G^{2}C-MT: Graph-Guided Context Selection for Document-Level Machine Translation

Baijun Ji[1,2*], **Zixuan Zhou**[1], **Xiangyu Duan**[1], **Yu Liu**[2], **Longbo Sun**[2], **Rupu Wei**[2] and **Bohong Zhao**[2]

[1]School of Computer Science and Technology, Soochow University
[2]Trip.com Group
cocaer.cl@gmail.com, zxzhou1213@stu.suda.edu.cn, xyduan@suda.edu.cn,
{liu.yub, lbsun, rpwei, bohongzhao}@trip.com

Abstract

Effective document-level machine translation (DocMT) requires capturing long-range discourse dependencies. Recent work has explored retrieval-based and discourse-aware context selection. However, these approaches often lack an explicit mechanism for modeling structured discourse dependencies between distant paragraphs in a document. In this paper, we propose G²C-MT (Graph-Guided Context for Machine Translation), which views DocMT context selection as a structured path discovery problem on a lightweight discourse graph, rather than retrieving unstructured context sets or relying on expensive LLM-based discourse modeling. In detail, we represent each paragraph as a node and model the relationship between each pair of nodes, considering their semantic similarity, adjacency, and keyword overlap. Furthermore, we propose a depth-biased random walk over the graph to sample a backward context path for each target paragraph. The context path will be used to prompt a large language model (LLM) for translation. This framework naturally supports multi-path context sampling, which can improve robustness by aggregating diverse translation candidates for discourse-ambiguous inputs. Experiments conducted across various domains show that G²C-MT outperforms strong baselines on multiple LLMs, including DeepSeek-V3, Gemini-2.5-Flash-lite, and the Qwen-2.5/3 series.

1 Introduction

High-quality document-level machine translation requires more than accurate sentence-level translation. It also needs the preservation of discourse phenomena, including lexical consistency and coreference resolution. Capturing long-range dependencies is therefore essential. Although recent advances in LLMs have shown success in handling long contexts [Liu *et al.*, 2023; Chen *et al.*, 2023; Gao *et al.*, 2025], translating an entire document in a single pass often leads to issues such as sentence omissions or context dilution [Wang *et al.*, 2025]. Moreover, exposing the model to the whole document context is inefficient, since the decoding cost of LLMs increases quadratically with the length of the input text. To address this issue, recent studies have employed retrieval-based and graph-based strategies to select prior translated paragraphs as context. Retrieval-based methods select historical translations according to semantic similarity to mitigate long-range dependencies [Wang *et al.*, 2025]. However, these methods often fail to preserve explicit discourse structures, as they treat sentences as an unstructured collection. Similarly, existing graph-based methods depend on expensive edge relation definitions via LLM-based relation classification, and they are usually limited to selecting first-order neighbors as historical context [Dutta *et al.*, 2025; Pham *et al.*, 2025]. This restriction prevents them from capturing deep and multi-hop discourse paths that define the global document structure.

To overcome the above problem, we propose G^{2}C-MT, a novel Graph-Guided Context for Machine Translation framework. Unlike retrieving sentences in isolation or relying on expensive graph construction processes, we model the document's discourse structure as a weighted directed acyclic graph (DAG) using a lightweight graph construction procedure. Specifically, each paragraph serves as a node, while edges denote the relationship between paragraphs. These relationships are quantified through a fusion score that derived from semantic similarity, sequential adjacency, and lexical overlap. This rich metric enables our framework to model the document as a graph with complex semantic relations rather than a simple linear chain.

Based on the constructed discourse graph, we apply a graph-driven method to select historical context for document-level translation dynamically. Specifically, when translating one target paragraph, we perform a backward and biased random walk starting from the corresponding node. Then, we identify a related context path, which is composed of previous paragraphs along with their translations. Unlike previous approaches that restrict context to immediate neighbors [Dutta *et al.*, 2025], we guide the traversal with two complementary signals: local edge weights that encode semantic and lexical relevance, and a global signal that encourages traversing nodes that can generate longer and richer discourse chains. This design enables the model to select a single, struc-

*Work done at Trip.com Group.

