MM-DyGraph: A Dataset, Benchmark, and Model for Multimodal Dynamic Graphs

Zongyuan Wu[1*] **Pengjie Wang**[1*] **Lingshan Chen**[2*] **Tianhang Wan**[1]
Yuan Meng[1†] **Xin Wang**[1†] **Ling Feng**[1†] **Wenwu Zhu**[1†]
[1]DCST [2]Zhili College
Tsinghua University, Beijing, China
{wuzongyu24,wpj24,cls24,wanth21}@mails.tsinghua.edu.cn
yuanmeng@mail.tsinghua.edu.cn
{xin_wang,fengling,wwzhu}@tsinghua.edu.cn

Abstract

Multimodal dynamic graphs are pervasive in real-world applications, where nodes and edges are accompanied by multimodal signals, including text and images. Despite their practical importance, benchmarks and methods that explicitly study multimodal learning in dynamic graphs remain scarce, hindering the full understanding and effective utilization of multimodal information for downstream tasks. To address this gap, we introduce **MM-DyGraph**[3] [4], to the best of our knowledge, the first benchmark for multimodal dynamic graph learning, consisting of diverse real-world datasets with fine-grained temporal graph evolution and rich multimodal attributes. **MM-DyGraph** provides three evaluation tasks to comprehensively assess model performance in dynamic settings. We benchmark representative state-of-the-art dynamic graph methods and reveal that naive feature concatenation yields *target-level inconsistencies*: multimodal inputs benefit certain prediction targets while harming others. This observation indicates that the modality informativeness depends on the specific prediction target over time. Building on these observations, we propose a **T**arget-aware **S**patial-**T**emporal **M**ultimodal **F**usion (**TST-MF**) framework, which jointly conditions multimodal fusion on the prediction target and spatial-temporal relations, enabling context-adaptive cross-modal interplay for target-specific representations. Extensive experiments over diverse datasets and tasks show that **TST-MF** delivers robust and competitive performance, achieving clear gains in multimodal settings.

1 Introduction

Dynamic graphs provide a natural abstraction for evolving relational systems, where nodes and edges evolve continuously over time. Such data are ubiquitous in high-impact applications including social platforms [1–4], recommendation and advertising systems [5–9], and finance [10–12], motivating a rich family of temporal graph neural networks to capture graph dynamics and spatiotemporal dependencies [13–19]. However, real-world dynamic graphs are rarely structure-only. Nodes and edges are often accompanied by multimodal content signals, such as product titles and images in e-commerce and advertising [20], short transaction descriptions and document-like evidence in payment ecosystems [21], and multimodal posts in social platforms that shape diffusion and engagement [22]. These multimodal attributes provide complementary semantics that play a vital role in improving

*These authors contributed equally to this research.
†Corresponding authors.
[3]Dataset is available on https://huggingface.co/datasets/LeslieWu999/MM-DyGraph
[4]Code is available on https://github.com/LeslieWu999/mm-dygraph

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Track on Evaluations and Datasets.

downstream task performance, particularly when structural connectivity is sparse, noisy, or rapidly changing [23, 24]. For example, a product's appearance and textual description may contribute differently to predicting a user's next purchase as their interests and interaction history change. Such applications call for the joint modeling of multimodal content and dynamic graph structure to capture how semantic information informs predictions in evolving relational contexts.

Despite this practical importance, benchmarks and methods that systematically study multimodal learning on dynamic graphs remain scarce. Existing benchmarks largely focus on structural and textual attributes [25–27], leaving visual signals underexplored in spatiotemporal evolution. To the best of our knowledge, no existing benchmark incorporates multimodal information for dynamic graphs, making it infeasible to develop and evaluate corresponding methods.

To address this gap, we introduce **MM-DyGraph**, a benchmark for multimodal dynamic graph learning. It consists of diverse real-world datasets with fine-grained temporal graph evolution and rich multimodal attributes and provides three evaluation tasks, *link prediction*, *edge classification*, and *node retrieval*, to comprehensively assess model performance in dynamic settings. Unlike prior benchmarks that largely restrict node and edge attributes to a single modality, **MM-DyGraph** enables systematic evaluation when multiple modalities (e.g., text and images) are simultaneously available, thereby facilitating fair comparisons and encouraging progress toward realistic multimodal dynamic graph learning.

With **MM-DyGraph**, we benchmark representative state-of-the-art dynamic graph learning methods and observe that naively concatenating multimodal features does not reliably improve performance. Furthermore, we find *target-level inconsistencies*: multimodal inputs may benefit certain prediction targets (e.g., specific nodes or edges) while harming others, indicating that modality utility is target-dependent and intertwined with spatial neighborhood and temporal history. Motivated by recent advances in context-adaptive multimodal modeling [28–31], we hypothesize that a key underlying reason is that unified fusion strategies treat multimodal signals as static, equally informative features, which can be insufficient for multimodal dynamic graphs, where the same modality may be informative for some targets but distracting for others under temporal evolution.

To address this problem, we propose **TST-MF**, a **T**arget-aware **S**patial-**T**emporal **M**ultimodal **F**usion framework for multimodal dynamic graphs. Rather than applying a fixed fusion rule, **TST-MF** performs context-adaptive cross-modal interplay by jointly conditioning fusion on (i) the prediction target, (ii) the target's spatial relations, and (iii) the target's temporal context. This design enables the model to emphasize task-relevant multimodal cues for each target while suppressing redundant or noisy signals, yielding more robust representations under spatiotemporal evolution. Extensive experiments across datasets and tasks demonstrate that **TST-MF** achieves clear gains over strong baselines, validating the effectiveness and generality of target- and context-conditioned fusion for multimodal dynamic graphs.

In summary, we summarize our contributions as follows:

1. **First Multimodal Dynamic Graph Benchmark:** To the best of our knowledge, we propose **MM-DyGraph**, the first benchmark for multimodal dynamic graph learning with diverse real-world datasets, rich multimodal attributes, and three downstream evaluation tasks. This benchmark will facilitate the development of multimodal dynamic graph learning methods and advance performance on downstream tasks.
2. **Key Observations:** We systematically benchmark representative state-of-the-art methods and reveal that naive multimodal fusion yields highly inconsistent effects across different targets. The same modality can be informative for some nodes but detrimental for others. This challenges the assumption that uniform fusion universally benefits all targets.
3. **Effective Model:** We propose **TST-MF**, a novel target-aware spatiotemporal multimodal fusion framework that strategically conditions cross-modal interplay on the prediction target and its spatiotemporal context. By seamlessly integrating multimodal information, **TST-MF** achieves substantial improvements across diverse datasets and tasks, offering an effective and principled solution for multimodal dynamic graph learning.

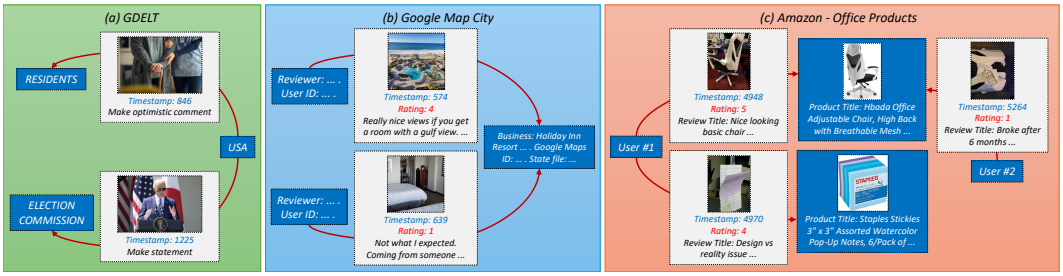


Figure 1: Visualization of **MM-DyGraph**, in which multimodal information can exist on *both* nodes and edges.

2 Related Work

2.1 Benchmarks for Dynamic Graph Representation Learning

Benchmarking has played a critical role in advancing dynamic graph learning by providing standardized datasets, evaluation protocols, and reproducible pipelines. Early efforts highlighted that evaluation for temporal link prediction is often confounded by edge recurrence and unrealistic data splits, motivating more rigorous protocols tailored to dynamic graphs [32–34]. More recently, the Temporal Graph Benchmark (TGB) [35] established a diverse suite of large-scale temporal interaction datasets spanning multiple domains, together with realistic evaluation protocols for node- and edge-level prediction tasks and an end-to-end benchmarking pipeline. To further support multi-relational temporal settings, TGB 2.0 [36] extended benchmarking to temporal knowledge graphs and heterogeneous temporal graphs, emphasizing scalability and reproducibility in large-scale multi-relational link prediction. In parallel, benchmarks have begun to incorporate richer attributes beyond structure. DTGB [27] focuses on dynamic text-attributed graphs and provides standardized evaluation under multiple real-world use cases, enabling systematic study of temporal graph methods with evolving textual attributes. While these benchmarks significantly improve the evaluation landscape for dynamic graphs, they predominantly target unimodal attributes. This limits the evaluation and model design biased toward unimodal cues, which can lead to suboptimal downstream performance in realistic settings where multimodal signals, especially visual information, are essential under temporal evolution.

2.2 Multimodal Graph Learning

Multimodal graph learning studies how to combine heterogeneous modalities (i.e., text, images, and other side information) with graph structure to improve downstream tasks [37]. Existing research in this area has primarily focused on static graphs, where multimodal node and edge attributes are fused through feature-level concatenation, modality-specific encoders followed by graph message passing, or cross-modal alignment modules that enforce semantic consistency across modalities [38]. A major line of work focuses on multimodal attributed graphs, where nodes and edges carry multimodal attributes that are encoded by modality-specific encoders and fused via concatenation, attention, or gating before or within graph propagation [39–43]. Another closely related direction enriches knowledge graphs with multimodal entity information, incorporating visual and textual signals into knowledge graph embeddings or scoring functions to improve knowledge graph completion and link prediction [40, 44–47]. Despite encouraging progress, most existing methods and benchmarks are designed for static graphs, leaving multimodal fusion under spatiotemporal evolution in dynamic graphs less explored.

3 MM-DyGraph

In this section, we present the proposed real-world datasets in **MM-DyGraph** (a visualization is presented in Figure 1). The remaining sections are structured as follows: the data sources are described in Section 3.1, the graph construction process is mentioned in Section 3.2, the evaluation tasks are given in Section 3.3, and the data statistics are introduced in Section 3.4.

3.1 Data Sources

GDELT [48] focuses on sociopolitical event extraction. It is modeled as a heterogeneous, non-bipartite and temporal knowledge graph, where nodes represent named entities, and edges denote directed sociopolitical interactions encoded via the CAMEO (Conflict and Mediation Event Observations) taxonomy. Visual information is introduced and obtained by extracting figures from the website through Beautiful Soup.

Google Map City (GMCT) [49, 50] focuses on point-of-interest recommendation, featuring spatial-temporal data mining and urban computing. It is modeled as a bipartite and highly sparse graph, containing multimodal features ranging from geo-coordinates to review text and other multimodal data. Visual information is introduced and obtained by downloading figures through links in the metadata. It provides a sparse interaction setting with multimodal content attached to edges and limited user-side attributes, complementing datasets with richer node-level features.

Amazon [51] focuses on E-commerce recommendation. It is also modeled as a bipartite and highly sparse graph, aggregating massive logs of user-item interactions metadata and user-written contents, which allow the modeling of sequential dynamics. Visual information is introduced and obtained by downloading figures through links provided by the metadata.

Yelp [52] is a dataset focusing on social network analysis and local business recommendation, whose graphs are heterogeneous and bipartite. It offers a comprehensive view of local business interactions. Visual information is originally provided by the dataset.

3.2 Dataset Construction

To ensure the quality of the graph data and the reliability of cross-modal associations, we implement a comprehensive post-processing pipeline following the initial data acquisition. This pipeline involves dataset-specific pruning to increase multimodal information density, and Vision-Language Model (VLM)-based quality assessment to filter out noisy or irrelevant figures.

The dataset-specific pruning procedure aims to accumulate sufficient cross-modal information for multimodal learning, enabling controlled evaluation without letting missing-media handling become the primary confounder. For Amazon, we exclude user nodes that have not contributed any image-containing reviews. For GMCT, we implement merchant-level pruning by removing business nodes whose edges lack visual documentation, whereas on Yelp, the pruning targets business nodes without attached images. GDELT remains unchanged.

After pruning, the datasets acquire a higher density of visual information, while also obtaining a size suitable for benchmarking. To support research under different modality-availability conditions, the published dataset includes both the pruned benchmark version and a complete version that retains all interactions within the same source scope prior to pruning. Details are given in Appendix A.1.

The VLM-based quality assessment focuses on cross-modal relevance, where we utilize Qwen3-VL-8B-Instruct [53]. Given an image-text pair, the VLM is prompted to generate a brief reasoning trace before assigning a relevance score for the pair, ranging from 0 to 10. The relevance score captures the broad semantic correspondence between the image and text, offering an accurate criterion to assess the cross-modal correlation. Manual supervision has validated the reliability of this VLM assessment (some case studies are provided in Appendix A.3). We filter out images with a relevance score lower than 6. If there still remain multiple images for a certain text after filtering, the one with the highest relevance score is retained. The prompt used in the assessment is presented in Appendix C.3.

To facilitate temporal sequence modeling, raw timestamps are transformed into discrete time steps via a temporal bucketing scheme, formed as $\tau = \left\lfloor \frac{t_{raw} - t_{min}}{\Delta t} \right\rfloor$, where Δt is the discretization granularity that is adapted based on the dataset characteristics to capture temporal dynamics inherent in different datasets. $\Delta t = 15$min is applied for datasets characterizing rapid state transitions, namely Google Map City and GDELT, while a coarser interval of $\Delta t = 24$h is utilized for datasets driven with daily periodicity and macro-level behavioral trends, namely Amazon and Yelp.

3.3 Tasks

Link Prediction evaluates models' performance on predicting whether there are links between provided node pairs. For each batch, real source-destination pairs are treated as positive examples,

while negative examples are generated by a negative edge sampler. The model computes temporal representations for both positive and negative pairs and predicts their link probabilities. Following DTGB [27], interactions are ordered chronologically and divided into training, validation, and test sets with a 70%/15%/15% split and adopt its transductive and inductive evaluation settings. Transductive evaluation measures performance on future interactions, while inductive evaluation focuses on interactions involving nodes absent from the training graph. Performance is measured by Average Precision (AP) and Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

Edge Classification evaluates models' performance on predicting the categorical label of an observed temporal edge. Given a node pair, the interaction timestamp, and the edge identifier, the model computes temporal node-pair representations and predicts an edge-label distribution with the classification head. Performance is evaluated using Precision, Recall, and F1.

Node Retrieval is a ranking evaluation built on the trained link prediction model. For each real source-destination interaction, the true destination node is treated as the ground truth, with 100 sampled ones as negative candidates. The model scores each source-destination pair with the same link probability predictor used for link prediction. Results on transductive and inductive settings are reported. Hits@k is utilized as the evaluation metric.

3.4 Statistics

We summarize the statistics of the previous and our proposed datasets in Appendix A.1. In our dataset, multimodal information is provided on both nodes and edges, where **GDELT** and **Google Map City** contain images only on edges, **Yelp** provides images only on nodes, and **Amazon** includes images on both nodes and edges. Other statistical analyses are presented in Appendix A.

4 Empirical Analysis

In this section, we first present the evaluation framework of **MM-DyGraph**, then conduct experiments to analyze the following three research questions:

(RQ1) Can existing dynamic graph methods work for multimodal situations?

(RQ2) Do additional modalities and additional information guarantee better performance?

(RQ3) Do different prediction targets benefit equally from different modalities?

4.1 Evaluation Settings

Embedding Method. Textual and visual information is encoded by a CLIP [54] model [5], which maps each piece of text and each image to a 512-dimensional vector. This model possesses excellent modality alignment capabilities that facilitate the subsequent graph learning process. Due to CLIP's intrinsic input text length limit, a chunking-and-averaging strategy is implemented to obtain the overall text embedding. We first encode the modality information and then provide the raw embeddings to the graph model. To enhance modal alignment, we normalize the raw embeddings for each modality separately.

Features. There are three feature configurations to evaluate the impact of multimodal integration: (i) random embedding baseline, (ii) unimodal (textual) embedding, and (iii) multimodal (visual-textual) representation. For the visual-textual configuration, we concatenate the text embedding and the image embedding to form a fused 1024-dimensional embedding. To ensure dimensional consistency across diverse input types, for the unimodal configuration, we duplicate the text embedding twice to match the fused feature space. In cases where the visual modality is not available, the information gap is mitigated by replacing the missing components with the corresponding textual embeddings. A comparison of alternative missing-modality imputation strategies is provided in Appendix D.

Baselines. We benchmark eight dynamic graph methods: DyGFormer [18], TGAT [3], GraphMixer [17], CAWN [15], TCL [16], TPNet [55], SALoM [56], and FreeDyG [57]. In addition, to evaluate the fusion mechanism independently of the temporal backbone, we implement three multimodal fusion strategies on the DyGFormer backbone: gated fusion, modality attention, and

[5] openai/clip-vit-base-patch16

MoE fusion [58]. Additional details regarding the evaluation framework and experimental settings are provided in Appendix C.

4.2 Results

The performance of the baseline models on the link prediction tasks is shown in Table 1 (and in Appendix D Table 7). The detailed results on the edge classification task are shown in Figure 2 (and Appendix D Table 8, Table 9, and Table 10). Performance on the node retrieval task is shown in Table 2 (and in Appendix D Table 11 and Table 12), and further assessment on a more challenging setting (historical sampling) are provided in Appendix D Table 24, Table 25, and Table 26. The bold numbers indicate the best results for each dataset. We observe that:

(RQ1) Existing dynamic graph methods work for multimodal situations. Across the experiments shown in the tables, the results suggest that naive concatenation achieves fairish results. It is often comparable to textual features and occasionally improves retrieval (e.g., DyGFormer on GDELT in Table 2). However, the effect is inconsistent, and in some cases multimodal concatenation can even substantially degrade performance (e.g., Figure 2 shows that in GDELT, naive concatenation underperforms even a random-feature baseline), motivating more principled fusion.

(RQ2) Additional modalities and additional information do not always guarantee better performance. Contrary to the intuition that additional information inherently improves representation, the results across the experiments in Table 1, Table 7, Figure 2, Table 2 and Figure 3 indicate that the effectiveness of multimodal signals is highly **dataset- and task-dependent**.

