detailed results of our experiments in Xx ⇒ En directions.

Setting	sCOMET	dCOMET	LTCR-1	LTCR-1_f
Baseline: Doc2Doc + Summary				
Doc2Doc + Summary	77.81	3.37	68.60	85.99
+ Theme	77.97	3.38	74.02	80.39
+ Topics	77.73	3.34	66.99	82.04
+ Transition	77.92	3.40	**77.56**	84.39
+ Theme + Topics	77.88	3.36	68.29	87.32
+ Theme + Transition	78.25	3.38	76.85	81.77
+ Topics + Transition	77.94	3.35	72.33	81.07
+ Theme + Topics + Transition	**78.31**	**3.41**	74.63	**87.80**
Baseline: Doc2Doc + Proper Nouns				
Doc2Doc + Proper Nouns	77.57	3.37	82.84	95.59
+ Theme	77.95	**3.39**	76.59	93.17
+ Topics	77.62	3.37	81.19	**96.53**
+ Transition	77.79	3.34	76.24	94.55
+ Theme + Topics	77.90	3.35	78.11	93.53
+ Theme + Transition	77.65	3.35	77.67	94.66
+ Topics + Transition	**78.02**	3.37	80.88	96.57
+ Theme + Topics + Transition	77.84	3.37	**83.98**	94.17
Baseline: Doc2Doc + Summary + Proper Nouns				
Doc2Doc + Summary + Proper Nouns	77.73	3.34	79.23	92.27
+ Theme	77.83	3.38	74.38	89.66
+ Topics	77.95	3.37	80.49	95.12
+ Transition	77.88	3.38	82.67	93.56
+ Theme + Topics	77.96	3.34	84.39	94.39
+ Theme + Transition	77.92	3.38	75.12	90.24
+ Topics + Transition	77.66	3.34	88.24	**96.57**
+ Theme + Topics + Transition	**78.06**	**3.42**	**89.05**	96.52

Table 9: Ablation study evaluating novel knowledges (Theme, Topics, Transition Hint) and prior knowledges (Bilingual Summary, Proper Nouns) on Zh $\Rightarrow$ En document-level translation in the Guofeng dataset.