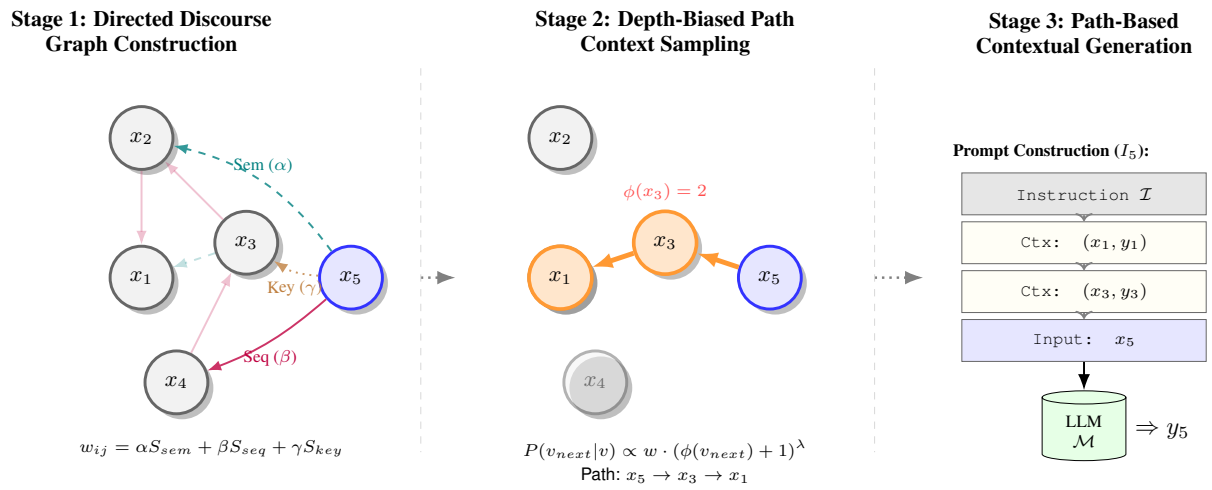

Figure 1: **Overview of the G^{2}C-MT Framework.** The process involves three stages: (1) Constructing a discourse graph considering semantic (α), sequential (β), and keyword (γ) cohesion; (2) Traversing a context path via Depth-Biased Random Walk (backtracking from target to history), where nodes with higher depth potential ϕ (e.g., x_3) and edge score attract the walker; (3) Formatting the path into a structured prompt for translation.

tured context path that captures long-range, non-linear dependencies while remaining computationally efficient. The selected path is subsequently formatted into a discourse-aware prompt, enabling the model to exploit structured contextual information without processing the entire text. Moreover, since the traversal is probabilistic rather than deterministic, G^{2}C-MT naturally supports multi-path sampling. By exploring multiple plausible discourse paths and aggregating the resulting translations, the framework can improve robustness in the presence of discourse-level ambiguity

Our main contributions are as follows:

- We propose G^{2}C-MT, a novel framework that models document context as a DAG, which can be built in a lightweight way. The graph can capture non-linear discourse dependencies effectively by modeling semantic and sequential correlations.
- We design a biased random walk mechanism that explicitly favors deeper context paths containing richer discourse information. Meanwhile, this probabilistic traversal can inherently support multi-path sampling, which can enhance robustness when the target paragraph involves discourse ambiguities.
- We conduct extensive experiments on various document-level translation benchmarks using LLMs of different scales. Our method outperforms strong baselines in both translation quality and coherence. Further analysis confirms the effectiveness of our graph-guided approach in capturing long-range dependencies.

2 Methodology

As illustrated in Figure 1 and Algorithm 1, the overall pipeline of our method can be divided into the following three stages:

1. Directed Discourse Graph Construction, which treats each paragraph as a node and builds edges between paragraphs considering their multi-dimensional relevance.
2. Depth-Biased Context Sampling, a stochastic process to backtrack previously translated paragraphs and favor deeper context paths.
3. Path-Based Contextual Generation, which formats these context paths as the discourse information to prompt LLMs.

2.1 Directed Discourse Graph Construction

We model the source document $D = \{x_1, x_2, \ldots, x_N\}$ as a weighted directed acyclic graph $G = (V, E)$ firstly. The directed edge e_{ij} connects a target node v_i to a previous node v_j where $j < i$. Note that future translation y_i (where $i > j$) are unavailable at step j, which is consistent with a human translating the document sentence by sentence. This definition of directed edges can also avoid the appearance of cyclic graphs, thus reducing complexity during traversing.

The weight w_{ij} of edge e_{ij} quantifies the relevance of paragraph x_j to x_i, calculated by a fusion of three discourse-related factors:

$$w_{ij} = \alpha \cdot S_{sem}(i,j) + \beta \cdot S_{seq}(i,j) + \gamma \cdot S_{key}(i,j), \quad (1)$$

where α, β, and γ are coefficients, that sum to 1 to balance each factor. The specific definitions are as follows:

Semantic Relevance (S_{sem}). Global coherence relies on thematic consistency. We map each paragraph x_i into a dense vector space via a pre-trained embedding model, denoted as $\mathbf{h}_i$. Then we compute the cosine similarity between these vectors. To prevent the graph from being too dense and introducing extra noise, we introduce a threshold τ_{sem} to truncate those edges with low correlation:

$$S_{sem}(i,j) = \max(0, \frac{\mathbf{h}_i^\top \mathbf{h}_j}{|\mathbf{h}_i||\mathbf{h}_j|} - \tau_{\text{sem}}) \quad (2)$$

Sequential Adjacency (S_{seq}). The paragraph being translated is most closely related to its adjacent paragraphs. For example, in dialogue questionnaires or background introductions, adjacent contextual information is essential for reference resolution and for preserving logical coherence. Therefore, an adjacent edge is naturally introduced and assigned a fixed weight:

$$S_{seq}(i,j) = \mathbb{I}(j = i-1) \tag{3}$$

where $\mathbb{I}(\cdot)$ is the indicator function. This guarantees that the local context is always considered a candidate.