First, structural dominance is observed in certain datasets. In scenarios like Amazon Office and GDELT, strong backbones utilizing *Random embeddings* already achieve high AUC/AP, which suggests that temporal and structural signals dominate the prediction in these contexts and the performance saturation limits the contribution of additional modalities. Consequently, moving from *Random embeddings* to *Text* or *Text+Image* yields mixed outcomes. Some methods show modest gains (e.g., GraphMixer on GMCT for link prediction), whereas others show negligible improvements or even degradation (e.g., TCL on GDELT and DyGFormer on Amazon - Office).

Second, there is a distinct task sensitivity regarding multimodal semantics. Importantly, *we observe a clear task dependence even within the same dataset.* For instance, on both Amazon Office and GDELT, link prediction performance is largely saturated by structural patterns and remains insensitive to multimodal inputs. In contrast, node retrieval, which requires distinguishing specific entity properties, shows a clearer performance gap between *Random* and *Text/Text+Image* features.

Overall, these observations indicate that naively injecting visual signals is not consistently helpful and can introduce noise or misalignment, especially when the visual modality is weakly related to the prediction objective or when the backbone lacks an effective mechanism to integrate multimodal cues with temporal neighborhood evidence.

(RQ3) Modality benefits vary from different prediction targets. While multimodal integration is generally beneficial, we observe that the gains are highly *target-dependent* across **MM-DyGraph**. We identify the **structural alignment** between the *visual signal source* and the *prediction target granularity* as the governing factor.

Specifically, tasks benefit most when the visual modality resides directly on the target structure. For instance, Amazon (containing node images) exhibits strong sensitivity in node retrieval tasks (high alignment). Conversely, improvements become inconsistent or negligible when alignment is weak, such as using edge-associated images (from GDELT/GMCT) for node-centric objectives, or node-centric images (from Yelp) for link prediction.

Furthermore, misaligned modalities act as confounders rather than features. Figure 3 reveals that within the same task and dataset, adding modalities improves some instances while degrading others. This risk is amplified under *inductive settings*, where distribution shifts turn misaligned visual cues into noisy artifacts, degrading performance compared to structure-only baselines.

Table 1: Average Precision (%) for link prediction. "*tr.*" and "*in.*" refer to the transductive and inductive setting, respectively. The best results are highlighted in bold, and the second-best results are underlined.

	Datasets	Feature	TGAT	CAWN	TCL	GraphMixer	FreeDyG	TPNet	SALoM	DyGFormer	TST-MF
tr.	GDELT	Random	89.15 ± 1.99	92.68 ± 0.36	91.73 ± 0.02	90.83 ± 0.30	90.63 ± 0.33	91.89 ± 0.58	91.21 ± 1.66	92.21 ± 0.12	**92.83 ± 0.18**
		Text	89.54 ± 1.37	92.94 ± 0.11	91.67 ± 0.16	91.65 ± 0.45	91.95 ± 0.06	92.23 ± 0.19	91.64 ± 0.55	92.48 ± 0.21	**94.03 ± 0.13**
		Text+Image	90.76 ± 0.34	93.08 ± 0.25	91.47 ± 0.19	91.38 ± 0.20	91.73 ± 0.10	92.57 ± 0.36	91.89 ± 0.12	92.46 ± 0.26	**93.98 ± 0.05**
	GMCT	Random	74.28 ± 3.36	78.19 ± 0.05	78.07 ± 0.17	78.26 ± 0.05	78.28 ± 0.11	79.53 ± 0.43	62.47 ± 1.35	**94.47 ± 0.78**	80.79 ± 3.27
		Text	82.04 ± 3.67	79.49 ± 0.09	78.55 ± 0.34	79.26 ± 0.30	78.88 ± 0.29	79.81 ± 0.29	60.96 ± 1.25	**95.73 ± 0.30**	89.77 ± 0.65
		Text+Image	81.62 ± 2.33	79.65 ± 0.05	78.50 ± 0.22	79.62 ± 0.48	78.93 ± 0.19	79.83 ± 0.26	60.04 ± 3.08	**95.54 ± 0.10**	89.08 ± 0.72
	Amazon - Industrial	Random	91.22 ± 0.77	91.08 ± 0.19	91.31 ± 0.53	92.27 ± 0.23	90.55 ± 0.81	91.73 ± 0.79	82.22 ± 2.85	89.88 ± 0.73	**92.44 ± 0.19**
		Text	93.74 ± 0.13	91.94 ± 0.03	93.11 ± 0.51	**94.10 ± 0.11**	92.25 ± 0.57	92.52 ± 0.18	83.33 ± 1.97	92.45 ± 0.31	94.04 ± 0.22
		Text+Image	93.97 ± 0.18	92.21 ± 0.18	93.11 ± 0.36	93.96 ± 0.37	91.95 ± 0.40	92.91 ± 0.51	83.64 ± 1.67	92.56 ± 0.24	**94.10 ± 0.12**
	Amazon - Office	Random	90.51 ± 1.72	89.28 ± 0.56	**91.87 ± 0.30**	83.53 ± 0.08	89.04 ± 1.13	89.90 ± 1.42	87.32 ± 1.70	90.82 ± 0.76	91.27 ± 0.20
		Text	**92.11 ± 1.25**	91.06 ± 0.41	91.36 ± 0.38	90.25 ± 2.46	91.78 ± 0.27	90.72 ± 1.69	87.70 ± 1.21	90.37 ± 1.98	92.04 ± 0.12
		Text+Image	91.23 ± 0.23	91.64 ± 0.29	91.73 ± 0.19	91.06 ± 2.56	91.74 ± 0.72	89.90 ± 0.34	86.21 ± 1.74	89.00 ± 0.23	**92.67 ± 0.10**
	Yelp	Random	80.91 ± 0.16	80.54 ± 0.13	80.31 ± 0.69	80.43 ± 0.52	81.90 ± 0.07	**91.08 ± 0.06**	79.30 ± 0.97	81.83 ± 0.07	82.11 ± 0.05
		Text	92.42 ± 0.16	92.27 ± 0.19	89.71 ± 1.96	87.32 ± 0.07	90.82 ± 0.08	91.49 ± 0.10	80.98 ± 0.54	88.75 ± 5.83	**92.61 ± 0.02**
		Text+Image	92.22 ± 0.42	91.94 ± 0.13	87.44 ± 5.74	87.26 ± 0.19	90.61 ± 0.10	91.76 ± 0.20	79.02 ± 0.75	88.55 ± 5.51	**92.65 ± 0.04**
in.	GDELT	Random	76.71 ± 0.43	79.65 ± 0.35	78.37 ± 0.10	77.37 ± 1.03	77.35 ± 0.44	77.41 ± 1.67	79.29 ± 0.27	**81.70 ± 0.30**	81.14 ± 0.36
		Text	76.91 ± 1.53	80.87 ± 0.47	78.46 ± 0.23	78.84 ± 0.61	80.71 ± 0.21	78.30 ± 0.20	79.40 ± 0.43	83.07 ± 0.28	**85.35 ± 0.17**
		Text+Image	78.38 ± 0.67	81.02 ± 0.41	78.33 ± 0.29	78.58 ± 0.34	80.33 ± 0.21	79.68 ± 1.05	80.03 ± 0.58	83.05 ± 0.28	**85.44 ± 0.14**
	GMCT	Random	60.42 ± 2.40	61.08 ± 0.19	61.11 ± 0.23	61.26 ± 0.14	61.27 ± 0.39	63.32 ± 0.37	52.88 ± 0.69	**79.51 ± 1.22**	62.26 ± 0.99
		Text	67.10 ± 6.29	62.12 ± 0.14	61.27 ± 0.16	62.23 ± 0.68	61.72 ± 0.44	62.93 ± 0.19	52.28 ± 0.15	**83.46 ± 2.84**	64.90 ± 1.33
		Text+Image	66.02 ± 4.08	62.30 ± 0.18	60.90 ± 0.23	62.83 ± 0.89	61.71 ± 0.37	63.28 ± 0.45	51.99 ± 1.07	**83.05 ± 3.36**	65.49 ± 1.00
	Amazon - Industrial	Random	74.73 ± 2.86	74.30 ± 0.67	74.92 ± 2.04	78.07 ± 0.66	72.30 ± 2.66	75.81 ± 2.67	65.67 ± 2.24	74.44 ± 0.24	**78.72 ± 0.72**
		Text	81.58 ± 0.42	75.89 ± 0.11	79.60 ± 1.88	**83.15 ± 0.39**	77.02 ± 1.83	77.93 ± 0.64	65.86 ± 2.09	76.47 ± 0.69	81.67 ± 0.77
		Text+Image	82.29 ± 0.70	76.74 ± 0.86	79.70 ± 1.09	**82.75 ± 1.15**	75.90 ± 1.29	79.32 ± 1.63	66.03 ± 1.63	76.33 ± 0.58	81.95 ± 0.54
	Amazon - Office	Random	74.14 ± 4.50	70.12 ± 1.59	**77.21 ± 0.74**	57.70 ± 0.10	69.19 ± 3.44	71.87 ± 4.13	71.87 ± 1.84	75.77 ± 1.74	75.91 ± 0.32
		Text	**78.07 ± 4.00**	74.62 ± 1.36	75.78 ± 1.31	72.29 ± 7.16	76.78 ± 0.67	73.65 ± 5.06	71.21 ± 1.47	73.50 ± 6.60	77.47 ± 0.29
		Text+Image	74.85 ± 0.80	76.70 ± 0.84	76.61 ± 0.54	75.02 ± 7.04	76.60 ± 2.16	71.25 ± 0.97	71.16 ± 2.10	68.64 ± 1.02	**78.16 ± 0.31**
	Yelp	Random	77.46 ± 0.22	76.57 ± 0.37	76.93 ± 0.30	76.83 ± 0.47	78.21 ± 0.09	**86.97 ± 0.10**	72.68 ± 0.67	77.99 ± 0.09	78.33 ± 0.06
		Text	89.09 ± 0.20	88.81 ± 0.25	86.05 ± 2.08	83.43 ± 0.13	87.23 ± 0.09	87.48 ± 0.08	74.02 ± 0.74	85.29 ± 6.09	**89.32 ± 0.04**
		Text+Image	88.93 ± 0.53	88.38 ± 0.20	83.91 ± 5.83	83.45 ± 0.16	86.98 ± 0.13	87.83 ± 0.25	71.63 ± 0.46	85.05 ± 5.64	**89.39 ± 0.05**

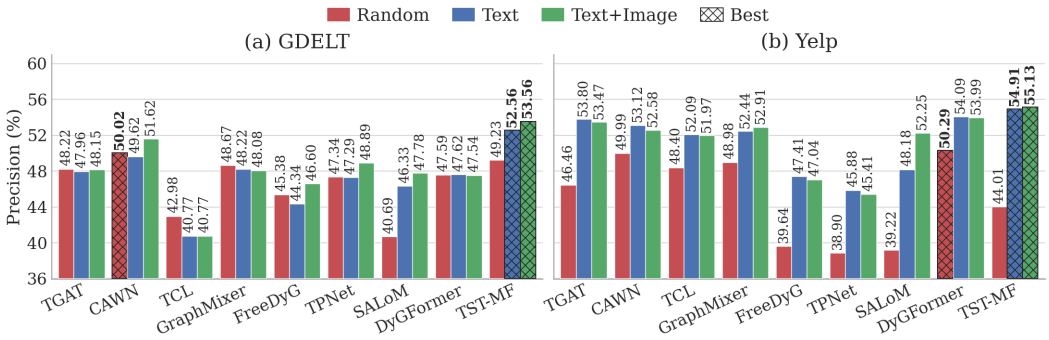


Figure 2: Precision (%) for edge classification on GDELT and Yelp. The best results are highlighted in shade.

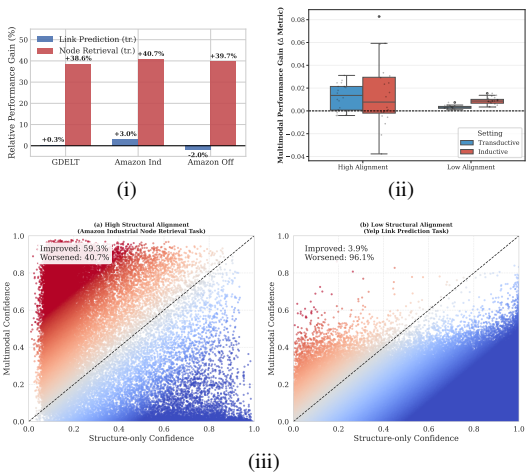


Figure 3: Misaligned modalities can behave as confounders rather than helpful features. (i) Relative performance gains from adding additional modalities over a structure-only baseline across link prediction (transductive) and node retrieval (transductive). Gains are significantly larger for node retrieval than link prediction across most datasets. (ii) Distribution of multimodal performance gains under high vs. low structural alignment, comparing transductive and inductive settings; gains are less reliable under inductive evaluation, where distribution shift can turn visual/text cues into noise. (iii) Instance-level analysis showing multimodal vs. structure-only confidence: points above the diagonal indicate improved confidence, while points below indicate degradation. Even within the same dataset and task, multimodal inputs help some instances but hurt many others. This exposes the risk of indiscriminate modality fusion, especially when task demands and provided modalities are misaligned.

Table 2: Hits@10 (%) for node retrieval. "*tr.*" and "*in.*" refer to the transductive and inductive setting, respectively. The best results are highlighted in bold, and the second-best results are underlined.

	Datasets	Feature	TGAT	CAWN	TCL	GraphMixer	FreeDyG	TPNet	SALoM	DyGFormer	TST-MF
tr.	GDELT	Random	67.77 ± 3.43	77.54 ± 0.69	75.42 ± 0.43	70.03 ± 0.99	69.38 ± 1.71	74.73 ± 1.55	73.62 ± 5.37	52.96 ± 4.82	**77.72 ± 0.36**
		Text	67.28 ± 3.85	78.81 ± 0.43	74.99 ± 0.42	72.45 ± 1.71	76.47 ± 0.66	75.57 ± 0.61	73.99 ± 2.02	69.78 ± 8.64	**81.21 ± 0.78**
		Text+Image	70.79 ± 1.04	79.08 ± 0.35	74.66 ± 0.43	72.41 ± 1.16	74.09 ± 1.77	76.21 ± 0.61	75.02 ± 0.21	71.34 ± 7.82	**81.62 ± 0.33**
	GMCT	Random	42.35 ± 5.98	50.29 ± 0.20	50.13 ± 0.30	50.22 ± 0.12	50.90 ± 0.19	50.79 ± 0.51	20.74 ± 0.06	**100.00 ± 0.00**	78.77 ± 24.87
		Text	54.01 ± 5.80	50.57 ± 0.11	50.34 ± 0.08	50.14 ± 0.41	51.21 ± 0.30	51.19 ± 0.24	20.89 ± 0.18	**100.00 ± 0.00**	**100.00 ± 0.00**
		Text+Image	53.02 ± 2.88	50.73 ± 0.02	50.27 ± 0.24	50.72 ± 0.69	51.18 ± 0.11	51.24 ± 0.23	20.69 ± 0.22	**100.00 ± 0.00**	**100.00 ± 0.00**
	Amazon - Industrial	Random	73.68 ± 0.80	73.23 ± 0.79	73.43 ± 0.91	75.47 ± 0.29	73.81 ± 0.58	74.79 ± 0.91	56.84 ± 3.80	**99.51 ± 0.79**	85.82 ± 9.03
		Text	79.40 ± 0.32	76.96 ± 0.38	79.02 ± 0.40	80.47 ± 0.09	76.78 ± 0.84	76.76 ± 0.41	58.62 ± 3.24	**100.00 ± 0.01**	**100.00 ± 0.00**
		Text+Image	79.92 ± 0.13	77.07 ± 0.57	78.61 ± 1.00	79.72 ± 1.43	76.29 ± 0.44	77.31 ± 0.74	61.84 ± 3.00	**100.00 ± 0.00**	**100.00 ± 0.00**
	Amazon - Office	Random	69.70 ± 3.29	67.33 ± 2.40	72.87 ± 0.95	66.83 ± 0.08	69.23 ± 1.05	70.64 ± 1.43	65.74 ± 2.27	**99.78 ± 0.15**	74.64 ± 3.84
		Text	74.33 ± 2.43	72.25 ± 1.16	71.68 ± 1.17	72.51 ± 3.38	73.68 ± 0.54	72.02 ± 2.07	67.13 ± 1.86	99.49 ± 0.63	**99.71 ± 0.34**
		Text+Image	73.38 ± 0.47	72.72 ± 1.24	73.01 ± 0.94	73.42 ± 3.39	73.80 ± 1.22	70.30 ± 0.48	64.35 ± 2.94	**99.99 ± 0.01**	**99.99 ± 0.01**
	Yelp	Random	48.18 ± 0.35	46.80 ± 0.19	46.91 ± 1.38	46.45 ± 1.52	49.86 ± 0.29	**71.57 ± 0.16**	52.60 ± 1.29	25.99 ± 14.39	50.04 ± 0.38
		Text	74.59 ± 0.38	73.91 ± 0.50	67.97 ± 4.75	61.89 ± 0.28	70.87 ± 0.26	72.22 ± 0.06	55.17 ± 1.61	67.41 ± 16.19	**75.19 ± 1.51**
		Text+Image	74.19 ± 1.22	72.94 ± 0.44	63.01 ± 13.26	61.72 ± 0.29	70.14 ± 0.36	72.80 ± 0.39	50.75 ± 1.58	67.54 ± 14.55	**79.11 ± 0.49**
in.	GDELT	Random	44.74 ± 1.32	47.36 ± 0.85	48.49 ± 0.18	44.83 ± 4.32	45.97 ± 0.73	43.84 ± 3.19	49.30 ± 0.62	33.75 ± 7.39	**50.38 ± 0.89**
		Text	45.41 ± 3.31	49.50 ± 0.94	48.48 ± 0.35	48.54 ± 1.07	51.95 ± 0.92	46.61 ± 0.61	48.87 ± 1.42	51.41 ± 5.13	**58.87 ± 0.63**
		Text+Image	46.86 ± 1.00	49.69 ± 0.94	48.55 ± 0.27	47.72 ± 1.15	51.29 ± 0.38	47.48 ± 1.80	50.18 ± 0.61	52.50 ± 3.86	**59.23 ± 0.58**
	GMCT	Random	20.64 ± 2.60	22.50 ± 0.17	22.64 ± 0.18	23.09 ± 0.24	41.63 ± 0.89	24.07 ± 0.23	12.59 ± 0.09	**100.00 ± 0.00**	67.55 ± 27.64
		Text	29.17 ± 7.33	22.81 ± 0.26	22.63 ± 0.39	21.66 ± 1.49	41.06 ± 0.34	23.69 ± 0.21	12.55 ± 0.03	**100.00 ± 0.00**	**100.00 ± 0.00**
		Text+Image	27.40 ± 4.31	23.23 ± 0.12	22.75 ± 0.33	23.15 ± 0.75	40.95 ± 0.18	24.42 ± 0.12	12.64 ± 0.14	**100.00 ± 0.01**	**100.00 ± 0.00**
	Amazon - Industrial	Random	41.04 ± 5.86	41.28 ± 0.93	42.23 ± 6.18	47.93 ± 0.77	45.23 ± 4.40	40.94 ± 5.67	28.05 ± 4.80	**78.07 ± 30.19**	58.66 ± 6.40
		Text	51.76 ± 1.26	39.12 ± 0.04	48.24 ± 3.21	54.64 ± 0.58	52.68 ± 4.78	45.41 ± 1.87	27.60 ± 1.93	**99.23 ± 0.19**	97.79 ± 0.35
		Text+Image	53.98 ± 1.23	41.28 ± 1.89	48.54 ± 2.34	54.10 ± 1.97	50.49 ± 1.85	48.08 ± 4.19	29.58 ± 4.71	99.51 ± 0.19	**99.94 ± 0.99**
	Amazon - Office	Random	40.20 ± 11.05	31.02 ± 2.38	48.22 ± 1.86	15.08 ± 1.54	36.86 ± 5.76	34.18 ± 8.18	37.27 ± 3.74	**83.63 ± 11.06**	46.55 ± 1.70
		Text	45.50 ± 9.16	38.60 ± 3.86	43.91 ± 3.72	34.02 ± 15.24	52.76 ± 1.60	37.67 ± 10.66	37.31 ± 1.80	**96.83 ± 0.32**	95.65 ± 8.96
		Text+Image	38.04 ± 1.99	43.72 ± 2.65	46.78 ± 1.55	40.61 ± 11.80	52.58 ± 5.56	31.96 ± 3.12	36.86 ± 3.21	98.10 ± 0.61	**98.99 ± 0.31**
	Yelp	Random	43.14 ± 0.26	41.60 ± 0.36	42.30 ± 0.61	42.00 ± 1.02	44.58 ± 0.13	**61.85 ± 0.10**	41.13 ± 0.91	26.12 ± 9.72	44.77 ± 0.31
		Text	65.50 ± 0.46	64.53 ± 0.61	58.46 ± 4.39	53.68 ± 0.25	61.75 ± 0.23	62.55 ± 0.22	42.92 ± 1.67	59.47 ± 14.07	**66.84 ± 0.38**
		Text+Image	65.24 ± 1.17	63.49 ± 0.58	54.84 ± 10.96	53.71 ± 0.20	61.22 ± 0.26	63.28 ± 0.30	38.81 ± 0.86	59.48 ± 12.06	**67.84 ± 0.48**

Table 3: Comparison of multimodal fusion strategies under the Text+Image setting (%). All methods share the DyGFormer backbone, and DyGFormer denotes naive feature concatenation. "*tr.*" and "*in.*" refer to the transductive and inductive setting, respectively. The best results are highlighted in bold, and the second-best results are underlined.

Task	Metric		Datasets	DyGFormer	Gated fusion	Modality attention	MoE fusion	TST-MF
LP	AP	*tr.*	GDELT	92.46 ± 0.26	92.74 ± 0.28	93.08 ± 0.10	92.92 ± 0.26	**93.98 ± 0.05**
			GMCT	95.54 ± 0.10	96.41 ± 0.17	96.30 ± 0.20	**96.59 ± 0.16**	89.08 ± 0.72
			Amazon - Industrial	92.56 ± 0.24	92.63 ± 0.26	92.61 ± 0.61	90.89 ± 2.25	**94.10 ± 0.12**
			Amazon - Office	89.00 ± 0.23	90.26 ± 1.41	90.73 ± 1.02	89.59 ± 0.94	**92.67 ± 0.10**
			Yelp	88.55 ± 5.51	91.25 ± 0.38	92.29 ± 0.17	88.56 ± 5.75	**92.65 ± 0.04**
		in.	GDELT	83.05 ± 0.28	83.29 ± 0.34	84.07 ± 0.31	84.14 ± 0.35	**85.44 ± 0.14**
			GMCT	**83.05 ± 3.36**	82.46 ± 1.83	81.92 ± 2.71	82.28 ± 1.28	65.49 ± 1.00
			Amazon - Industrial	76.33 ± 0.58	77.35 ± 0.83	76.44 ± 2.17	74.98 ± 1.07	**81.95 ± 0.54**
			Amazon - Office	68.64 ± 1.02	72.38 ± 4.30	74.01 ± 3.95	70.59 ± 3.07	**78.16 ± 0.31**
			Yelp	85.05 ± 5.64	87.73 ± 0.35	88.96 ± 0.21	85.00 ± 5.95	**89.39 ± 0.05**
NR	Hits@10	*tr.*	GDELT	71.34 ± 7.82	62.18 ± 6.57	74.46 ± 1.58	71.33 ± 2.94	**81.62 ± 0.33**
			GMCT	**100.00 ± 0.00**	**100.00 ± 0.00**	**100.00 ± 0.00**	99.96 ± 0.07	**100.00 ± 0.00**
			Amazon - Industrial	**100.00 ± 0.00**	**100.00 ± 0.00**	**100.00 ± 0.01**	99.66 ± 0.58	**100.00 ± 0.00**
			Amazon - Office	**99.99 ± 0.01**	99.25 ± 1.27	99.98 ± 0.03	98.45 ± 1.52	**99.99 ± 0.01**
			Yelp	67.54 ± 14.55	75.02 ± 1.72	57.09 ± 35.30	66.89 ± 17.38	**79.11 ± 0.49**
		in.	GDELT	52.50 ± 3.86	46.75 ± 7.15	54.62 ± 0.68	54.57 ± 1.51	**59.23 ± 0.58**
			GMCT	**100.00 ± 0.01**	**100.00 ± 0.00**	**100.00 ± 0.00**	99.88 ± 0.21	**100.00 ± 0.00**
			Amazon - Industrial	99.51 ± 0.19	98.25 ± 1.59	98.65 ± 0.86	96.51 ± 4.64	**99.94 ± 0.99**
			Amazon - Office	98.10 ± 0.61	96.58 ± 1.56	96.36 ± 1.14	95.89 ± 2.37	**98.99 ± 0.31**
			Yelp	59.48 ± 12.06	65.29 ± 1.92	55.40 ± 22.19	58.69 ± 15.22	**67.84 ± 0.48**
EC	Precision		GDELT	47.54 ± 0.10	48.49 ± 0.71	48.85 ± 0.89	47.74 ± 0.25	**53.56 ± 1.70**
			GMCT	56.77 ± 1.72	**57.69 ± 6.60**	56.36 ± 3.38	54.81 ± 2.50	53.45 ± 1.57
			Amazon - Industrial	**59.38 ± 1.85**	52.81 ± 0.68	52.53 ± 0.57	53.50 ± 0.20	50.32 ± 1.86
			Amazon - Office	**56.02 ± 0.50**	53.98 ± 0.86	53.73 ± 0.12	53.91 ± 0.44	54.42 ± 0.02
			Yelp	53.99 ± 0.18	52.65 ± 0.30	53.91 ± 0.49	52.71 ± 1.03	**55.13 ± 0.87**

5 Target-aware Spatial-Temporal Modality Fusion

5.1 Observation and Motivation

The main results reveal that the effectiveness of multimodal input in dynamic graphs is neither universal nor monotonic. Although multimodal attributes can provide complementary semantics, their utility is not globally fixed and instead varies with the evolving *spatiotemporal context*. In several situations, competitive performance can already be achieved with minimal or even random features, suggesting that temporal structural patterns dominate the signal and naively adding modalities may introduce noise. In contrast, on other datasets and objectives, informative attributes lead to noticeable gains, yet the improvements remain inconsistent across backbones, indicating that simple concatenation fails to

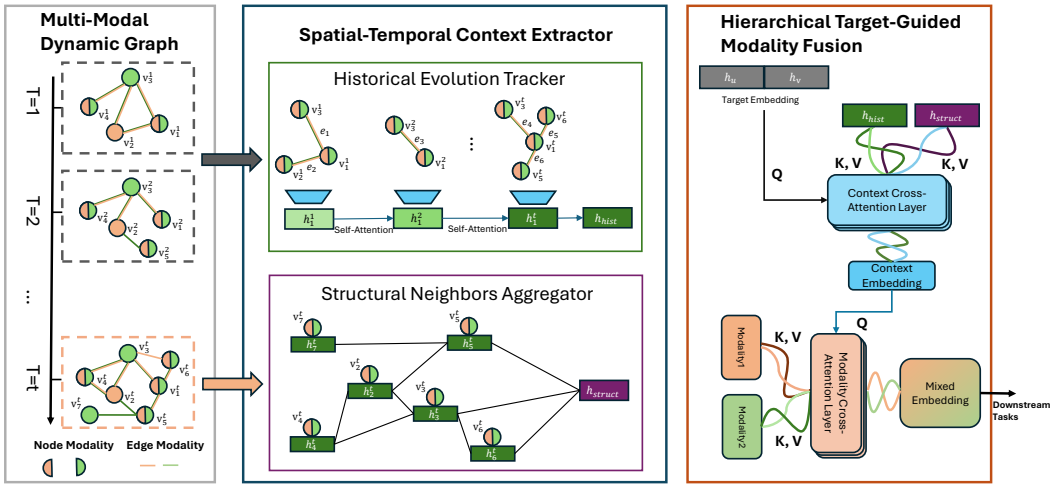


Figure 4: Overview of the proposed TST-MF architecture. Given a prediction query, the Spatial-Temporal Context Extractor encodes (i) the interaction history with a sequential Transformer to obtain a history embedding, and (ii) the local neighborhood with a graph Transformer to obtain a structural embedding. The Hierarchical Target-Guided Modality Fusion module then performs two-stage cross-attention to perform a context-adaptive fused representation for downstream prediction.

capture *when* and *where* each modality becomes useful. These observations motivate a *context-aware fusion* mechanism that conditions multimodal integration on the neighborhood (spatial context) and history (temporal context), rather than applying a static fusion rule.

Beyond dataset- and task-level variation, we further observe strong *target-dependent* heterogeneity. In dynamic graphs, the prediction target, the particular nodes being queried, determines what evidence is relevant. The same modality may be informative for some targets but distracting for others, even within the same dataset and task. Therefore, effective multimodal learning calls for *target-aware fusion*, where cross-modal interplay is guided by the target representation, enabling adaptive modality selection and combination for each prediction instance.

Motivated by these findings, we propose **TST-MF**, a target-aware Spatial-Temporal Multimodal Fusion framework for dynamic graphs.

5.2 Model Architecture

As shown in Figure 4, our framework consists of two main components: a *Spatial-Temporal Context Extractor* and a *Hierarchical Target-Guided Modality Fusion* module. Given a prediction query regarding a node group $\mathcal{A}$, which is the target, on a dynamic graph $\mathcal{G}_{<t}$, we first extract its temporal and structural context, and then perform target-guided multimodal fusion to get a task-conditioned representation $\mathbf{z}_r(\mathcal{A}, t)$ for $r \in \mathcal{A}$.

Spatial-Temporal Context Extractor. This component produces two complementary context representations for each query. (i) A *historical evolution tracker* encodes the target's interaction history as a temporal sequence and captures long-range dependencies using a sequential transformer, yielding a history embedding $H_r = H^{\text{his}}(\mathcal{H}_r(t))$ where $\mathcal{H}_r(t)$ is the history sequence. (ii) A *structural neighbor aggregator* summarizes the target's local neighborhood via a structural attention aggregator, yielding a structural embedding $S_r = S^{\text{struct}}(\mathcal{N}_r(t))$ where $\mathcal{N}_r(t)$ is the node components in $\mathcal{H}_r(t)$. Together, these embeddings characterize the nodes from both temporal evolution and spatial structure.

Hierarchical Target-Guided Modality Fusion. We fuse modalities in a hierarchical, target-conditioned manner. First, we construct a *target embedding* $q = f_{\text{task}}(\mathcal{A})$ from the raw embeddings of the queried entities. Using this target embedding as a query, we perform cross-attention over the history and structural embeddings to obtain a unified *context embedding* $c_r = g^{\text{context}}(q; H_r, S_r)$, which captures not only spatiotemporal context but also target-specific intent. Second, the context embedding is used to query modality-specific features (e.g., text and image representations) through another cross-attention stage, producing context-adaptive fused representation $m_r = h^{\text{modal}}(c_r; \mathcal{A})$.

This two-stage design enables the model to (1) adaptively retrieve relevant spatiotemporal evidence for the specific prediction target and (2) selectively integrate modalities conditioned on that evidence, rather than relying on naive concatenation.

The pseudocode for the proposed method is provided in Appendix B.

5.3 Experiment Results

To verify the efficacy of our proposed **TST-MF**, we conduct experiments on the aforementioned **MM-DyGraph**. The main results are reported in Table 1, Figure 2, and Table 2. Overall, **TST-MF** achieves strong performance across datasets and tasks, frequently obtaining the best or second-best results among strong dynamic graph baselines. This demonstrates that conditioning multimodal fusion on both the prediction target and its spatiotemporal context yields more robust representations than target-agnostic or naive fusion strategies.

To isolate the contribution of the fusion mechanism from that of the temporal backbone, we further compare **TST-MF** with three representative multimodal fusion strategies, namely gated fusion, modality attention, and the mixture-of-experts fusion [58], all built on the same DyGFormer backbone with comparable hidden dimensions. Results are shown in Table 3 and Appendix D Table 13. **TST-MF** outperforms all three strategies in most datasets and tasks, suggesting that target-guided fusion is more effective than target-agnostic alternatives.

5.4 Ablation Study

We conduct ablation studies to disentangle the contribution of each design choice in **TST-MF**. Specifically, we consider four variants that remove (i) the target guidance in the context cross-attention layer, (ii) the hierarchical architecture in the target-guided modality fusion, and either (iii) the spatial or (iv) historical context branch in the spatiotemporal context extractor. Furthermore, to test whether a modality is equally useful for every prediction, we mask it at test time and measure the change in each instance's loss. Due to page limit, the results and detailed analysis are provided in Appendix E.

6 Conclusion

In this work, we study multimodal learning on dynamic graphs and find that leveraging heterogeneous signals (e.g., text and images) under temporal evolution remains underexplored. To facilitate systematic evaluation, we introduce **MM-DyGraph**, a benchmark comprising diverse real-world multimodal dynamic graph datasets with fine-grained temporal graph evolution, and providing three representative tasks covering link prediction, edge classification, and node retrieval. Benchmarking representative state-of-the-art dynamic graph methods on MM-DyGraph reveals that naively incorporating multimodal inputs does not consistently improve performance and can even be detrimental, indicating that existing models underutilize cross-modal complementarity. Motivated by these findings, we propose **TST-MF**, a target-aware spatiotemporal multimodal fusion framework that conditions fusion on the queried prediction target together with its neighborhood and historical context, enabling adaptive modality integration over time. Extensive experiments demonstrate that **TST-MF** outperforms strong baselines across datasets and tasks, validating the effectiveness and generality of our approach and establishing a solid foundation for future research on multimodal dynamic graph learning.

Acknowledgements

This work was supported by Beijing National Research Center for Information Science and Technology under Grant No.BNR2026TD03005 and National Natural Science Foundation of China under Grant No.62402264.

References

[1] Rakshit Trivedi, Mehrdad Farajtabar, Prasenjeet Biswal, and Hongyuan Zha. Dyrep: Learning representations over dynamic graphs. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=HyePrhR5KX.

[2] Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael Bronstein. Temporal graph networks for deep learning on dynamic graphs, 2020. URL https://arxiv.org/abs/2006.10637.

[3] Da Xu, Chuanwei Ruan, Evren Körpeoglu, Sushant Kumar, and Kannan Achan. Inductive representation learning on temporal graphs. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=rJeW1yHYwH.

[4] Srijan Kumar, Xikun Zhang, and Jure Leskovec. Predicting dynamic embedding trajectory in temporal interaction networks. In Ankur Teredesai, Vipin Kumar, Ying Li, Rómer Rosales, Evimaria Terzi, and George Karypis, editors, *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 1269–1278. ACM, 2019. doi: 10.1145/3292500.3330895. URL https://doi.org/10.1145/3292500.3330895.

[5] Zeyang Zhang, Xin Wang, Haibo Chen, Haoyang Li, and Wenwu Zhu. Disentangled dynamic graph attention network for out-of-distribution sequential recommendation. *ACM Trans. Inf. Syst.*, 43(1), December 2024. ISSN 1046-8188. doi: 10.1145/3701988. URL https://doi.org/10.1145/3701988.

[6] Mengqi Zhang, Shu Wu, Xueli Yu, Qiang Liu, and Liang Wang. Dynamic graph neural networks for sequential recommendation. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2022. doi: 10.1109/TKDE.2022.3151618.

[7] Panyu Zhai, Yanwu Yang, and Chunjie Zhang. Periodic graph neural networks for click-through rate prediction in online advertising. *ACM Trans. Inf. Syst.*, 44(1), 2025. ISSN 1046-8188. doi: 10.1145/3769103. URL https://doi.org/10.1145/3769103.

[8] Farzaneh Rajabi and Jack Siyuan He. Click-through rate prediction using graph neural networks and online learning, 2021. URL https://arxiv.org/abs/2105.03811.

[9] Chunjing Xiao, Wanlin Ji, Yuxiang Zhang, and Shenkai Lv. Pidkg: Propagating interaction influence on the dynamic knowledge graph for recommendation. *ACM Trans. Web*, 18(2), 2024. ISSN 1559-1131. doi: 10.1145/3593314. URL https://doi.org/10.1145/3593314.

[10] Omkar Kulkarni and Rohitash Chandra. Dynberg: Dynamic bert-based graph neural network for financial fraud detection, 2025. URL https://arxiv.org/abs/2511.00047.

[11] Yifan Duan, Guibin Zhang, Shilong Wang, Xiaojiang Peng, Wang Ziqi, Junyuan Mao, Hao Wu, Xinke Jiang, and Kun Wang. Cat-gnn: Enhancing credit card fraud detection via causal temporal graph neural networks, 2024. URL https://arxiv.org/abs/2402.14708.