Keyword Overlap (S_{key}). Semantic-based retrieval may miss some paragraphs that contain overlapping keywords but have low semantic similarity. These omissions may lead to inconsistency in the translation of terms, such as some proper nouns. We alleviate this problem by introducing $S_{key}(i,j)$ of keyword overlap. Specifically, $\mathcal{K}_i$ denotes the set of top-K keywords in x_i extracted via TF-IDF. Then we calculate the lexical score through the degree of keyword overlap:

$$S_{key}(i,j) = \sum_{t \in \mathcal{K}_{ij}} \frac{\psi(t,x_i) + \psi(t,x_j)}{2} \tag{4}$$

where $\mathcal{K}_{ij} = \mathcal{K}_i \cap \mathcal{K}_j$ denotes the intersection of keywords, and $\psi(t,x)$ represents the TF-IDF score of term t in paragraph x.

2.2 Depth-Biased Context Path Sampling

Once the discourse graph is built, we can backtrack the context path, starting from the given target paragraph x_i. The backtrack strategy is also important. The most straightforward method is greedy search, which always selects the neighbor node with the highest weight. However, we find that this method sometimes terminates prematurely, or tends to select some nodes with certain types of edges ,such as repeatedly traversing edges with highly similar semantics, resulting in redundancy. To address this, we propose a sampling strategy to balance the edge relevance with the depth of context path to exploit global structural information.

Depth Heuristic

We introduce a concept of *Depth Heuristic* to estimate the informational richness of a context node. The Depth Heuristic $\phi(v_j)$ represents the longest path backtracked from node v_j. We argue that this backtrace depth represents how much historical translation context can be provided from a given node v_i. Specifically, we can efficiently calculate $\phi(v_j)$ via dynamic programming.

$$\phi(v_j) = 1 + \max_{v_k \in \mathcal{B}(v_j)} \phi(v_k) \tag{5}$$

where $\phi(v_{start}) = 1$ and $\mathcal{B}(v_j)$ denotes the set of backward neighbors of v_j.

Probabilistic Sampling

We apply the random walk mechanism for context selection to construct the path $\mathcal{P}_i = (v_{p_1}, v_{p_2}, \ldots, v_{p_L})$, starting from the target $v_{p_1} = v_i$. Given the current node v_{curr}, the walker transitions to a previous node v_{next}, which is sampled from its neighborhood $\mathcal{N}(v_{curr})$ defined above. The transition probability is as follows:

$$P(v_{next}|v_{curr}) = \frac{w_{curr,next} \cdot (\phi(v_{next}) + 1)^{\lambda}}{Z} \tag{6}$$

where Z is the partition function for normalization, and the term $(\phi(v_{next}) + 1)^{\lambda}$ introduces a bias towards deeper structures. The hyperparameter $\lambda \geq 0$ affects the strength of the deep bias. When $\lambda = 0$, the traversal process degrades to a standard random walk based on edge weights. Increasing λ will favor nodes with higher depth, which encourages the retrieval of long-range context.

2.3 Path-Based Contextual Generation

After completing the traversal, we can employ this sampled discourse path $\mathcal{P}_i$ for in-context learning by prompting LLMs. We reverse the path to restore the natural document order, and the final prompt I_i is constructed as follows:

$$I_i = \mathcal{I} \oplus [(x_{p_L}, y_{p_L}) \oplus \cdots \oplus (x_{p_2}, y_{p_2})] \oplus x_i \tag{7}$$

where $\mathcal{I}$ denotes the translation instruction and $\oplus$ represents string concatenation. The pair (x_k, y_k) denotes the previous source paragraph and its translation corresponding to the node v_k.

Multi-Path Sampling. Since our method is based on a random walk, each traversal can yield a different context path. We can sample K independent context paths $\{\mathcal{P}_i^{(1)}, \ldots, \mathcal{P}_i^{(K)}\}$ for the target paragraph x_i and then generate K candidate translations $\{y_i^{(1)}, \ldots, y_i^{(K)}\}$ accordingly. The final translation can be determined via a majority voting mechanism or by selecting the candidate with the lowest perplexity. In this paper, we cluster the K candidates via k-means and select the representative candidate closest to the cluster centroid, which also proves to be a simple yet effective strategy.

Complexity Analysis. The time cost of graph construction primarily lies in embedding computation and TF-IDF keyword matching, leading to an overall time complexity of $O(N^2)$. In practice, this is a one-time cost of $<$10 s per typical document (e.g., $N \approx 200$ sentences on a single CPU), and each random walk completes in milliseconds. At inference time, G^{2}C-MT requires exactly N LLM calls for N paragraphs—the same as the Window-Based baseline—since graph construction and the walk are pre-processing steps that do not invoke the LLM. Moreover, paragraphs below the similarity cutoff τ_{sem} are pruned, making the graph sparse.

3 Experiments

3.1 Experimental Setup

Datasets We evaluate our models on two standard benchmark test sets for DocMT. One of them is the SAP test set. Derived from the WAT 2020 and 2021 shared tasks, the set contains documents in the IT domain. For this test set, experiments are conducted on six translation directions: English↔Vietnamese, English↔Chinese, and English↔Indonesian. The sentences are organized

Algorithm 1 G^{2}C-MT: Graph-Guided Contextual Translation

Input: Source Document $D = \{x_1, \ldots, x_N\}$, LLM $\mathcal{M}$
Output: Translated Document Y
Stage 1: Directed Discourse Graph Construction
Initialize $G = (V, E)$ with nodes $V = \{1, \ldots, N\}$
for $i = 1$ **to** N **do**
  for $j = 1$ **to** $i - 1$ **do**
    Calc weight w_{ij} via semantic/seq/keyword scores
    if $w_{ij} > 0$ **then**
      Add edge (i, j) to E with weight w_{ij}
    end if
  end for
end for
Stage 2: Depth Heuristic Calculation
Compute $\phi(v)$ for all $v \in V$ via dynamic programming
Stage 3: Path-Based Generation
$Y \leftarrow \emptyset$
for $i = 1$ **to** N **do**
  Sample path $\mathcal{P}_{\text{back}}$ from x_i on G
  $\mathcal{P}_{\text{ctx}} \leftarrow \text{Reverse}(\mathcal{P}_{\text{back}} \setminus \{x_i\})$
  Construct prompt I_i using $\mathcal{P}_{\text{ctx}}$ and x_i
  $y_i \leftarrow \mathcal{M}(I_i)$
  Append y_i to Y
end for
Return Y

into separate documents, comprising approximately 2,000 sentences per direction. We also employ the tst2017 test set from the IWSLT 2017 translation task, comprising parallel TED talk documents. For this test set, we evaluate on the following eight translation directions: English↔Chinese, English↔French, English↔German, and English↔Japanese. Each translation direction consists of 10 to 12 sentence-aligned parallel documents, totaling approximately 1,500 sentences. In both benchmarks, each document paragraph consists of a single sentence; thus each graph node corresponds to one sentence.

Baselines We compare our model against the following baselines:

- **Sentence**: Each segment is translated independently any contextual information.
- **Window-Based**: A fixed-size sliding window of preceding sentences is used as context for translating the current sentence.
- **Semantic-Based**: Context paragraphs are selected based on embedding-based semantic similarity to the source paragraph.
- **GRAFT** [Dutta *et al.*, 2025]: A multi-agent framework for DocMT. The framework first determines the most relevant context paragraph and then extracts key alignment information, such as pronouns, entities, and phrases, to aid document-level translation.
- **DelTA** [Wang *et al.*, 2025]: An agentic framework aimed at translation consistency, featuring a multi-level memory architecture responsible for proper nouns, summaries, and variable-term contexts.

Settings For the backbone translation models, we utilize the APIs for Gemini-2.5-Flash-lite, DeepSeek-v3-0324, Qwen-2.5-72B-Instruct and Qwen3-235B-A22B-Instruct. To ensure reproducibility and deterministic outputs, we set the decoding temperature to 0 for all LLMs. For Graph Construction, we employ `text-embedding-3-small` provided by OpenAI for the semantic edge calculation (S_{sem}). The semantic similarity threshold τ_{sem} is empirically set to 0.6. The hyperparameters governing edge weight contribution are set to $\alpha = 0.2$ (semantic), $\beta = 0.3$ (sequential), and $\gamma = 0.5$ (keyword), prioritizing lexical overlap and semantic coherence based on performance on the validation set. We verify term importance using TF-IDF, where each paragraph is treated as a document to compute the IDF within the scope of the source text. For the Biased Random Walk, we set the path depth bias $\lambda = 2.0$ to encourage the selection of deeper context nodes. The maximum context path length (number of previous paragraphs included) is capped at $L = 4$ and stop the walk when the distance exceeds 100. For Window-Based and Semantic-Based baselines, we also set the context size to 4 paragraphs for a fair comparison.