[12] Yejin Kim, Youngbin Lee, Minyoung Choe, Sungju Oh, and Yongjae Lee. Temporal graph networks for graph anomaly detection in financial networks, 2024. URL https://arxiv.org/abs/2404.00060.

[13] Zeyang Zhang, Xin Wang, Ziwei Zhang, Haoyang Li, Zhou Qin, and Wenwu Zhu. Dynamic graph neural networks under spatio-temporal distribution shift. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 6074–6089. Curran Associates, Inc., 2022. doi: 10.52202/068431-0440. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/2857242c9e97de339ce642e75b15ff24-Paper-Conference.pdf.

[14] Zeyang Zhang, Xin Wang, Ziwei Zhang, Zhou Qin, Weigao Wen, Hui Xue, Haoyang Li, and Wenwu Zhu. Spectral invariant learning for dynamic graphs under distribution shifts, 2024. URL https://arxiv.org/abs/2403.05026.

[15] Yanbang Wang, Yen-Yu Chang, Yunyu Liu, Jure Leskovec, and Pan Li. Inductive representation learning in temporal networks via causal anonymous walks. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=KYPz4YsCPj.

[16] Lu Wang, Xiaofu Chang, Shuang Li, Yunfei Chu, Hui Li, Wei Zhang, Xiaofeng He, Le Song, Jingren Zhou, and Hongxia Yang. Tcl: Transformer-based dynamic graph modelling via contrastive learning, 2021. URL https://arxiv.org/abs/2105.07944.

[17] Weilin Cong, Si Zhang, Jian Kang, Baichuan Yuan, Hao Wu, Xin Zhou, Hanghang Tong, and Mehrdad Mahdavi. Do we really need complicated model architectures for temporal networks? In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=ayPPc0SyLv1.

[18] Le Yu, Leilei Sun, Bowen Du, and Weifeng Lv. Towards better dynamic graph learning: New architecture and unified library. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/d611019afba70d547bd595e8a4158f55-Abstract-Conference.html.

[19] Zeyang Zhang, Ziwei Zhang, Xin Wang, Yijian Qin, Zhou Qin, and Wenwu Zhu. Dynamic heterogeneous graph attention neural architecture search. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):11307–11315, Jun. 2023. doi: 10.1609/aaai.v37i9.26338. URL https://ojs.aaai.org/index.php/AAAI/article/view/26338.

[20] Xiang-Rong Sheng, Feifan Yang, Litong Gong, Biao Wang, Zhangming Chan, Yujing Zhang, Yueyao Cheng, Yong-Nan Zhu, Tiezheng Ge, Han Zhu, Yuning Jiang, Jian Xu, and Bo Zheng. Enhancing taobao display advertising with multimodal representations: Challenges, approaches and insights, 2024. URL https://arxiv.org/abs/2407.19467.

[21] Berke Oral and Gülşen Eryiğit. Fusion of visual representations for multimodal information extraction from unstructured transactional documents. 25(3):187–205, 2022. ISSN 1433-2833. doi: 10.1007/s10032-022-00399-3. URL https://doi.org/10.1007/s10032-022-00399-3.

[22] Wanqiu Cui, Junping Du, Dawei Wang, Feifei Kou, and Zhe Xue. Mvgan: Multi-view graph attention network for social event detection. *ACM Trans. Intell. Syst. Technol.*, 12(3), 2021. ISSN 2157-6904. doi: 10.1145/3447270. URL https://doi.org/10.1145/3447270.

[23] Xin Wang, Hong Chen, and Wenwu Zhu. Multimodal disentangled representation for recommendation. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2021. doi: 10.1109/ICME51207.2021.9428193.

[24] Yanbin Wei, Shuai Fu, Weisen Jiang, Zejian Zhang, Zhixiong Zeng, Qi Wu, James T. Kwok, and Yu Zhang. GITA: graph to visual and textual integration for vision-language graph reasoning. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/00295cede6e1600d344b5cd6d9fd4640-Abstract-Conference.html.

[25] Bowen Jin, Gang Liu, Chi Han, Meng Jiang, Heng Ji, and Jiawei Han. Large language models on graphs: A comprehensive survey, 2023. URL https://arxiv.org/abs/2312.02783.

[26] Hao Yan, Chaozhuo Li, Ruosong Long, Chao Yan, Jianan Zhao, Wenwen Zhuang, Jun Yin, Peiyan Zhang, Weihao Han, Hao Sun, Weiwei Deng, Qi Zhang, Lichao Sun, Xing Xie, and Senzhang Wang. A comprehensive study on text-attributed graphs: Benchmarking and rethinking. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on*

Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/37d00f567a18b478065f1a91b95622a0-Abstract-Datasets_and_Benchmarks.html.

[27] Jiasheng Zhang, Jialin Chen, Menglin Yang, Aosong Feng, Shuang Liang, Jie Shao, and Rex Ying. DTGB: A comprehensive benchmark for dynamic text-attributed graphs. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/a65d054a407f94c34ecfb598fb540a0d-Abstract-Datasets_and_Benchmarks_Track.html.

[28] Yufei Guo, Yinan Wu, Haoran Duan, Guiguang Ding, and Jungong Han. Phaselora: Control-regime-conditioned low-rank adaptation for continuous-action vision-language-action policies. *arXiv preprint arXiv:2608.15285*, 2026.

[29] Kang He, Yuzhe Ding, Xinrong Wang, Fei Li, Chong Teng, and Donghong Ji. Enhance-then-balance modality collaboration for robust multimodal sentiment analysis, 2026. URL https://arxiv.org/abs/2604.12518.

[30] Jason Wu, Yuyang Yuan, Kang Yang, Lance Kaplan, and Mani Srivastava. Admn: A layer-wise adaptive multimodal network for dynamic input noise and compute resources, 2025. URL https://arxiv.org/abs/2502.07862.

[31] Konyul Park, Yecheol Kim, Daehun Kim, and Jun Won Choi. Resilient sensor fusion under adverse sensor failures via multi-modal expert fusion, 2025. URL https://arxiv.org/abs/2503.19776.

[32] Farimah Poursafaei, Shenyang Huang, Kellin Pelrine, and Reihaneh Rabbany. Towards better evaluation for dynamic link prediction. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/d49042a5d49818711c401d34172f9900-Abstract-Datasets_and_Benchmarks.html.

[33] Farimah Poursafaei, Shenyang Huang, Kellin Pelrine, and Reihaneh Rabbany. Towards better evaluation for dynamic link prediction. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/d49042a5d49818711c401d34172f9900-Abstract-Datasets_and_Benchmarks.html.

[34] Tao Zou, Yuhao Mao, Junchen Ye, and Bowen Du. Repeat-aware neighbor sampling for dynamic graph learning. In Ricardo Baeza-Yates and Francesco Bonchi, editors, *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pages 4722–4733. ACM, 2024. doi: 10.1145/3637528.3672001. URL https://doi.org/10.1145/3637528.3672001.

[35] Shenyang Huang, Farimah Poursafaei, Jacob Danovitch, Matthias Fey, Weihua Hu, Emanuele Rossi, Jure Leskovec, Michael M. Bronstein, Guillaume Rabusseau, and Reihaneh Rabbany. Temporal graph benchmark for machine learning on temporal graphs. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/066b98e63313162f6562b35962671288-Abstract-Datasets_and_Benchmarks.html.

[36] Julia Gastinger, Shenyang Huang, Michael Galkin, Erfan Loghmani, Ali Parviz, Farimah Poursafaei, Jacob Danovitch, Emanuele Rossi, Ioannis Koutis, Heiner Stuckenschmidt,

Reihaneh Rabbany, and Guillaume Rabusseau. TGB 2.0: A benchmark for learning on temporal knowledge graphs and heterogeneous graphs. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/fda026cf2423a01fcbcf1e1e43ee9a50-Abstract-Datasets_and_Benchmarks_Track.html.

[37] Xin Wang, Zeyang Zhang, Linxin Xiao, Haibo Chen, Chendi Ge, and Wenwu Zhu. Towards multimodal graph large language model, 2025. URL https://arxiv.org/abs/2506.09738.

[38] Ciyuan Peng, Jiayuan He, and Feng Xia. Learning on multimodal graphs: A survey, 2024. URL https://arxiv.org/abs/2402.05322.

[39] Hao Yan, Chaozhuo Li, Jun Yin, Zhigang Yu, Weihao Han, Mingzheng Li, Zhengxin Zeng, Hao Sun, and Senzhang Wang. When graph meets multimodal: Benchmarking and meditating on multimodal attributed graphs learning, 2024. URL https://arxiv.org/abs/2410.09132.

[40] Jing Zhu, Yuhang Zhou, Shengyi Qian, Zhongmou He, Tong Zhao, Neil Shah, and Danai Koutra. Mosaic of modalities: A comprehensive benchmark for multimodal graph learning, 2024. URL https://arxiv.org/abs/2406.16321.

[41] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1437–1445, 2019.

[42] Jianing Li, Chaoqun Yang, Guanhua Ye, and Quoc Viet Hung Nguyen. Graph neural networks with deep mutual learning for designing multi-modal recommendation systems. *Information Sciences*, 654:119815, 2024. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2023.119815. URL https://www.sciencedirect.com/science/article/pii/S0020025523014007.

[43] Jie Cai, Xin Wang, Haoyang Li, Ziwei Zhang, and Wenwu Zhu. Multimodal graph neural architecture search under distribution shifts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(8):8227–8235, Mar. 2024. doi: 10.1609/aaai.v38i8.28663. URL https://ojs.aaai.org/index.php/AAAI/article/view/28663.

[44] Xin Wang, Benyuan Meng, Hong Chen, Yuan Meng, Ke Lv, and Wenwu Zhu. Tiva-kg: A multimodal knowledge graph with text, image, video and audio. 2023.

[45] Enqiang Wang, Qing Yu, Yelin Chen, Wushouer Slamu, and Xukang Luo. Multi-modal knowledge graphs representation learning via multi-headed self-attention. *Information Fusion*, 88:78–85, 2022. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2022.07.008. URL https://www.sciencedirect.com/science/article/pii/S1566253522000689.

[46] Yuxing Lu, Weichen Zhao, Nan Sun, and Jinzhuo Wang. Enhancing multimodal knowledge graph representation learning through triple contrastive learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pages 5963–5971. ijcai.org, 2024. URL https://www.ijcai.org/proceedings/2024/659.

[47] Ban Tran and Thanh Le. Mesn: A multimodal knowledge graph embedding framework with expert fusion and relational attention. *Knowledge-Based Systems*, 318:113541, 2025. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2025.113541. URL https://www.sciencedirect.com/science/article/pii/S0950705125005878.

[48] Kalev Leetaru and Philip Schrodt. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA Annual Convention*, San Diego, CA, 2013. URL https://www.gdeltproject.org/. Data available at https://www.gdeltproject.org/.

[49] Jiacheng Li, Jingbo Shang, and Julian McAuley. UCTopic: Unsupervised contrastive learning for phrase representations and topic mining. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6159–6169, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.426. URL https://aclanthology.org/2022.acl-long.426.

[50] An Yan, Zhankui He, Jiacheng Li, Tianyang Zhang, and Julian John McAuley. Personalized showcases: Generating multi-modal explanations for recommendations. In Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete, editors, *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 2251–2255. ACM, 2023. doi: 10.1145/3539618.3592036. URL https://doi.org/10.1145/3539618.3592036.

[51] Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*, 2024.

[52] Yelp Inc. Yelp open dataset: An all-purpose dataset for learning, 2025. URL https://www.yelp.com/dataset. Accessed: 2026-01-27.

[53] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

[55] Xiaodong Lu, Leilei Sun, Tongyu Zhu, and Weifeng Lv. Improving temporal link prediction via temporal walk matrix projection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=Ti3ciyqlS3.

[56] Hanwen Liu, Longjiao Zhang, Rui Wang, Tongya Zheng, Sai Wu, Chang Yao, and Mingli Song. SALom: Structure aware temporal graph networks with long-short memory updater. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=hvaEMv4MVD.

[57] Yuxing Tian, Yiyan Qi, and Fan Guo. Freedyg: Frequency enhanced continuous-time dynamic graph model for link prediction. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=82Mc5ilInM.

[58] Yufei He, Yuan Sui, Xiaoxin He, Yue Liu, Yifei Sun, and Bryan Hooi. Unigraph2: Learning a unified embedding space to bind multimodal graphs. In *THE WEB CONFERENCE 2025*, 2025. URL https://openreview.net/forum?id=lEQEKUpXt6.

A Dataset

A.1 License

All used codes and datasets are publicly available and permit usage for research purposes under either the MIT License or Apache License 2.0.

The statistics of our dataset and comparison with previous datasets are summarized in Table 4.

Table 4: Statistics of existing and our datasets. Uniquely, our datasets provide multimodal attributes for *both* nodes and edges.

	Datasets	# Nodes	# Edges	# Timestamps	Node Figures	Edge Figures	Domain	Bipartite
Previous Dynamic Graph	tgbn-trade	255	468,245	32	✗	✗	Trade	✗
	tgbl-wiki	9,227	157,474	152,757	✗	✗	Interaction	✗
	tgbl-review	352,637	4,873,540	6,865	✗	✗	E-commerce	✓
	MOOC	7,144	411,749	345,600	✗	✗	Interaction	✓
	Enron	42,711	797,907	1,006	✗	✗	Business	✗
	ICEWS1819	31,796	1,100,071	730	✗	✗	Knowledge graph	✗
	Stack elec	397,702	1,262,225	5,224	✗	✗	Multi-round dialogue	✓
	Stack ubuntu	674,248	1,497,006	4,972	✗	✗	Multi-round dialogue	✓
Previous Multi-modal Graph	Goodreads	685,294	7,235,084	-	✓	✗	Books	✗
	Ele-fasion	97,766	199,602	-	✓	✗	E-commerce	✗
	Movies	16,672	218,390	-	✓	✗	E-commerce	✗
	Toys	20,695	126,886	-	✓	✗	E-commerce	✗
	Grocery	17,074	171,340	-	✓	✗	E-commerce	✗
	Reddit	99,638	1,167,188	-	✓	✗	Business	✗
Ours	GDELT	6,546	123,365	2,710	✗	✓	Sociopolitics	✗
	GDELT (full)	18,918	1,458,728	32,694	✗	✓	Sociopolitics	✗
	Google Map City	353,255	377,077	928	✗	✓	Point of Interest	✓
	Google Map City (full)	1,058,101	2,874,839	5,856	✗	✓	Point of Interest	✓
	Amazon - Industrial and Scientific	365,817	555,552	4,867	✓	✓	E-commerce	✓
	Amazon - Office Products	815,855	1,362,819	6,575	✓	✓	E-commerce	✓
	Yelp	535,362	1,169,612	730	✓	✗	Business	✓
	Yelp (full)	2,138,275	6,990,280	6,036	✓	✗	Business	✓

A.2 Statistical Analysis

We present a comprehensive statistical analysis of the datasets, examining the distributions of text/image attributes and temporal dynamics in Figure 5.

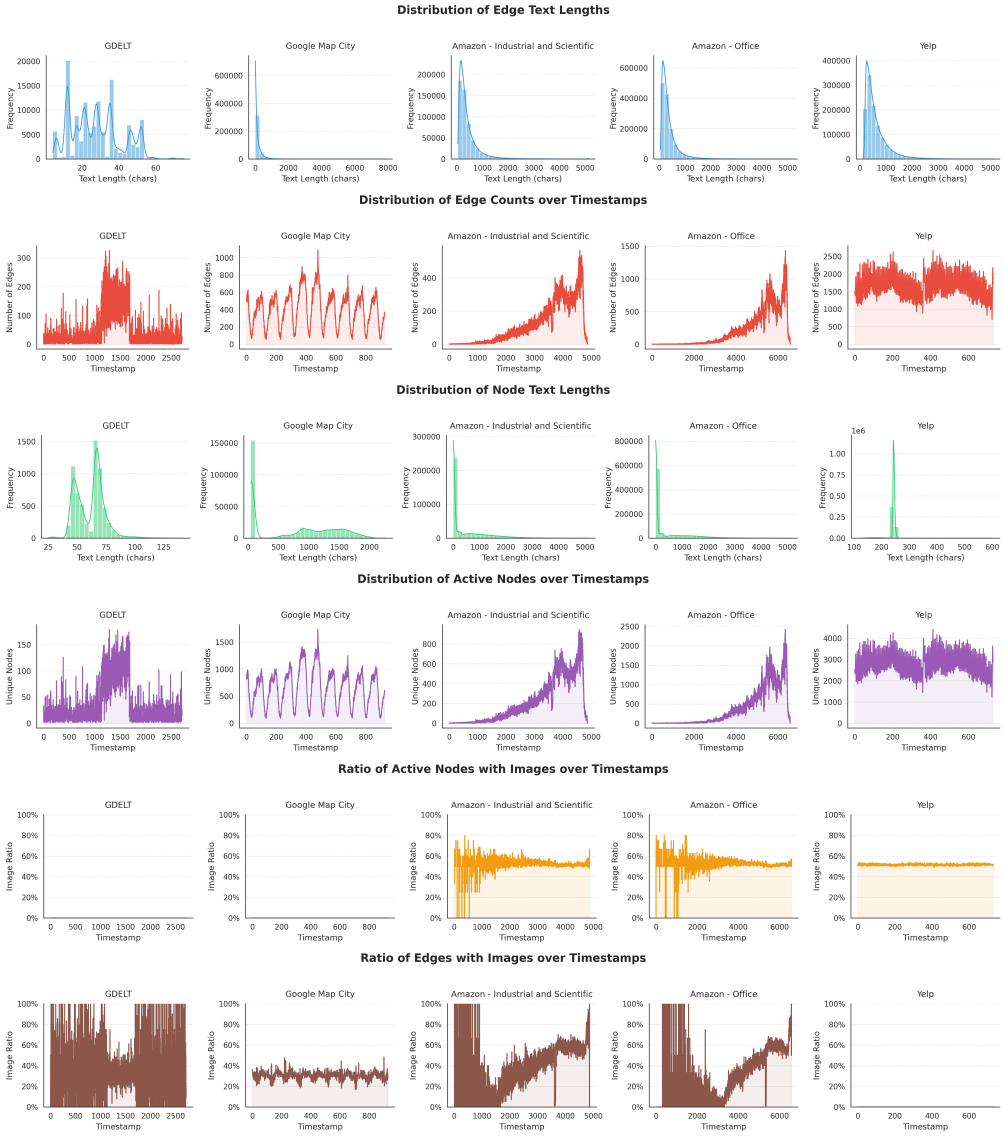
The distribution of text lengths for edges and nodes reveals distinct domain-specific characteristics. (1) **Edge Text Lengths**: GDELT presents a multimodal distribution concentrated at shorter lengths, consistent with the heterogeneous endogeneity. Google Map City displays a heavy-tailed distribution, featuring sharp peaks with a sparse tail. Yelp and Amazon exhibit unimodal, right-skewed distributions. (2) **Node Text Lengths**: GDELT displays a distinct bimodal distribution, indicating two dominant types of node attributes. Meanwhile, the bipartite datasets (Amazon, Yelp, and GMCT) show highly skewed distributions with sharp peaks, indicating homogeneity among text attributes.