3.2 Main Results

Performance on Technical Documentation

The evaluation results of all translation directions and backbones on SAP datasets are presented in Table 1. Firstly, we observe that the discourse context plays an important role in enhancing translation quality for DocMT. The **Sentence** baseline consistently underperformes all context-aware methods, trailing the simple **Window-Based** context by an average of 2–3 d-BLEU. Secondly, G^{2}C-MT consistently achieves the highest d-BLEU scores across all translation directions and backbones, validating the robustness and generality of our graph-based approach regardless of the underlying model architecture. Thirdly, we observe that the **Window-Based** approach usually outperforms the **Semantic-Based** approach in this domain. This is intuitive for technical documentation, where logical progression (e.g., step 1, step 2) often implies that the adjacent sentences is the most relevant context. However, G^{2}C-MT surpasses the **Window-Based** baseline by substantial margins—for instance, achieving a gain of **+0.9 d-BLEU** on EN→VI and **+1.2 d-BLEU** on VI→EN using Gemini-2.5-Flash-lite. This indicates that even in highly sequential documents, our *Depth-Biased Sampling* successfully retrieves necessary long-range constraints (such as terminology defined earlier in the document) that a fixed window misses, without introducing the noise associated with pure semantic retrieval. We also note the unstable performance of the **Semantic-Based** method, which sometimes even underperforms the Sentence baseline (EN → ID). This suggests that treating context as a set of independent fragments may have the negative effect of disrupting the logical flow, leading to incoherent translations.

Performance on Narrative Discourse

Table 2 shows the results on the IWSLT 2017 benchmark using Qwen-2.5-72B-Instruct. This dataset is challenging for its loose conversational structure and long-range thematic dependencies. First, G^{2}C-MT significantly outperforms

Model	Method	d-BLEU Score					
		EN → VI	EN→ZH	EN → ID	VI → EN	VI → ZH	ID → EN
Gemini-2.5-Flash-lite	Sentence	65.7	40.0	55.9	53.3	37.4	52.7
	Window-Based	66.4	44.9	57.4	56.0	43.1	56.6
	Semantic-Based	65.8	39.9	55.8	55.1	41.2	56.3
	G^{2}C-MT	**67.3**	**45.1**	**58.2**	**57.2**	**43.8**	**57.2**
DeepSeek-v3-0324	Sentence	63.7	36.3	55.0	51.4	36.9	51.4
	Window-Based	64.9	38.1	56.2	56.1	40.9	55.7
	Semantic-Based	64.5	37.7	54.6	54.7	39.8	54.2
	G^{2}C-MT	**65.5**	**39.2**	**56.9**	**56.6**	**41.8**	**56.9**
Qwen3-235B-A22B-Instruct	Sentence	63.0	39.6	52.8	52.7	37.6	52.7
	Window-Based	63.8	40.8	54.4	55.4	40.9	**55.3**
	Semantic-Based	63.2	40.9	54.3	53.3	39.8	54.9
	G^{2}C-MT	**64.3**	**42.5**	**55.0**	**55.8**	**41.7**	54.6

Table 1: d-BLEU scores on the SAP benchmark across three different LLM backbones. Bold indicates the best performance. Results averaged over 3 random walk seeds ($\sigma \leq 0.3$). Improvements of G^{2}C-MT over Window-Based are significant ($p < 0.05$, bootstrap).

Method	d-BLEU Score							
	EN→ZH	EN→FR	EN→DE	EN→ JP	ZH→EN	FR→EN	DE → EN	JP → EN
NLLB-3.3B	30.8	34.7	24.3	13.7	26.4	42.8	33.1	16.6
Google Translate	30.5	40.8	22.3	16.2	27.5	27.4	22.7	21.0
Sentence	36.5	41.4	28.0	17.6	28.3	42.4	32.8	19.3
Window-Based	**36.9**	42.0	28.7	17.6	29.4	42.7	33.8	20.4
Semantic-Based	36.6	41.7	28.4	17.8	29.1	43.1	34.3	17.9
DelTA [Wang *et al.*, 2025]	35.6	41.1	28.9	16.5	29.8	**43.6**	33.9	20.5
GRAFT [Dutta *et al.*, 2025]	36.3	**44.8**	28.5	16.1	30.1	43.1	34.3	17.9
G^{2}C-MT	**36.9**	42.0	**29.1**	**18.1**	**30.6**	43.4	**34.5**	**21.0**

Table 2: d-BLEU scores on the IWSLT 2017 test set based on Qwen-2.5-72B-Instruct for fair comparison with prior work.

the commercial and supervised baselines (NLLB-3.3B and Google Translate) across all directions, confirming the efficacy of LLMs for document-level translation when prompted with appropriate context. Second, our method proves superior to recent state-of-the-art agentic frameworks, which achieves comparable or higher d-BLEU scores (e.g., **+1.4** over GRAFT in EN→ZH and **+2.1** over DelTA in JP→EN). This result is significant given the computational efficiency of our method, as GRAFT and DelTA rely on computationally expensive, multi-step LLM calls to curate context or maintain memory modules.

Revisiting Context: Structure vs. Similarity. A notable pattern across both datasets is that the **Semantic-Based** baseline frequently fails to outperform the simple **Window-Based** approach, particularly on SAP (Table 1) and several IWSLT directions (e.g., EN→DE, JP→EN). This highlights a critical fact in document translation: local cohesion often outweighs global topical relevance. Pure semantic retrieval may break the linear narrative required for immediate syntactic dependency handling (e.g., pronoun resolution). G^{2}C-MT avoids this trade-off by rooting retrieval in the discourse structure. By explicitly modeling sequential edges (S_{seq}) alongside semantic ones, our graph traversal prioritizes the immediate history while the depth-biased sampling allows the model to trace logical threads back to earlier context.

4 Analysis

In this section, we conduct a deeper analysis to investigate: (1) whether G^{2}C-MT effectively handles discourse phenomena; (2) how the model exploits long-range dependencies; (3) the potential of multi-path sampling for robustness; and (4) the contribution of each component through ablation studies.

Model	BlonDe	d-Prism	d-Comet	d-BLEU
Sentence	44.77	-1.97	2.29	47.8
Window-Based	50.74	-2.00	2.35	51.9
Semantic-Based	49.87	-1.97	2.36	50.9
G^{2}C-MT	**51.35**	**-1.88**	**2.43**	**52.7**

Table 3: Evaluation of discourse Metric on the SAP dataset (EN → XX). Higher is better for all metrics.

4.1 Evaluation of Discourse Metric

We adopt three specialized metrics to further investigate the discourse translation performance: **BlonDe** [Jiang *et al.*, 2022], which explicitly tracks discourse phenomena such as entities, tenses, and pronouns; **d-Prism** [Thompson and Post, 2020; Vernikos *et al.*, 2022], which measures semantic consistency using probability scores; and **d-Comet** [Rei *et al.*,

2022; Vernikos *et al.*, 2022], which evaluates translation quality by considering the preceding context.

As shown in Table 3, G^{2}C-MT achieves the best performance across all metrics. **Window-Based** achieves high BlonDe scores but obtains lower d-Prism scores than **Semantic-Based**. This indicates that the **Window-Based** method can capture local connections well, while the **Semantic-Based** method is good at finding relevant topics in long-range context. G^{2}C-MT mitigates this trade-off by obtaining the highest BlonDe score (51.35) while maintaining strong semantic consistency through the combination of sequential and semantic edges.

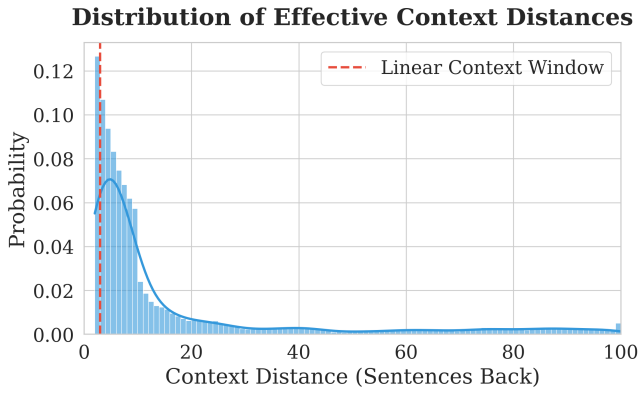

Figure 2: **Distribution of Context Distances.** The red dashed line represents the hard cutoff of standard Window-based approach (Window=4). The blue histogram shows the context selected by G^{2}C-MT. Our method retains strong attention to local context (peak at $\Delta < 10$) while maintaining a long tail of retrieval capabilities extending up to 100 sentences back, capturing long-range dependencies that linear models miss.

4.2 Properites of Graph-Guided Context Selection

To better understand the efficacy of the proposed pipeline, we characterize the context selection behavior of G^{2}C-MT compared to the **Semantic-Based** and **Window-Based** methods. We define the *Context Distance* $\Delta = i - k$ and visualize the distribution of the selected nodes in Figure 2. We can observe two key properties of G^{2}C-MT from the figure:

- **Beyond Fixed Windows:** The **Window-Based** method uses a fixed window size for context selection, where any dependency beyond the window size L is inaccessible. Our method G^{2}C-MT can walk through semantic edges (S_{sem}) to skip over intermediate nodes. We can see that in Figure 2, G^{2}C-MT effectively captures context at distances $\Delta > 20$.
- **Beyond Isolated Retrieval (Coherence over Similarity):** the **Semantic-Based** method retrieves context based on similarity scores, resulting in a *bag of sentences* $\{x_k\}$ without structural connections. However, the contexts selected by G^{2}C-MT are the components of a *connected path* $\mathcal{P}_i$. There exists a strong connection between nodes, since these edges of the path are weighted by discourse factors (Semantic, Sequential, and Lexical). The explicitly modeled sequential edges (S_{seq}) act as glue, so the distribution of context distances in G^{2}C-MT still peaks at small Δ values.

Strategy	d-BLEU
Single Path ($K = 1$)	67.3
Multi-Path ($K = 3$)	67.8
Multi-Path ($K = 5$)	**67.9**

Table 4: Effect of Multi-Path Exploration on translation quality (En→Vi).