The temporal evolution of the datasets, depicted in Figure 5.2 and Figure 5.4, highlights four highly distinct dynamic patterns. (1) Event-Driven Bursts: Instead of simple stationary volatility, GDELT exhibits a prominent complanate peak in the middle of the timeline, suggesting a surge driven by major real-world exogenous events. (2) Periodic Dynamics: Google Map City demonstrates strong, regular periodicity, reflecting clear cyclical patterns in human mobility or platform usage. (3) Continuous Stability: Unlike typical platform adoption curves, Yelp maintains a relatively stable and uniform interaction volume across the entire observed period. (4) Evolutionary Growth: The Amazon datasets exhibit an evolutionary phase: a "cold start" period followed by exponential growth and subsequent structural shifts. Furthermore, across all datasets, the distribution of active nodes closely tracks the edge distribution, indicating simultaneity between node activation and edge formation.

Furthermore, the temporal evolution of image ratios across active nodes and edges reveals complex structural dynamics. Regarding the node image ratios, we observe distinct stabilization behaviors: while Yelp demonstrates long-term stability with a consistent ratio throughout the timeline, the Amazon datasets initially display significant volatility before converging to a stable equilibrium. Conversely, the edge image ratios exhibit dynamic trends corresponding to the underlying interaction volumes. Notably, GDELT presents a prominent complanate valley in its edge image ratio that inversely mirrors the complanate peak observed in its temporal volume, suggesting a dilution of multimedia content. Google Map City exhibits regular periodic oscillations in its edge image ratio,

aligning with its cyclical temporal dynamics. Meanwhile, the Amazon datasets exhibit an initial volatile decline followed by a steady, continuous ascent.

Figure 5: Dataset Statistical Analysis.



A.3 Case Study

To visualize the performance of the VLM-based quality assessment, we demonstrate some case studies on Amazon datasets in Figure 6. Across the four case studies, the VLM relevance scores are generally consistent with human judgment and effectively reflect the degree of semantic alignment between each image-text pair. In case (a), the image directly depicts the M&M'S Caramel Chocolate Candy product, and key information such as the brand, flavor, packaging, and sharing-size format is clearly aligned with the text, which justifies the maximum relevance score. In contrast, case (b) receives a very low relevance score because the image is blurry and visually uninformative. The text mentions a damaged board, while the image does not provide sufficient evidence to identify the object or support the review content. Case (c) shows a brand-only image with limited product-level information. Although the image displays the "STACK! CO" logo, which matches the brand name in the text, it does not show the actual half-barrel keg dolly or any of the product attributes described in the text, such as the wheels, locking caster, diameter, or drainage design. Therefore, the VLM

assigns a low relevance score, indicating that merely matching the brand identity is insufficient for establishing a strong image-text correspondence. Finally, in case (d), the image clearly shows brass-colored threaded fittings that match the review content about machining quality, internal reduction, and air/liquid flow, resulting in a high relevance score. Overall, these examples suggest that the VLM-based relevance assessment provides reliable judgments. The VLM assigns high scores to image-text pairs with clear semantic alignment while penalizing blurry, uninformative, or only weakly related images. This supports our use of the relevance score as the criterion for filtering image-text pairs.

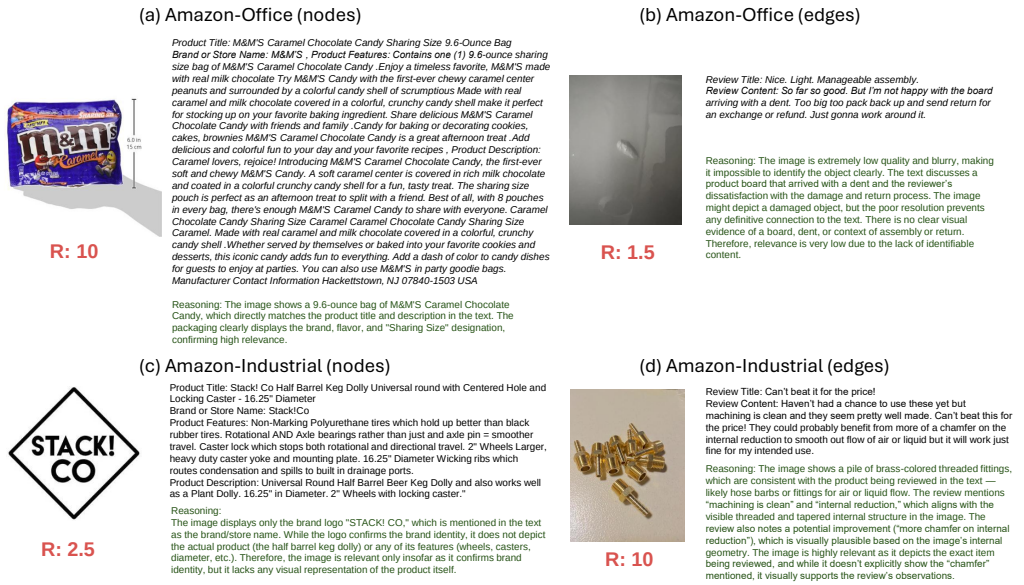


Figure 6: Case studies of VLM-based image-text relevance assessment. The four examples illustrate how the VLM assigns relevance scores to different image-text pairs, including highly matched product images and low-quality visual inputs. These cases show that the VLM can effectively identify semantically aligned image-text pairs, supporting its use as a relevance-based filtering criterion.

B Pseudocode of TST-MF

Algorithm 1 TST-MF: Task-Guided Spatio-Temporal Multimodal Fusion

Require: A batch of temporal interactions $\mathcal{B} = \{(u_i, v_i, t_i)\}_{i=1}^{B}$; historical dynamic graph $\mathcal{G}_{<t}$; node modality features $\{\mathbf{X}^{(m)}\}_{m=1}^{M}$; edge features $\mathbf{E}$; patch size p.

Ensure: Source and destination embeddings $\mathbf{z}_u, \mathbf{z}_v$ for downstream tasks.

$\mathbf{q} \leftarrow \mathrm{TaskProj}([\mathbf{x}_u \,\|\, \mathbf{x}_v])$ ▷ Construct target-guided task query.

for $r \in \{u, v\}$ **do** ▷ Encode source and destination nodes symmetrically.

  $(\mathcal{N}_r, \mathcal{E}_r, \mathcal{T}_r) \leftarrow \mathrm{SampleTemporalNeighbors}(r, t, \mathcal{G}_{<t})$ ▷ Retrieve historical context.

  $\mathcal{H}_r \leftarrow \mathrm{BuildPatchSequence}(\mathcal{N}_r, \mathcal{E}_r, \mathcal{T}_r, \mathbf{X}, \mathbf{E}, p)$ ▷ Patch related signals.

  $\mathbf{H}_r^{\mathrm{hist}} \leftarrow \mathrm{HistoricalEvolutionTracker}(\mathcal{H}_r)$ ▷ Model temporal evolution by self-attention.

  $\mathbf{S}_r^{\mathrm{struct}} \leftarrow \mathrm{StructuralNeighborAggregator}(\mathbf{x}_r, \{\mathbf{x}_j : j \in \mathcal{N}_r\})$ ▷ Aggregate structural context.

  $\mathbf{c}_r \leftarrow \mathrm{ContextCrossAttn}(\mathbf{Q} = \mathbf{q}, \mathbf{K}, \mathbf{V} = \{\mathbf{H}_r^{\mathrm{hist}}, \mathbf{S}_r^{\mathrm{struct}}\})$ ▷ Fuse historical and structural context.

  $\mathbf{m}_r \leftarrow \mathrm{ModalityCrossAttn}(\mathbf{Q} = \mathbf{c}_r, \mathbf{K}, \mathbf{V} = \{\mathbf{x}_r^{(m)}\}_{m=1}^{M})$ ▷ Select task-relevant modalities.

  $\mathbf{z}_r \leftarrow \mathrm{MLP}(\mathbf{m}_r)$ ▷ Project to final node representation.

end for

return $\mathbf{z}_u, \mathbf{z}_v$ ▷ Used by the task-specific prediction head.

C Evaluation Settings

C.1 Environment

We conduct training with the following hardware configurations:

- Operating System: Ubuntu 20.04.6 LTS
- CPU: AMD EPYC 7763 64-Core Processor
- GPU: NVIDIA A100-SXM4-40GB GPU and NVIDIA A100-SXM4-80GB GPU

C.2 Hyperparameters

We involved models DyGFormer, TGAT, GraphMixer, CAWN, FreeDyG, TCL, TPNet, SALoM, Gated Fusion, Modality Attention, MoE fusion, and TST-MF, resulting in a total of 12 models. The model architectures and training configurations are summarized in Table 5 and 6. For each method, we conducted *three* runs with different random seeds 0, 1, 2.

Table 5: Common hyperparameters, except that GraphMixer uses a dropout rate of 0.5 on edge classification.

Hyperparameter	**Value**
Batch size	256
Optimizer	Adam
Learning rate	1×10^{-4}
Weight decay	0.0
Dropout	0.1
Max epochs	200
Early stopping patience	5
Test interval (epochs)	1
#Runs	3
Validation ratio	0.15
Test ratio	0.15

Table 6: Model-specific hyperparameter. † marks a maximum input sequence length, and SALoM lists both. For CAWN, #Layers and #Heads are the walk length and the walk-encoder heads. The fusion module of modality attention adds 4 heads.

Model	**#Neighbors**	**#Layers**	**#Heads**	**Sampling**	**#Params**
TGAT	20	2	2	recent	50.7M
CAWN	32	1	8	interval	161.3M
TCL	20	2	2	recent	9.3M
GraphMixer	20	2	–	recent	7.1M
FreeDyG	32	2	–	recent	57.0M
TPNet	20	2	–	recent	57.4M
SALoM	10/512†	1	2	recent	276.2M
DyGFormer	48†	2	2	recent	6.2M
Gated fusion	48†	2	2	recent	5.9M
Modality attention	48†	2	2	recent	7.1M
MoE fusion	48†	2	2	recent	23.7M
TST-MF	48†	2	2	recent	12.0M

C.3 Prompt

The prompt used in VLM-based quality assessment is shown in Listing 1.

Listing 1: Prompt used in VLM-based quality assessment.

````
### Task
Evaluate the relationship between the provided image and the following text.

### Text
{text}

{extra_image_description (optional)}

### Evaluation Criteria
We want to measure the following aspect:
- Relevance: How related is the image to the text? (e.g., they depict the same item
or topic. The image might provide extra information not mentioned in the text, which
 is completely fine and should not lower the relevance score).

### Output Instructions
Please explain your reasoning in a few sentences first.
After your reasoning, provide two scores, each ranging from 0 to 10 (you can use
decimals like 7.3):
- relevance_score
- relevance_confidence (Your confidence in the Relevance Score)

### Output Format
Output your final response STRICTLY in the following format. Do NOT copy the
placeholder numbers, make your own judgments. Do not add any text after the JSON
block.

Reasoning:
[Your reasoning here]

```json
{{"relevance_score": [0-10 score], "relevance_confidence": [0-10 score]}}
```
````

D Supplementary Empirical Results

The AUC-ROC results of different methods on the link prediction task are provided in Table 7. The Precision, Recall and F-score results on the edge classification task are provided in Table 8, Table 9 and Table 10. The Hits@1 and Hits@3 results are provided in Table 11 and Table 12, respectively. The comparison of multimodal fusion strategies on all metrics is reported in Table 13.

Our pruning process increases multimodal information density, but also changes the retained graph size and modality availability. To assess the sensitivity of our findings to this preprocessing, we additionally evaluate models on the complete GDELT and GMCT datasets, preserving the interactions available before pruning. The results are provided in Tables 14–16. Results on the complete Yelp dataset are omitted because its scale makes evaluation infeasible for most evaluated models under single NVIDIA-A100 GPU setup. The results show that our main observations remain broadly consistent after preserving the original modality sparsity.

Moreover, we examine how different missing modality processing choices affects the benchmark results, we compare four imputation strategies: (i) zero filling, which replaces a missing embedding with a zero vector; (ii) mean filling, which uses the mean of the available embeddings; (iii) learned embedding, which uses a trainable vector to represent the missing modality; and (iv) duplication, which uses the corresponding other modality embedding. The results are reported in Tables 17–23. This comparison examines both the sensitivity of absolute performance to imputation and the stability of the relative performance of different methods. Results indicate that the different imputation strategies yield comparable performance, and the relative performance trends remain consistent.

What's more, our default node-retrieval protocol ranks the true destination against 100 randomly sampled negative destinations. Although this provides a consistent evaluation setting, random negatives can be easy to distinguish on some datasets, leading to near-saturated Hits@10 scores. We therefore additionally evaluate historical negative sampling, which constructs negative candidates from nodes that have recently interacted with the query source before the query time, excluding the current positive destination. This setting emphasizes historically relevant alternatives and provides a more challenging ranking evaluation. And to distinguish the retrieval of previously observed connections from the discovery of new ones, we further report recurrence-aware results. A test interaction is classified as repeated if its source–destination pair has appeared before the query time, and as novel otherwise. Table 24 reports overall Hits@10 across all test interactions, while Tables 25 and 26 report results on repeated and novel interactions, respectively. Under the stricter protocol, the metric no longer approaches saturation, confirming that the historical negative sampling is substantially more discriminative.

Table 7: AUC-ROC (%) for link prediction. "*tr.*" and "*in.*" refer to the transductive and inductive setting, respectively. The best results are highlighted in bold and the second-best results are underlined.

	Datasets	Feature	TGAT	CAWN	TCL	GraphMixer	FreeDyG	TPNet	SALoM	DyGFormer	TST-MF
tr.	GDELT	Random	88.30 ± 1.96	92.22 ± 0.46	91.07 ± 0.03	90.70 ± 0.12	90.73 ± 0.31	91.61 ± 0.54	90.26 ± 1.93	91.56 ± 0.08	**92.23 ± 0.25**
		Text	88.82 ± 1.21	92.60 ± 0.08	91.08 ± 0.22	91.36 ± 0.35	91.54 ± 0.34	91.76 ± 0.21	90.92 ± 0.60	92.05 ± 0.21	**93.70 ± 0.11**
		Text+Image	90.16 ± 0.27	92.73 ± 0.26	90.84 ± 0.32	91.15 ± 0.14	91.47 ± 0.18	92.22 ± 0.22	91.18 ± 0.15	91.93 ± 0.24	**93.71 ± 0.08**
	GMCT	Random	73.63 ± 2.25	76.69 ± 0.02	76.54 ± 0.14	76.72 ± 0.01	76.77 ± 0.01	78.46 ± 0.61	61.68 ± 2.90	**94.36 ± 0.54**	81.17 ± 4.79
		Text	81.96 ± 3.60	79.40 ± 0.15	77.34 ± 0.85	79.28 ± 0.35	78.11 ± 0.31	79.08 ± 0.46	58.75 ± 2.88	**95.35 ± 0.66**	91.89 ± 0.76
		Text+Image	81.60 ± 2.33	79.56 ± 0.05	77.42 ± 0.60	79.48 ± 0.66	78.29 ± 0.25	78.78 ± 0.19	57.47 ± 6.18	**95.08 ± 0.27**	92.49 ± 0.66
	Amazon - Industrial	Random	91.92 ± 0.25	91.80 ± 0.17	91.79 ± 0.25	**92.49 ± 0.18**	91.69 ± 0.30	92.40 ± 0.33	82.33 ± 3.29	89.27 ± 1.11	92.48 ± 0.05
		Text	94.09 ± 0.11	92.88 ± 0.03	93.48 ± 0.24	94.23 ± 0.07	92.90 ± 0.31	93.07 ± 0.16	83.34 ± 2.23	92.73 ± 0.26	**94.26 ± 0.05**
		Text+Image	94.23 ± 0.13	93.00 ± 0.13	93.43 ± 0.30	94.09 ± 0.32	92.73 ± 0.32	93.30 ± 0.34	84.21 ± 1.21	92.94 ± 0.29	**94.26 ± 0.16**
	Amazon - Office	Random	91.88 ± 0.81	91.27 ± 0.38	**92.48 ± 0.23**	90.33 ± 0.07	91.36 ± 0.29	91.84 ± 0.40	86.82 ± 1.96	91.04 ± 0.68	92.01 ± 0.17
		Text	**93.07 ± 0.59**	92.40 ± 0.23	92.22 ± 0.14	92.39 ± 0.97	92.83 ± 0.18	92.41 ± 0.70	87.45 ± 1.60	91.85 ± 0.58	92.81 ± 0.09
		Text+Image	92.74 ± 0.09	92.58 ± 0.24	92.49 ± 0.13	92.64 ± 1.08	92.87 ± 0.33	92.02 ± 0.10	85.36 ± 1.89	91.41 ± 0.16	**93.23 ± 0.43**
	Yelp	Random	81.22 ± 0.17	80.98 ± 0.07	80.71 ± 0.69	80.90 ± 0.46	81.84 ± 0.07	**91.04 ± 0.07**	73.72 ± 1.00	81.65 ± 0.11	82.01 ± 0.10
		Text	92.80 ± 0.14	92.71 ± 0.17	90.40 ± 1.81	87.63 ± 0.03	90.87 ± 0.06	91.55 ± 0.11	75.30 ± 0.62	89.00 ± 6.08	**93.00 ± 0.02**
		Text+Image	92.63 ± 0.36	92.40 ± 0.16	87.98 ± 5.82	87.57 ± 0.16	90.65 ± 0.08	91.85 ± 0.24	73.51 ± 0.57	88.84 ± 5.74	**93.83 ± 0.02**
in.	GDELT	Random	74.26 ± 0.21	78.18 ± 0.66	76.23 ± 0.13	75.71 ± 0.86	76.17 ± 0.70	76.25 ± 1.61	76.52 ± 0.62	79.23 ± 0.16	**79.38 ± 0.49**
		Text	74.86 ± 1.33	79.34 ± 0.34	76.32 ± 0.11	77.14 ± 0.65	78.85 ± 0.59	77.10 ± 0.15	77.14 ± 0.39	81.37 ± 0.42	**84.26 ± 0.03**
		Text+Image	76.86 ± 0.77	79.52 ± 0.30	76.31 ± 0.24	76.85 ± 0.45	78.53 ± 0.47	78.15 ± 0.73	77.67 ± 0.39	81.29 ± 0.34	**84.51 ± 0.11**
	GMCT	Random	60.20 ± 2.42	60.22 ± 0.64	60.34 ± 0.30	60.82 ± 0.62	60.62 ± 0.15	63.48 ± 0.65	51.29 ± 1.89	**80.14 ± 1.38**	63.14 ± 1.71
		Text	67.58 ± 6.10	62.73 ± 0.29	60.96 ± 0.69	63.00 ± 0.92	62.01 ± 0.47	63.31 ± 0.37	49.90 ± 0.61	**82.68 ± 2.68**	67.65 ± 1.42
		Text+Image	66.61 ± 4.08	62.84 ± 0.27	60.44 ± 0.63	63.46 ± 0.93	62.14 ± 0.55	63.28 ± 0.57	49.33 ± 2.51	**82.39 ± 2.37**	67.77 ± 1.48
	Amazon - Industrial	Random	71.81 ± 1.72	71.44 ± 0.51	72.27 ± 1.21	74.42 ± 0.79	70.22 ± 1.48	72.97 ± 1.87	66.06 ± 1.28	70.84 ± 0.40	**74.93 ± 0.67**
		Text	79.22 ± 0.45	74.85 ± 0.16	77.58 ± 1.25	**81.00 ± 0.30**	74.52 ± 1.34	75.23 ± 0.54	66.23 ± 2.28	72.95 ± 0.73	79.62 ± 0.71
		Text+Image	79.80 ± 0.60	75.30 ± 0.73	77.27 ± 1.05	**80.17 ± 1.50**	73.69 ± 1.12	76.37 ± 1.42	66.73 ± 1.22	73.11 ± 1.06	80.00 ± 0.60
	Amazon - Office	Random	71.84 ± 2.71	68.68 ± 1.68	**73.99 ± 0.84**	63.89 ± 0.21	69.11 ± 1.55	70.79 ± 2.38	70.14 ± 1.00	73.59 ± 1.09	72.84 ± 0.14
		Text	**75.89 ± 3.55**	73.04 ± 0.90	72.87 ± 0.92	72.51 ± 4.44	74.52 ± 0.72	72.50 ± 2.94	70.11 ± 1.48	71.97 ± 3.86	74.81 ± 0.20
		Text+Image	73.79 ± 0.43	74.21 ± 1.05	73.95 ± 0.68	73.85 ± 4.73	74.66 ± 1.47	70.83 ± 0.53	69.05 ± 2.35	68.53 ± 0.07	**75.38 ± 1.60**
	Yelp	Random	76.09 ± 0.18	75.56 ± 0.26	75.75 ± 0.21	75.48 ± 0.43	76.51 ± 0.10	**86.27 ± 0.15**	66.12 ± 0.66	76.26 ± 0.12	76.65 ± 0.12
		Text	89.06 ± 0.22	88.88 ± 0.27	86.15 ± 2.12	82.96 ± 0.15	86.76 ± 0.12	86.93 ± 0.07	67.34 ± 0.86	84.80 ± 6.97	**89.36 ± 0.05**
		Text+Image	88.92 ± 0.51	88.43 ± 0.20	83.69 ± 6.53	83.00 ± 0.10	86.48 ± 0.11	87.35 ± 0.37	64.82 ± 0.42	84.57 ± 6.50	**89.10 ± 0.05**

Table 8: Precision (%) for edge classification. The best results are highlighted in bold and the second-best results are underlined.

Datasets	Feature	TGAT	CAWN	TCL	GraphMixer	FreeDyG	TPNet	SALoM	DyGFormer	TST-MF
GDELT	Random	48.22 ± 0.33	**50.02 ± 3.28**	42.98 ± 3.83	48.67 ± 1.10	45.38 ± 1.53	47.34 ± 2.56	40.69 ± 0.00	47.59 ± 0.17	49.23 ± 1.43
	Text	47.96 ± 0.39	49.62 ± 2.52	40.77 ± 0.00	48.22 ± 0.15	44.34 ± 3.19	47.29 ± 1.25	46.33 ± 1.78	47.62 ± 0.06	**52.56 ± 1.28**
	Text+Image	48.15 ± 0.09	51.62 ± 5.99	40.77 ± 0.00	48.08 ± 0.31	46.60 ± 0.21	48.89 ± 2.63	47.78 ± 0.17	47.54 ± 0.10	**53.56 ± 1.70**
GMCT	Random	55.26 ± 0.56	55.42 ± 0.02	54.36 ± 0.20	58.09 ± 0.10	50.46 ± 1.80	50.34 ± 0.66	45.13 ± 0.00	**59.44 ± 0.13**	51.60 ± 0.31
	Text	54.68 ± 0.04	56.46 ± 1.92	54.56 ± 0.20	58.81 ± 1.13	47.12 ± 3.45	52.66 ± 0.31	45.13 ± 0.00	**59.32 ± 0.32**	47.90 ± 4.80
	Text+Image	55.45 ± 0.29	56.64 ± 2.07	54.40 ± 0.19	**58.58 ± 0.76**	47.88 ± 4.75	49.49 ± 3.77	47.71 ± 4.46	56.77 ± 1.72	53.45 ± 1.57
Amazon - Industrial	Random	48.05 ± 0.87	53.14 ± 3.78	47.27 ± 0.28	**57.29 ± 1.93**	46.05 ± 0.37	45.95 ± 0.63	42.97 ± 2.16	56.46 ± 2.26	47.96 ± 0.01
	Text	48.34 ± 0.41	55.81 ± 1.19	47.67 ± 0.19	55.66 ± 0.83	53.64 ± 1.34	50.54 ± 2.03	50.42 ± 1.79	**56.32 ± 1.42**	50.18 ± 1.77
	Text+Image	48.50 ± 0.12	**59.83 ± 2.25**	48.05 ± 0.42	58.12 ± 0.51	50.28 ± 3.44	50.57 ± 3.10	47.88 ± 0.07	59.38 ± 1.85	50.32 ± 1.86
Amazon - Office	Random	55.09 ± 0.29	**57.84 ± 1.42**	50.48 ± 4.68	55.36 ± 0.34	48.23 ± 0.13	47.94 ± 0.51	45.08 ± 3.57	55.83 ± 0.66	49.43 ± 0.06
	Text	54.66 ± 0.72	**57.62 ± 1.81**	49.98 ± 4.80	56.38 ± 0.28	55.70 ± 0.64	54.13 ± 0.70	51.11 ± 1.89	56.25 ± 0.81	54.13 ± 0.03
	Text+Image	53.51 ± 0.27	**57.67 ± 0.90**	50.40 ± 5.15	56.58 ± 0.86	54.18 ± 0.10	52.75 ± 0.39	53.41 ± 0.18	56.02 ± 0.50	54.42 ± 0.02
Yelp	Random	46.46 ± 3.75	49.99 ± 0.31	48.40 ± 1.25	48.98 ± 0.11	39.64 ± 1.47	38.90 ± 0.81	39.22 ± 0.23	**50.29 ± 0.18**	44.01 ± 1.33
	Text	53.80 ± 1.32	53.12 ± 0.51	52.09 ± 0.64	52.44 ± 0.11	47.41 ± 0.39	45.88 ± 1.16	48.18 ± 2.00	54.09 ± 0.07	**54.91 ± 0.21**
	Text+Image	53.47 ± 0.13	52.58 ± 0.90	51.97 ± 0.84	52.91 ± 0.55	47.04 ± 0.88	45.41 ± 1.53	52.25 ± 0.14	53.99 ± 0.18	**55.13 ± 0.87**

Table 9: Recall (%) for edge classification. The best results are highlighted in bold and the second-best results are underlined.

Datasets	Feature	TGAT	CAWN	TCL	GraphMixer	FreeDyG	TPNet	SALoM	DyGFormer	TST-MF
GDELT	Random	64.07 ± 0.04	64.20 ± 0.08	63.85 ± 0.00	**64.32 ± 0.22**	63.32 ± 0.51	62.55 ± 1.04	63.79 ± 0.00	63.91 ± 0.15	62.50 ± 1.35
	Text	63.98 ± 0.08	64.23 ± 0.12	63.85 ± 0.00	**64.25 ± 0.05**	63.58 ± 0.27	63.82 ± 0.03	63.88 ± 0.11	63.94 ± 0.09	63.70 ± 0.04
	Text+Image	64.05 ± 0.03	64.24 ± 0.12	63.85 ± 0.00	**64.26 ± 0.11**	63.36 ± 0.60	63.30 ± 0.88	63.96 ± 0.09	63.93 ± 0.08	63.77 ± 0.07
GMCT	Random	67.10 ± 0.39	67.45 ± 0.09	67.01 ± 0.27	67.36 ± 0.31	66.91 ± 0.22	61.93 ± 3.65	67.18 ± 0.00	**67.86 ± 0.03**	53.66 ± 1.57
	Text	67.16 ± 0.06	**67.60 ± 0.30**	66.39 ± 0.92	67.46 ± 0.25	67.16 ± 0.03	66.39 ± 0.65	67.18 ± 0.00	67.39 ± 0.21	65.65 ± 2.66
	Text+Image	67.36 ± 0.15	**67.60 ± 0.23**	66.61 ± 0.54	67.48 ± 0.35	66.64 ± 0.94	66.53 ± 1.01	67.18 ± 0.00	67.47 ± 0.14	67.18 ± 0.00
Amazon - Industrial	Random	64.48 ± 0.71	65.15 ± 0.19	64.66 ± 0.38	63.96 ± 0.30	61.76 ± 2.36	63.03 ± 2.26	64.59 ± 0.00	**65.36 ± 0.14**	58.26 ± 1.42
	Text	64.65 ± 0.39	64.97 ± 0.19	64.94 ± 0.07	64.36 ± 0.32	65.14 ± 0.18	64.50 ± 0.17	63.88 ± 0.34	**65.42 ± 0.06**	64.69 ± 0.03
	Text+Image	64.81 ± 0.44	64.96 ± 0.21	64.60 ± 0.55	64.32 ± 0.06	65.16 ± 0.37	64.23 ± 0.67	65.01 ± 0.06	**65.20 ± 0.11**	**65.20 ± 0.01**
Amazon - Office	Random	66.36 ± 0.44	66.64 ± 0.01	65.71 ± 0.96	66.23 ± 0.30	63.37 ± 2.34	62.44 ± 0.47	65.60 ± 0.02	**66.76 ± 0.06**	64.19 ± 2.02
	Text	66.39 ± 0.27	66.66 ± 0.03	66.06 ± 0.67	66.26 ± 0.24	66.67 ± 0.09	66.66 ± 0.08	66.51 ± 0.24	**66.94 ± 0.06**	66.51 ± 0.01
	Text+Image	66.09 ± 0.17	66.70 ± 0.06	66.02 ± 0.53	66.06 ± 0.48	66.52 ± 0.06	66.12 ± 0.21	66.31 ± 0.31	**66.90 ± 0.03**	66.59 ± 0.01
Yelp	Random	54.63 ± 0.73	55.29 ± 0.15	55.03 ± 0.25	54.75 ± 0.15	50.65 ± 0.26	49.19 ± 1.68	50.82 ± 0.02	**55.69 ± 0.03**	51.68 ± 0.04
	Text	57.94 ± 0.86	58.02 ± 0.19	55.63 ± 3.59	57.54 ± 0.28	54.47 ± 0.10	53.68 ± 0.06	55.17 ± 0.43	**58.52 ± 0.06**	57.64 ± 0.03
	Text+Image	57.93 ± 0.17	57.54 ± 0.48	57.63 ± 0.22	57.72 ± 0.44	54.28 ± 0.26	53.79 ± 0.25	55.96 ± 0.28	58.53 ± 0.19	**58.72 ± 1.11**

Table 10: F1 (%) for edge classification. The best results are highlighted in bold and the second-best results are underlined.

Datasets	Feature	TGAT	CAWN	TCL	GraphMixer	FreeDyG	TPNet	SALoM	DyGFormer	TST-MF
GDELT	Random	50.95 ± 0.23	**53.45 ± 0.06**	49.78 ± 0.02	52.54 ± 0.75	49.82 ± 0.18	50.39 ± 0.40	49.69 ± 0.00	52.96 ± 0.05	53.32 ± 0.04
	Text	50.48 ± 0.35	**53.50 ± 0.06**	49.77 ± 0.00	52.10 ± 0.29	50.83 ± 1.01	50.11 ± 0.25	50.04 ± 0.45	52.93 ± 0.12	53.38 ± 0.04
	Text+Image	50.81 ± 0.16	53.45 ± 0.10	49.77 ± 0.00	52.66 ± 0.86	51.78 ± 0.85	51.19 ± 1.39	50.72 ± 0.88	52.81 ± 0.07	**53.64 ± 0.03**
GMCT	Random	59.20 ± 1.08	59.35 ± 0.11	56.99 ± 2.53	**59.93 ± 0.17**	54.18 ± 0.22	54.48 ± 0.12	53.99 ± 0.00	58.40 ± 0.06	52.12 ± 0.08
	Text	58.24 ± 0.21	59.27 ± 0.02	58.73 ± 0.45	**59.98 ± 0.16**	53.99 ± 0.00	55.32 ± 0.84	53.99 ± 0.00	59.68 ± 0.35	54.72 ± 1.27
	Text+Image	59.41 ± 0.31	59.33 ± 0.05	58.48 ± 0.65	59.88 ± 0.17	54.57 ± 1.00	54.83 ± 1.15	54.04 ± 0.08	**59.89 ± 0.06**	56.32 ± 0.21
Amazon - Industrial	Random	54.80 ± 0.86	55.79 ± 0.34	53.41 ± 0.64	**56.58 ± 0.13**	50.86 ± 0.36	51.13 ± 0.08	50.70 ± 0.00	55.83 ± 0.48	51.00 ± 0.04
	Text	55.11 ± 0.42	55.75 ± 0.16	53.80 ± 0.48	**56.56 ± 0.50**	55.24 ± 0.33	54.97 ± 0.07	55.26 ± 0.23	55.59 ± 0.13	55.48 ± 0.01
	Text+Image	55.24 ± 0.10	55.64 ± 0.09	54.58 ± 0.51	**56.78 ± 0.28**	54.34 ± 0.45	54.26 ± 0.50	52.91 ± 0.95	55.72 ± 0.44	54.89 ± 0.12
Amazon - Office	Random	55.37 ± 1.56	57.11 ± 0.31	55.26 ± 1.04	**58.04 ± 0.28**	52.40 ± 0.05	52.71 ± 0.13	52.00 ± 0.06	56.73 ± 0.27	52.90 ± 0.03
	Text	55.64 ± 0.74	57.62 ± 0.20	54.53 ± 2.82	**58.17 ± 0.30**	56.04 ± 0.16	55.70 ± 0.30	55.30 ± 0.83	57.33 ± 0.25	56.16 ± 0.02
	Text+Image	55.44 ± 0.46	57.73 ± 0.17	53.99 ± 1.90	**58.11 ± 0.43**	56.17 ± 0.03	56.35 ± 0.18	55.54 ± 1.00	57.40 ± 0.34	56.97 ± 0.02
Yelp	Random	46.10 ± 1.80	48.30 ± 0.17	46.44 ± 0.98	**49.01 ± 0.43**	38.06 ± 1.01	38.96 ± 1.11	34.89 ± 0.41	48.54 ± 0.06	42.46 ± 0.03
	Text	50.94 ± 1.10	51.63 ± 0.09	46.76 ± 8.08	52.00 ± 0.20	45.56 ± 0.31	44.95 ± 0.40	44.46 ± 1.77	52.17 ± 0.22	**52.83 ± 0.21**
	Text+Image	50.98 ± 0.34	51.31 ± 0.22	50.93 ± 0.75	52.37 ± 0.62	45.60 ± 0.34	44.78 ± 0.40	46.58 ± 0.76	52.34 ± 0.66	**52.43 ± 2.95**

Table 11: Hits@1 (%) for node retrieval. "*tr.*" and "*in.*" refer to the transductive and inductive setting, respectively. The best results are highlighted in bold and the second-best results are underlined.