4.3 Multi-Path Sampling for Robustness

In this section, we analyze the effect of multi-path sampling in G^{2}C-MT. As introduced in Section 2.3, we first sample multiple context paths for the same target paragraph by random walks, and then select the final translation from multiple transslation candidates. We use the embedding model to vectorize each candidate translation, and then select the candidate closest to the cluster centroid as the final output. We experimented with sampling $K = \{1, 3, 5\}$ paths on the difficult En→Vi subset, leaving larger K values for future work due to inference cost. As shown in Table 4, there is a consistent improvements in d-BLEU as we increase the number of sampled paths K. When $K = 5$, we achieve the best performance of 67.9 d-BLEU, which is +0.6 higher than the single-path baseline. This suggests that when translating ambiguous sentences, different context paths may provide complementary information, leading to more robust translations. Although this comes at the cost of increased inference latency, it offers a flexible trade-off for scenarios where quality is preferable.

Method	En → XX	Δ
G^{2}C-MT (Full)	**56.9**	-
w/o Keyword Edge (S_{key})	56.4	-0.5
w/o Semantic Edge (S_{sem})	56.7	-0.2
w/o Sequential Edge (S_{seq})	56.5	-0.4
w/o Biased Random Walk	56.4	-0.5

Table 5: Ablation study on edge components and search strategy on Gemini-2.5-Flash-lite.

4.4 Ablation Study

We conduct an ablation study on the SAP dataset to analyze the effect of different edge types and our proposed search strategies. The results are summarized in Table 5. From the table, we can observe that removing the **Keyword Edge** or the **Biased Random Walk** results in the largest performance drops (-0.5). This result is intuitive, as terminological consistency is crucial in technical documents, especially for the SAP dataset which is in the IT domain. The **Biased Random Walk** also plays an important role in preventing the model from settling for shallow, uninformative context. Another noticeable observation is that the absence of the **Sequential**

Edge (S_{seq}) causes a larger performance drop (-0.4) compared to removing the **Semantic Edge** (S_{sem}) (-0.2). This indicates that local sequential order is crucial for immediate discourse coherence (e.g., pronoun resolution), while the less impact of semantic edges may imply that the **Keyword Edge** has already captured much of the necessary topical relevance.

4.5 Case Study

Table 6 demonstrates how G^{2}C-MT handles long-range ambiguity. The source term *posting* is polysemous: it means publishing but refers to accounting entry in this specific ERP context. The defining context appears ten paragraphs earlier, causing the window-based baseline to default to the generic, incorrect translation. In contrast, G^{2}C-MT successfully retrieves the prior paragraph guided by keyword overlap edges—and generates the correct domain terminology.

Contextual Information	
Distant Context (Dist: -10 paras)	*Scenario: Automatic **Posting*** Represents a business context where... processes can trigger automatic **postings**, for example, elimination **posting**, adjustment **posting**...
Source Input (x_i)	If **jobs** have been scheduled for **posting** periods, you can change them by choosing this button.
Translations	
Reference	如果已为过账期间计划作业，则可通过选择此按钮进行更改。
Baseline *(Window-based)*	如果已为发布期间安排了任务，则可以通过选择此按钮来更改它们。 *(Fails to access distinct context, treating "posting" as "publish")*
G^{2}C-MT *(Ours)*	如果已为过账期间安排了作业，则可以通过选择此按钮来更改它们。 *(Retrieves context via graph, correct domain terminology)*

Table 6: A case study on long-range lexical disambiguation in SAP dataset.

5 Related Work

5.1 LLM-based Document-Level MT

LLMs have demonstrated strong in-context learning and long-context modeling capabilities, which makes them suitable for DocMT [Wang *et al.*, 2023; Wu *et al.*, 2024; Cui *et al.*, 2024]. One recent line of work focuses on combining graph structures with LLMs to generate discourse-aware translations since the graph can naturally model discourse dependencies. Dutta *et al.* proposed GRAFT, which uses an LLM agent to segment documents into discourse units, identify context dependencies, and form a directed acyclic discourse graph. TransGraph [Pham *et al.*, 2025] similarly models inter-chunk discourse relations via LLM-based relation classification. Our work differs from both along three axes: **(i) Graph construction cost**—GRAFT and TransGraph require $O(N^2)$ LLM calls for pairwise edge classification, whereas G^{2}C-MT constructs edges using lightweight embedding similarity and TF-IDF in a single pass; **(ii) Context depth**—prior methods restrict context selection to first-order neighbors, while our depth-biased random walk discovers multi-hop paths spanning up to 100 nodes; **(iii) Multi-path robustness**—stochastic traversal enables multi-path sampling (Table 4), a capability absent in prior graph-based DocMT. Another line of work focuses on building memory mechanisms [Wang *et al.*, 2025] to maintain document-level consistency during translation, but these approaches tend to underemphasize explicit discourse structure modeling.