	Datasets	Feature	TGAT	CAWN	TCL	GraphMixer	FreeDyG	TPNet	SALoM	DyGFormer	TST-MF
tr.	GDELT	Random	28.12 ± 5.67	34.88 ± 0.69	31.63 ± 0.23	30.04 ± 1.52	28.29 ± 0.28	33.58 ± 2.69	32.54 ± 1.34	24.67 ± 5.01	**36.72 ± 0.29**
		Text	28.79 ± 3.26	33.23 ± 0.06	31.70 ± 0.11	31.53 ± 0.78	30.90 ± 1.44	33.56 ± 0.38	32.25 ± 1.04	33.99 ± 3.87	**38.25 ± 0.56**
		Text+Image	30.42 ± 1.64	33.65 ± 1.09	31.53 ± 0.31	31.11 ± 0.64	31.04 ± 0.49	36.34 ± 1.01	32.06 ± 0.77	34.19 ± 4.03	**38.31 ± 0.74**
	GMCT	Random	7.45 ± 2.79	11.74 ± 0.21	11.24 ± 0.47	11.09 ± 0.14	34.74 ± 1.26	12.85 ± 0.06	5.46 ± 0.06	**93.35 ± 3.69**	31.33 ± 18.55
		Text	15.04 ± 3.97	11.15 ± 0.33	11.42 ± 0.13	9.09 ± 2.51	35.77 ± 0.51	12.87 ± 0.42	5.50 ± 0.08	**88.78 ± 4.31**	70.15 ± 7.03
		Text+Image	13.63 ± 2.32	11.64 ± 0.20	11.58 ± 0.40	10.51 ± 1.45	35.98 ± 0.19	13.63 ± 0.27	5.30 ± 0.19	**93.29 ± 1.56**	73.27 ± 3.06
	Amazon - Industrial	Random	28.28 ± 3.48	27.74 ± 1.22	27.37 ± 4.13	32.82 ± 0.07	34.31 ± 2.70	30.65 ± 3.59	14.21 ± 3.75	**44.61 ± 18.94**	37.99 ± 3.49
		Text	35.99 ± 1.28	25.73 ± 0.18	33.22 ± 2.69	38.36 ± 0.55	38.99 ± 2.54	34.23 ± 1.01	13.22 ± 1.73	63.68 ± 3.07	**64.44 ± 0.29**
		Text+Image	37.95 ± 0.71	27.24 ± 1.35	33.27 ± 2.11	38.32 ± 1.45	37.92 ± 1.83	35.73 ± 2.32	14.64 ± 3.15	62.78 ± 4.46	**65.49 ± 6.91**
	Amazon - Office	Random	25.96 ± 7.56	21.59 ± 1.32	30.27 ± 1.17	7.99 ± 1.54	28.70 ± 4.49	24.14 ± 5.76	18.51 ± 3.20	**49.77 ± 9.00**	29.78 ± 3.53
		Text	29.74 ± 6.42	24.21 ± 3.12	28.06 ± 1.14	20.90 ± 11.34	36.94 ± 0.72	26.12 ± 6.89	18.51 ± 1.61	**65.07 ± 3.65**	55.63 ± 5.83
		Text+Image	24.51 ± 1.88	27.54 ± 1.14	28.33 ± 0.64	26.39 ± 8.54	36.78 ± 2.12	22.97 ± 2.67	17.74 ± 2.31	69.52 ± 1.77	**69.83 ± 1.73**
	Yelp	Random	12.64 ± 0.36	12.39 ± 0.60	12.11 ± 0.72	12.55 ± 0.52	15.34 ± 0.14	**28.32 ± 0.27**	18.64 ± 0.67	4.48 ± 5.44	15.67 ± 1.93
		Text	28.55 ± 0.34	27.73 ± 0.57	22.05 ± 3.45	19.81 ± 0.16	26.93 ± 0.22	29.05 ± 0.49	20.62 ± 0.66	24.76 ± 8.77	**29.71 ± 0.52**
		Text+Image	28.30 ± 1.40	26.96 ± 0.49	20.09 ± 6.56	19.61 ± 0.32	26.46 ± 0.42	29.75 ± 0.27	18.77 ± 1.15	24.07 ± 7.01	**29.94 ± 0.47**
in.	GDELT	Random	11.29 ± 0.69	15.81 ± 0.55	10.87 ± 0.08	11.03 ± 0.05	14.78 ± 0.68	14.03 ± 1.49	16.26 ± 1.21	15.94 ± 3.33	**16.73 ± 0.19**
		Text	11.51 ± 1.26	17.32 ± 0.44	10.66 ± 0.20	11.64 ± 1.02	17.08 ± 0.62	13.39 ± 0.23	16.13 ± 0.83	20.39 ± 1.52	**23.16 ± 1.23**
		Text+Image	12.92 ± 1.65	16.94 ± 0.51	10.87 ± 0.11	11.16 ± 0.30	16.06 ± 0.50	16.43 ± 1.29	15.73 ± 2.01	21.12 ± 1.80	**23.80 ± 0.73**
	GMCT	Random	3.41 ± 1.22	4.07 ± 0.14	3.95 ± 0.43	4.39 ± 0.18	28.98 ± 11.18	4.37 ± 0.23	2.05 ± 0.04	**67.32 ± 16.74**	7.73 ± 4.44
		Text	5.18 ± 1.90	3.78 ± 0.41	3.70 ± 0.18	2.91 ± 0.93	37.79 ± 1.26	4.43 ± 0.05	2.11 ± 0.08	**50.83 ± 8.65**	28.93 ± 5.33
		Text+Image	4.34 ± 1.21	3.33 ± 0.11	3.71 ± 0.24	3.20 ± 0.80	38.28 ± 0.19	4.96 ± 0.48	1.98 ± 0.09	**65.20 ± 7.17**	32.64 ± 2.59
	Amazon - Industrial	Random	13.36 ± 1.42	12.66 ± 1.35	11.91 ± 2.70	15.85 ± 0.40	**17.86 ± 0.83**	17.36 ± 1.29	4.59 ± 2.31	14.55 ± 8.88	16.73 ± 0.59
		Text	19.75 ± 0.45	13.19 ± 0.53	17.00 ± 2.32	20.45 ± 0.39	21.46 ± 1.32	20.59 ± 0.65	3.58 ± 0.76	27.07 ± 2.84	**32.02 ± 1.79**
		Text+Image	20.36 ± 0.56	13.17 ± 0.95	16.47 ± 2.09	20.54 ± 0.52	20.40 ± 1.59	20.52 ± 0.78	4.12 ± 1.27	23.93 ± 3.99	**38.92 ± 1.99**
	Amazon - Office	Random	14.14 ± 2.76	12.51 ± 1.49	14.14 ± 0.72	3.37 ± 1.50	18.63 ± 2.56	14.80 ± 2.99	6.00 ± 1.45	**23.92 ± 9.14**	13.82 ± 3.45
		Text	15.76 ± 3.00	12.21 ± 0.98	13.60 ± 1.04	10.68 ± 6.68	20.39 ± 0.36	15.18 ± 2.86	5.20 ± 0.99	**37.41 ± 4.95**	30.58 ± 5.16
		Text+Image	13.76 ± 1.27	13.68 ± 0.50	12.43 ± 0.36	14.83 ± 4.81	20.26 ± 0.06	14.42 ± 1.13	5.20 ± 1.12	44.47 ± 4.14	**48.45 ± 2.43**
	Yelp	Random	11.36 ± 0.37	11.02 ± 0.69	11.02 ± 0.44	11.52 ± 0.26	13.79 ± 0.13	**21.98 ± 0.39**	12.73 ± 0.60	3.46 ± 4.48	13.71 ± 0.60
		Text	23.20 ± 0.37	22.12 ± 0.52	18.41 ± 2.54	16.53 ± 0.05	21.97 ± 0.26	22.67 ± 0.59	14.23 ± 0.76	20.76 ± 6.60	**23.67 ± 0.53**
		Text+Image	23.12 ± 1.12	21.43 ± 0.40	16.88 ± 4.98	16.56 ± 0.28	21.73 ± 0.31	23.21 ± 0.24	13.16 ± 0.56	20.26 ± 5.04	**24.58 ± 0.45**

Table 12: Hits@3 (%) for node retrieval. "*tr.*" and "*in.*" refer to the transductive and inductive setting, respectively. The best results are highlighted in bold, and the second-best results are underlined.

	Datasets	Feature	TGAT	CAWN	TCL	GraphMixer	FreeDyG	TPNet	SALoM	DyGFormer	TST-MF
tr.	GDELT	Random	47.26 ± 6.65	55.88 ± 0.56	53.42 ± 0.19	48.30 ± 3.94	46.34 ± 1.30	51.59 ± 2.51	55.21 ± 2.00	35.27 ± 6.80	**57.46 ± 0.46**
		Text	49.37 ± 4.36	55.20 ± 0.54	53.64 ± 0.14	52.96 ± 1.39	53.50 ± 1.01	54.73 ± 0.82	55.01 ± 1.21	51.40 ± 6.58	**58.98 ± 1.08**
		Text+Image	51.56 ± 2.13	55.65 ± 1.06	53.46 ± 0.21	51.71 ± 2.01	51.67 ± 0.63	54.84 ± 2.20	56.03 ± 0.24	52.06 ± 6.41	**59.05 ± 0.36**
	GMCT	Random	20.00 ± 5.05	25.25 ± 0.31	25.20 ± 0.35	25.57 ± 0.35	39.36 ± 0.84	26.29 ± 0.07	11.35 ± 0.13	**100.00 ± 0.01**	59.31 ± 31.43
		Text	30.51 ± 6.38	25.00 ± 0.18	24.90 ± 0.29	23.08 ± 2.21	38.90 ± 0.37	26.56 ± 0.16	11.60 ± 0.16	**100.00 ± 0.01**	99.63 ± 0.34
		Text+Image	29.04 ± 4.09	25.53 ± 0.18	25.24 ± 0.17	25.07 ± 1.56	38.84 ± 0.27	27.28 ± 0.17	11.31 ± 0.22	**100.00 ± 0.01**	99.69 ± 0.69
	Amazon - Industrial	Random	44.03 ± 4.84	44.17 ± 0.85	46.34 ± 4.80	50.38 ± 0.84	46.78 ± 3.58	44.94 ± 4.79	31.47 ± 4.48	**80.80 ± 21.52**	63.54 ± 8.20
		Text	55.30 ± 1.25	44.90 ± 0.11	52.88 ± 2.32	58.30 ± 0.29	53.31 ± 4.03	48.67 ± 1.29	31.81 ± 2.45	**96.64 ± 0.52**	95.54 ± 0.24
		Text+Image	57.01 ± 0.78	46.79 ± 1.57	52.72 ± 2.19	57.74 ± 1.93	51.58 ± 1.63	51.10 ± 3.48	33.93 ± 5.36	**97.08 ± 0.70**	95.71 ± 2.92
	Amazon - Office	Random	41.34 ± 10.38	33.65 ± 2.04	49.68 ± 2.07	20.16 ± 1.15	38.43 ± 4.67	36.77 ± 6.98	40.13 ± 3.41	**82.55 ± 7.23**	48.20 ± 1.51
		Text	47.65 ± 8.81	41.53 ± 3.16	45.57 ± 3.30	37.64 ± 13.81	51.97 ± 1.50	39.96 ± 8.87	40.98 ± 2.43	93.37 ± 0.75	**95.08 ± 6.30**
		Text+Image	40.87 ± 1.54	45.82 ± 2.08	48.84 ± 1.91	43.01 ± 10.75	51.92 ± 4.77	34.99 ± 2.83	39.00 ± 3.57	94.75 ± 0.75	**97.02 ± 1.93**
	Yelp	Random	25.47 ± 0.28	24.36 ± 0.39	24.53 ± 1.13	24.48 ± 1.19	27.95 ± 0.11	**49.31 ± 0.12**	34.41 ± 1.01	9.60 ± 10.64	28.17 ± 0.33
		Text	51.05 ± 0.48	49.85 ± 0.63	41.65 ± 5.43	37.15 ± 0.28	47.51 ± 0.41	50.10 ± 0.44	36.94 ± 1.35	43.99 ± 14.74	**51.91 ± 0.58**
		Text+Image	50.58 ± 1.65	48.54 ± 0.69	38.16 ± 11.46	37.01 ± 0.40	46.80 ± 0.51	50.99 ± 0.23	33.13 ± 1.77	42.84 ± 12.07	**53.91 ± 0.94**
in.	GDELT	Random	25.12 ± 0.70	28.32 ± 0.39	27.00 ± 0.30	26.93 ± 0.62	27.49 ± 0.17	27.18 ± 2.41	29.53 ± 0.40	24.04 ± 5.95	**31.40 ± 0.21**
		Text	26.19 ± 2.26	29.08 ± 0.62	27.62 ± 0.07	27.76 ± 0.61	30.65 ± 1.72	27.09 ± 0.53	29.04 ± 0.97	33.72 ± 2.11	**38.95 ± 0.70**
		Text+Image	26.96 ± 1.22	29.43 ± 0.89	27.23 ± 0.12	27.70 ± 0.35	29.69 ± 0.55	30.54 ± 1.31	29.74 ± 0.73	34.12 ± 2.69	**39.08 ± 0.49**
	GMCT	Random	8.32 ± 2.54	9.62 ± 0.25	9.43 ± 0.51	10.10 ± 0.34	39.53 ± 0.34	10.17 ± 0.14	5.13 ± 0.08	**97.58 ± 1.77**	31.73 ± 19.82
		Text	13.15 ± 4.08	9.38 ± 0.23	9.58 ± 0.10	7.51 ± 2.33	39.16 ± 0.36	10.11 ± 0.18	5.14 ± 0.09	**94.26 ± 3.08**	71.47 ± 9.85
		Text+Image	11.71 ± 2.41	9.42 ± 0.10	9.70 ± 0.19	8.30 ± 1.87	39.36 ± 0.09	11.08 ± 0.28	5.11 ± 0.15	**97.81 ± 1.18**	79.29 ± 3.29
	Amazon - Industrial	Random	27.29 ± 3.54	26.00 ± 2.04	23.26 ± 4.98	30.74 ± 0.26	36.15 ± 2.29	29.89 ± 3.69	11.02 ± 4.21	**36.34 ± 21.51**	33.18 ± 2.03
		Text	34.85 ± 0.84	23.03 ± 0.70	30.60 ± 2.80	35.87 ± 0.64	39.27 ± 2.88	34.14 ± 1.25	9.88 ± 1.45	**62.35 ± 5.12**	61.00 ± 0.83
		Text+Image	36.56 ± 0.92	24.05 ± 1.52	31.24 ± 2.42	36.06 ± 0.97	38.47 ± 2.38	35.41 ± 2.33	11.26 ± 2.67	60.52 ± 6.87	**62.75 ± 8.84**
	Amazon - Office	Random	26.88 ± 6.58	21.81 ± 1.82	29.15 ± 2.02	6.47 ± 1.76	31.67 ± 5.72	24.48 ± 6.49	15.05 ± 3.60	**47.16 ± 9.69**	33.60 ± 5.73
		Text	29.37 ± 6.16	23.38 ± 3.36	27.74 ± 1.19	19.64 ± 12.04	39.05 ± 0.39	26.32 ± 7.87	14.20 ± 1.33	64.78 ± 5.43	**64.81 ± 7.93**
		Text+Image	24.08 ± 2.53	26.80 ± 0.86	26.38 ± 0.63	26.45 ± 9.14	38.61 ± 1.29	23.40 ± 2.90	14.36 ± 2.66	70.02 ± 1.85	**73.66 ± 3.00**
	Yelp	Random	22.83 ± 0.39	21.54 ± 0.68	22.08 ± 0.67	22.49 ± 0.65	24.99 ± 0.10	**39.81 ± 0.24**	24.78 ± 0.67	7.76 ± 8.73	25.07 ± 0.66
		Text	42.19 ± 0.50	40.83 ± 0.61	34.73 ± 4.08	31.26 ± 0.18	39.23 ± 0.37	40.48 ± 0.53	26.68 ± 1.26	36.87 ± 11.43	**43.12 ± 0.41**
		Text+Image	41.90 ± 1.38	39.65 ± 0.58	31.92 ± 8.53	31.26 ± 0.36	38.78 ± 0.48	41.27 ± 0.25	23.76 ± 0.83	36.34 ± 8.74	**44.44 ± 0.65**

Table 13: Comparison of multimodal fusion strategies on all metrics under the Text+Image setting (%). All methods share the DyGFormer backbone, and DyGFormer denotes naive feature concatenation. "*tr.*" and "*in.*" refer to the transductive and inductive setting, respectively. The best results are highlighted in bold, and the second-best results are underlined.