5.2 Structured Reasoning and Self-Consistency in LLMs

LLMs' reasoning capabilities can be effectively invoked through sophisticated prompting techniques instead of simple instruction-response paradigms. Chain-of-Thought (CoT) prompting [Wei *et al.*, 2022] shows the potential of handling complex reasoning tasks by generating intermediate reasoning steps. Building on this work, non-linear reasoning frameworks have been proposed, such as Tree of Thoughts (ToT) [Yao *et al.*, 2023] and Graph of Thoughts (GoT) [Besta *et al.*, 2024], further improving reasoning performance by exploring multiple reasoning paths. Beyond structural exploration, single-path generation can be extended to multi-path generation to enhance robustness via self-consistency [Wang *et al.*, 2022]. This paradigm is particularly suitable for DocMT, since the translation of a given paragraph often relies on ambiguous discourse clues that may permit multiple valid interpretations. In this work, we bridge these mechanisms by integrating graph-based structural modeling with multi-path sampling for DocMT.

5.3 Retrieval-based Machine Translation

Early works in NMT have explored retrieval-based methods to improve domain adaptation and translation quality. One representative line of work [Khandelwal *et al.*, 2021; Meng *et al.*, 2022] retrieves token-level examples from a vector datastore to calibrate the model's output distribution. With the powerful in-context learning ability of LLMs, retrieving few-shot examples in sentence-level for LLM has become a dominant paradigm [Agrawal *et al.*, 2023; Ji *et al.*, 2024; Zebaze *et al.*, 2025]. [Wang *et al.*, 2025; Cui *et al.*, 2024] work on document-level translation by retrieving relevant examples but still lack explicit modeling of discourse structure and primarily consider semantic similarity, treating the document as a bag of unrelated sentences.

6 Conclusion

In this paper, we presented G^{2}C-MT, a novel framework for document-level machine translation. Our method models the document as a directed graph to capture structured discourse dependencies. By using depth-biased random walks, we select high-quality context paths as discourse context for prompting LLMs. Experiments across technical and narrative domains show that G^{2}C-MT significantly outperforms strong baselines. Furthermore, our approach is computationally efficient compared to complex agent-based methods. We believe this graph-guided strategy provides a robust solution for long-text translation tasks.

Acknowledgments

We would like to thank the anonymous reviewers for the helpful comments. This work was supported by National Nat-

ural Science Foundation of China (Grant No. 62276179, 62537001) and Project Funded by the Priority Academic Program Development of Jiangsu Higher EducationInstitutions. Xiangyu Duan is the corresponding author.

References

[Agrawal *et al.*, 2023] Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, 2023.

[Besta *et al.*, 2024] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 17682–17690, 2024.

[Chen *et al.*, 2023] Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. Longlora: Efficient fine-tuning of long-context large language models. *arXiv preprint arXiv:2309.12307*, 2023.

[Cui *et al.*, 2024] Menglong Cui, Jiangcun Du, Shaolin Zhu, and Deyi Xiong. Efficiently exploring large language models for document-level machine translation with in-context learning. *arXiv preprint arXiv:2406.07081*, 2024.

[Dutta *et al.*, 2025] Himanshu Dutta, Sunny Manchanda, Prakhar Bapat, Meva Ram Gurjar, and Pushpak Bhattacharyya. Graft: A graph-based flow-aware agentic framework for document-level machine translation. *arXiv preprint arXiv:2507.03311*, 2025.

[Gao *et al.*, 2025] Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. How to train long-context language models (effectively). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7376–7399, 2025.

[Ji *et al.*, 2024] Baijun Ji, Xiangyu Duan, Zhenyu Qiu, Tong Zhang, Junhui Li, Hao Yang, and Min Zhang. Submodular-based in-context example selection for llms-based machine translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15398–15409, 2024.

[Jiang *et al.*, 2022] Yuchen Eleanor Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 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, July 2022. Association for Computational Linguistics.

[Khandelwal *et al.*, 2021] Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Nearest neighbor machine translation. In *International Conference on Learning Representations (ICLR)*, 2021.

[Liu *et al.*, 2023] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts, 2023.

[Meng *et al.*, 2022] Yuxian Meng, Xiaoya Li, Xiayu Zheng, Fei Wu, Xiaofei Sun, Tianwei Zhang, and Jiwei Li. Fast nearest neighbor machine translation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 555–565, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[Pham *et al.*, 2025] Viet-Thanh Pham, Minghan Wang, Hao-Han Liao, and Thuy-Trang Vu. Discourse graph guided document translation with large language models. *arXiv preprint arXiv:2511.07230*, 2025.

[Rei *et al.*, 2022] 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. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[Thompson and Post, 2020] Brian Thompson and Matt Post. Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. In *Proceedings of the Fifth Conference on Machine Translation (Volume 1: Research Papers)*, Online, November 2020. Association for Computational Linguistics.

[Vernikos *et al.*, 2022] Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 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*, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[Wang *et al.*, 2022] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[Wang *et al.*, 2023] Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. Document-level machine translation with large language models. *ArXiv*, abs/2304.02210, 2023.

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

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

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

[Yao *et al.*, 2023] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

[Zebaze *et al.*, 2025] Armel Randy Zebaze, Benoît Sagot, and Rachel Bawden. In-context example selection via similarity search improves low-resource machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1222–1252, 2025.