Task	Metric		Datasets	DyGFormer	Gated fusion	Modality attention	MoE fusion	TST-MF
LP	AP	*tr.*	GDELT	92.46 ± 0.26	92.74 ± 0.28	93.08 ± 0.10	92.92 ± 0.26	**93.98 ± 0.05**
			GMCT	95.54 ± 0.10	96.41 ± 0.17	96.30 ± 0.20	**96.59 ± 0.16**	89.08 ± 0.72
			Amazon - Industrial	92.56 ± 0.24	92.63 ± 0.26	92.61 ± 0.61	90.89 ± 2.25	**94.10 ± 0.12**
			Amazon - Office	89.00 ± 0.23	90.26 ± 1.41	90.73 ± 1.02	89.59 ± 0.94	**92.67 ± 0.10**
			Yelp	88.55 ± 5.51	91.25 ± 0.38	92.29 ± 0.17	88.56 ± 5.75	**92.65 ± 0.04**
		in.	GDELT	83.05 ± 0.28	83.29 ± 0.34	84.07 ± 0.31	84.14 ± 0.35	**85.44 ± 0.14**
			GMCT	**83.05 ± 3.36**	82.46 ± 1.83	81.92 ± 2.71	82.28 ± 1.28	65.49 ± 1.00
			Amazon - Industrial	76.33 ± 0.58	77.35 ± 0.83	76.44 ± 2.17	74.98 ± 1.07	**81.95 ± 0.54**
			Amazon - Office	68.64 ± 1.02	72.38 ± 4.30	74.01 ± 3.95	70.59 ± 3.07	**78.16 ± 0.31**
			Yelp	85.05 ± 5.64	87.73 ± 0.35	88.96 ± 0.21	85.00 ± 5.95	**89.39 ± 0.05**
	AUC-ROC	*tr.*	GDELT	91.93 ± 0.24	92.23 ± 0.26	92.53 ± 0.10	92.50 ± 0.30	**93.71 ± 0.08**
			GMCT	95.08 ± 0.27	96.30 ± 0.14	96.01 ± 0.06	**96.44 ± 0.10**	92.49 ± 0.66
			Amazon - Industrial	92.94 ± 0.29	92.76 ± 0.19	93.28 ± 0.40	90.53 ± 3.54	**94.26 ± 0.16**
			Amazon - Office	91.41 ± 0.16	91.84 ± 0.49	91.96 ± 0.22	91.48 ± 0.21	**93.23 ± 0.43**
			Yelp	88.84 ± 5.74	91.66 ± 0.36	92.68 ± 0.14	88.83 ± 6.03	**93.83 ± 0.02**
		in.	GDELT	81.29 ± 0.34	81.30 ± 0.31	82.52 ± 0.62	82.54 ± 0.59	**84.51 ± 0.11**
			GMCT	**82.39 ± 2.37**	81.48 ± 1.52	80.77 ± 2.11	81.25 ± 1.31	67.77 ± 1.48
			Amazon - Industrial	73.11 ± 1.06	73.95 ± 0.91	74.52 ± 1.86	71.24 ± 1.73	**80.00 ± 0.60**
			Amazon - Office	68.53 ± 0.07	70.85 ± 2.39	72.02 ± 2.50	69.30 ± 1.29	**75.38 ± 1.60**
			Yelp	84.57 ± 6.50	87.72 ± 0.30	88.94 ± 0.19	84.48 ± 6.88	**89.10 ± 0.05**
NR	Hits@1	*tr.*	GDELT	34.19 ± 4.03	30.80 ± 2.98	35.54 ± 0.74	34.60 ± 0.16	**38.31 ± 0.74**
			GMCT	93.29 ± 1.56	92.49 ± 2.20	94.52 ± 5.38	**95.18 ± 3.04**	73.27 ± 3.06
			Amazon - Industrial	62.78 ± 4.46	62.16 ± 3.94	64.09 ± 3.69	58.61 ± 11.32	**65.49 ± 6.91**
			Amazon - Office	69.52 ± 1.77	66.26 ± 3.30	66.26 ± 4.20	67.81 ± 5.56	**69.83 ± 1.73**
			Yelp	24.07 ± 7.01	27.26 ± 1.62	20.29 ± 17.48	23.74 ± 9.53	**29.94 ± 0.47**
		in.	GDELT	21.12 ± 1.80	20.87 ± 1.00	22.50 ± 0.81	22.40 ± 0.69	**23.80 ± 0.73**
			GMCT	65.20 ± 7.17	64.36 ± 9.81	76.14 ± 21.34	**76.48 ± 14.75**	32.64 ± 2.59
			Amazon - Industrial	23.93 ± 3.99	26.55 ± 6.03	29.34 ± 4.50	22.93 ± 8.17	**38.92 ± 1.99**
			Amazon - Office	44.47 ± 4.14	39.56 ± 5.77	40.49 ± 8.58	44.26 ± 7.46	**48.45 ± 2.43**
			Yelp	20.26 ± 5.04	22.49 ± 1.32	16.68 ± 13.99	19.73 ± 7.56	**24.58 ± 0.45**
	Hits@3	*tr.*	GDELT	52.06 ± 6.41	45.42 ± 6.02	54.69 ± 0.33	52.61 ± 1.30	**59.05 ± 0.36**
			GMCT	**100.00 ± 0.01**	**100.00 ± 0.00**	**100.00 ± 0.00**	99.88 ± 0.21	99.69 ± 0.69
			Amazon - Industrial	**97.08 ± 0.70**	95.44 ± 1.83	96.06 ± 1.24	94.09 ± 4.97	95.71 ± 2.92
			Amazon - Office	94.75 ± 0.75	93.23 ± 1.55	93.23 ± 0.89	92.73 ± 2.36	**97.02 ± 1.93**
			Yelp	42.84 ± 12.07	48.65 ± 2.34	36.21 ± 30.63	42.74 ± 15.26	**53.91 ± 0.94**
		in.	GDELT	34.12 ± 2.69	32.92 ± 2.90	36.85 ± 0.69	36.43 ± 0.61	**39.08 ± 0.49**
			GMCT	97.81 ± 1.18	97.91 ± 0.85	97.96 ± 2.26	**98.57 ± 1.10**	79.29 ± 3.29
			Amazon - Industrial	60.52 ± 6.87	59.91 ± 6.30	**63.57 ± 4.87**	54.92 ± 18.48	62.75 ± 8.84
			Amazon - Office	70.02 ± 1.85	66.42 ± 3.32	65.33 ± 4.67	67.40 ± 6.83	**73.66 ± 3.00**
			Yelp	36.34 ± 8.74	40.28 ± 1.95	30.96 ± 23.21	35.41 ± 12.50	**44.44 ± 0.65**
	Hits@10	*tr.*	GDELT	71.34 ± 7.82	62.18 ± 6.57	74.46 ± 1.58	71.33 ± 2.94	**81.62 ± 0.33**
			GMCT	**100.00 ± 0.00**	**100.00 ± 0.00**	**100.00 ± 0.00**	99.96 ± 0.07	**100.00 ± 0.00**
			Amazon - Industrial	**100.00 ± 0.00**	**100.00 ± 0.00**	**100.00 ± 0.01**	99.66 ± 0.58	**100.00 ± 0.00**
			Amazon - Office	**99.99 ± 0.01**	99.25 ± 1.27	99.98 ± 0.03	98.45 ± 1.52	**99.99 ± 0.01**
			Yelp	67.54 ± 14.55	75.02 ± 1.72	57.09 ± 35.30	66.89 ± 17.38	**79.11 ± 0.49**
		in.	GDELT	52.50 ± 3.86	46.75 ± 7.15	54.62 ± 0.68	54.57 ± 1.51	**59.23 ± 0.58**
			GMCT	**100.00 ± 0.01**	**100.00 ± 0.00**	**100.00 ± 0.00**	99.88 ± 0.21	**100.00 ± 0.00**
			Amazon - Industrial	99.51 ± 0.19	98.25 ± 1.59	98.65 ± 0.86	96.51 ± 4.64	**99.94 ± 0.99**
			Amazon - Office	98.10 ± 0.61	96.58 ± 1.56	96.36 ± 1.14	95.89 ± 2.37	**98.99 ± 0.31**
			Yelp	59.48 ± 12.06	65.29 ± 1.92	55.40 ± 22.19	58.69 ± 15.22	**67.84 ± 0.48**
EC	Precision		GDELT	47.54 ± 0.10	48.49 ± 0.71	48.85 ± 0.89	47.74 ± 0.25	**53.56 ± 1.70**
			GMCT	56.77 ± 1.72	**57.69 ± 6.60**	56.36 ± 3.38	54.81 ± 2.50	53.45 ± 1.57
			Amazon - Industrial	**59.38 ± 1.85**	52.81 ± 0.68	52.53 ± 0.57	53.50 ± 0.20	50.32 ± 1.86
			Amazon - Office	**56.02 ± 0.50**	53.98 ± 0.86	53.73 ± 0.12	53.91 ± 0.44	54.42 ± 0.02
			Yelp	53.99 ± 0.18	52.65 ± 0.30	53.91 ± 0.49	52.71 ± 1.03	**55.13 ± 0.87**
	Recall		GDELT	63.93 ± 0.08	63.78 ± 0.64	**64.03 ± 0.41**	**64.03 ± 0.12**	63.77 ± 0.07
			GMCT	**67.47 ± 0.14**	66.63 ± 0.61	67.19 ± 0.00	66.17 ± 0.58	67.18 ± 0.00
			Amazon - Industrial	**65.20 ± 0.11**	65.02 ± 0.17	64.51 ± 0.62	65.06 ± 0.12	**65.20 ± 0.01**
			Amazon - Office	**66.90 ± 0.03**	66.66 ± 0.08	66.73 ± 0.10	66.72 ± 0.03	66.59 ± 0.01
			Yelp	58.53 ± 0.19	57.84 ± 0.14	58.21 ± 0.63	57.90 ± 0.66	**58.72 ± 1.11**
	F1		GDELT	52.81 ± 0.07	53.26 ± 0.17	**53.70 ± 0.12**	52.87 ± 0.30	53.64 ± 0.03
			GMCT	**59.89 ± 0.06**	55.26 ± 1.12	54.02 ± 0.02	55.91 ± 0.50	56.32 ± 0.21
			Amazon - Industrial	**55.72 ± 0.44**	55.39 ± 0.02	55.58 ± 0.13	55.02 ± 0.21	54.89 ± 0.12
			Amazon - Office	**57.40 ± 0.34**	56.25 ± 0.22	56.33 ± 0.07	56.07 ± 0.03	56.97 ± 0.02
			Yelp	52.34 ± 0.66	51.05 ± 0.25	51.64 ± 0.83	51.15 ± 1.30	**52.43 ± 2.95**

Table 14: Link prediction results (%) on the complete GDELT and GMCT datasets. "*tr.*" and "*in.*" refer to the transductive and inductive settings, respectively. AP refers to Average Precision. The best results are highlighted in bold and the second-best results are underlined.

	Datasets	Metric	Feature	TGAT	CAWN	TCL	GraphMixer	DyGFormer	TST-MF
tr.	GDELT	AP	Random	96.19 ± 0.27	97.04 ± 0.04	96.20 ± 0.11	96.06 ± 0.20	97.25 ± 0.03	**97.36 ± 0.03**
			Text	97.40 ± 0.07	97.37 ± 0.13	96.09 ± 0.15	96.96 ± 0.11	97.71 ± 0.02	**97.87 ± 0.02**
			Text+Image	97.42 ± 0.07	97.27 ± 0.28	96.19 ± 0.25	97.03 ± 0.08	97.62 ± 0.09	**97.85 ± 0.04**
		AUC-ROC	Random	95.97 ± 0.26	96.91 ± 0.04	96.02 ± 0.15	95.94 ± 0.14	97.01 ± 0.07	**97.18 ± 0.03**
			Text	97.30 ± 0.07	97.31 ± 0.13	95.93 ± 0.18	96.87 ± 0.15	97.63 ± 0.03	**97.80 ± 0.02**
			Text+Image	97.33 ± 0.08	97.20 ± 0.29	96.06 ± 0.28	96.97 ± 0.08	97.53 ± 0.11	**97.78 ± 0.06**
	GMCT	AP	Random	72.24 ± 0.43	74.92 ± 0.07	72.11 ± 3.38	74.96 ± 0.01	**78.65 ± 0.05**	76.11 ± 0.55
			Text	79.22 ± 12.22	**93.67 ± 0.80**	73.71 ± 0.94	79.99 ± 3.53	78.90 ± 0.20	82.51 ± 9.11
			Text+Image	93.16 ± 0.67	**94.08 ± 0.51**	75.21 ± 0.14	83.86 ± 2.98	80.00 ± 0.05	87.90 ± 8.91
		AUC-ROC	Random	74.25 ± 0.36	76.26 ± 0.02	72.94 ± 2.51	76.34 ± 0.02	**77.38 ± 0.27**	76.71 ± 0.05
			Text	80.69 ± 11.17	**93.97 ± 0.66**	74.75 ± 1.87	80.75 ± 2.35	77.88 ± 0.08	83.98 ± 8.18
			Text+Image	93.58 ± 0.73	**94.38 ± 0.39**	77.17 ± 0.58	84.61 ± 2.57	80.45 ± 0.13	88.97 ± 8.06
in.	GDELT	AP	Random	88.18 ± 0.23	88.33 ± 0.10	88.73 ± 0.13	87.33 ± 0.66	**91.92 ± 0.10**	90.86 ± 0.41
			Text	91.85 ± 0.05	90.85 ± 0.38	88.57 ± 0.26	90.70 ± 0.07	93.23 ± 0.11	**93.26 ± 0.03**
			Text+Image	91.78 ± 0.19	90.55 ± 0.80	88.68 ± 0.27	90.71 ± 0.19	92.79 ± 0.30	**93.27 ± 0.06**
		AUC-ROC	Random	87.47 ± 0.41	87.82 ± 0.31	87.85 ± 0.14	86.69 ± 0.63	**90.81 ± 0.19**	90.29 ± 0.37
			Text	91.72 ± 0.09	90.55 ± 0.38	87.73 ± 0.26	90.37 ± 0.08	92.80 ± 0.11	**92.94 ± 0.06**
			Text+Image	91.70 ± 0.20	90.21 ± 0.98	87.91 ± 0.34	90.34 ± 0.28	92.21 ± 0.42	**92.91 ± 0.06**
	GMCT	AP	Random	58.89 ± 0.23	60.81 ± 0.12	59.31 ± 3.05	60.75 ± 0.02	**62.25 ± 0.34**	62.08 ± 1.05
			Text	68.61 ± 16.96	**87.77 ± 1.56**	59.92 ± 0.61	68.31 ± 5.78	62.22 ± 0.49	70.66 ± 14.45
			Text+Image	87.26 ± 0.52	**88.41 ± 0.96**	61.06 ± 0.77	73.21 ± 5.15	64.04 ± 0.59	78.61 ± 13.54
		AUC-ROC	Random	59.29 ± 0.23	60.94 ± 0.14	58.54 ± 1.67	60.83 ± 0.02	60.40 ± 0.27	**61.41 ± 0.19**
			Text	69.00 ± 16.89	**88.25 ± 1.31**	60.08 ± 1.15	67.99 ± 4.83	60.61 ± 0.24	71.32 ± 14.29
			Text+Image	88.00 ± 0.84	**88.80 ± 0.71**	61.19 ± 0.39	73.47 ± 4.76	62.54 ± 0.20	79.41 ± 13.59

Table 15: Edge classification results (%) on the complete GDELT and GMCT datasets. The best results are highlighted in bold and the second-best results are underlined.

Datasets	Metric	Feature	TGAT	CAWN	TCL	GraphMixer	DyGFormer	TST-MF
GDELT	Precision	Random	60.92 ± 2.70	**61.85 ± 0.58**	46.09 ± 4.48	60.68 ± 0.94	57.56 ± 0.26	61.23 ± 0.14
		Text	49.57 ± 0.25	**63.85 ± 1.26**	45.92 ± 4.34	58.91 ± 10.07	60.51 ± 3.81	61.79 ± 0.67
		Text+Image	49.43 ± 0.13	**62.49 ± 1.92**	43.50 ± 4.47	55.61 ± 5.35	59.74 ± 4.03	61.13 ± 2.25
	Recall	Random	64.69 ± 0.03	65.01 ± 0.02	64.43 ± 0.40	**65.09 ± 0.06**	**65.09 ± 0.02**	64.18 ± 0.34
		Text	64.39 ± 0.35	65.04 ± 0.04	64.27 ± 0.35	65.00 ± 0.10	65.20 ± 0.09	**65.41 ± 0.09**
		Text+Image	64.50 ± 0.21	64.94 ± 0.02	64.20 ± 0.39	64.97 ± 0.18	65.15 ± 0.22	**65.48 ± 0.01**
	F1	Random	54.73 ± 0.20	55.24 ± 0.07	52.12 ± 2.00	**55.66 ± 0.02**	55.17 ± 0.14	54.91 ± 0.21
		Text	53.84 ± 3.37	55.21 ± 0.10	51.41 ± 1.73	**55.91 ± 0.12**	55.45 ± 0.16	55.63 ± 0.14
		Text+Image	55.78 ± 0.14	55.04 ± 0.07	50.95 ± 1.80	**55.98 ± 0.09**	55.32 ± 0.55	55.73 ± 0.08
GMCT	Precision	Random	57.87 ± 1.45	60.95 ± 0.18	55.46 ± 3.12	60.81 ± 0.11	**61.32 ± 0.11**	53.79 ± 1.19
		Text	58.52 ± 0.69	**61.06 ± 0.05**	53.77 ± 0.41	61.00 ± 0.15	61.00 ± 0.95	55.39 ± 0.33
		Text+Image	60.06 ± 2.20	61.20 ± 0.11	51.75 ± 1.29	60.99 ± 0.07	**62.17 ± 0.89**	54.56 ± 0.80
	Recall	Random	66.59 ± 0.36	**68.09 ± 0.03**	66.65 ± 0.77	67.18 ± 0.66	**68.09 ± 0.02**	64.51 ± 0.89
		Text	66.74 ± 0.34	68.12 ± 0.03	66.29 ± 0.11	67.12 ± 0.83	**68.19 ± 0.10**	65.99 ± 0.04
		Text+Image	67.48 ± 1.05	68.15 ± 0.08	66.07 ± 0.13	66.96 ± 0.77	**68.23 ± 0.10**	66.02 ± 0.03
	F1	Random	57.87 ± 0.96	61.05 ± 0.19	56.09 ± 4.78	**62.27 ± 0.54**	61.38 ± 0.25	55.43 ± 0.45
		Text	58.58 ± 0.46	60.98 ± 0.11	54.77 ± 1.43	**62.49 ± 0.49**	62.18 ± 0.18	54.58 ± 0.33
		Text+Image	61.24 ± 1.57	61.17 ± 0.55	53.75 ± 2.17	**62.67 ± 0.16**	62.49 ± 0.03	54.33 ± 0.81

Table 16: Node retrieval results (%) on the complete GDELT and GMCT datasets. "*tr.*" and "*in.*" refer to the transductive and inductive settings, respectively. The best results are highlighted in bold and the second-best results are underlined.

	Datasets	Metric	Feature	TGAT	CAWN	TCL	GraphMixer	DyGFormer	TST-MF
tr.	GDELT	Hits@1	Random	51.39 ± 1.29	55.65 ± 0.32	50.30 ± 0.38	50.14 ± 0.95	31.33 ± 7.78	**58.59 ± 0.05**
			Text	58.39 ± 0.65	55.33 ± 1.70	50.00 ± 0.61	54.93 ± 0.39	59.63 ± 0.19	**61.35 ± 0.46**
			Text+Image	58.48 ± 0.19	55.66 ± 2.01	49.96 ± 1.00	55.08 ± 0.20	52.54 ± 7.70	**61.18 ± 0.35**
		Hits@3	Random	73.00 ± 2.03	76.29 ± 0.13	73.47 ± 0.25	70.30 ± 2.69	45.47 ± 6.98	**79.51 ± 0.11**
			Text	79.73 ± 0.49	76.93 ± 1.73	72.49 ± 0.49	76.69 ± 0.39	79.91 ± 0.34	**82.26 ± 0.25**
			Text+Image	79.79 ± 0.20	77.25 ± 1.64	73.37 ± 0.91	76.84 ± 0.43	71.04 ± 9.37	**82.01 ± 0.28**
		Hits@10	Random	88.47 ± 0.86	91.07 ± 0.68	89.17 ± 0.10	88.23 ± 0.89	70.42 ± 5.32	**93.15 ± 0.18**
			Text	92.83 ± 0.12	92.12 ± 0.98	88.54 ± 0.75	91.64 ± 0.40	92.37 ± 0.49	**94.42 ± 0.22**
			Text+Image	92.99 ± 0.26	92.34 ± 1.12	89.02 ± 1.11	91.81 ± 0.40	87.47 ± 5.18	**94.17 ± 0.07**
	GMCT	Hits@1	Random	6.33 ± 0.50	5.34 ± 0.20	5.82 ± 3.69	5.30 ± 0.03	**45.72 ± 1.75**	24.34 ± 1.05
			Text	32.75 ± 2.61	33.40 ± 3.54	3.65 ± 1.32	11.64 ± 5.25	**44.03 ± 34.72**	42.68 ± 17.21
			Text+Image	30.89 ± 1.93	34.37 ± 2.58	6.24 ± 2.26	15.60 ± 3.18	44.14 ± 4.06	**51.51 ± 21.19**
		Hits@3	Random	14.74 ± 1.05	15.77 ± 0.12	14.27 ± 5.33	15.83 ± 0.03	**94.41 ± 0.12**	70.75 ± 0.59
			Text	57.31 ± 2.50	57.56 ± 4.04	12.43 ± 2.06	23.92 ± 6.72	65.21 ± 44.96	**86.92 ± 3.86**
			Text+Image	55.46 ± 3.01	58.96 ± 2.63	16.01 ± 1.55	29.86 ± 5.16	**94.27 ± 1.62**	88.09 ± 9.65
		Hits@10	Random	33.33 ± 1.52	41.03 ± 0.02	35.71 ± 3.62	41.22 ± 0.07	99.78 ± 0.37	**99.94 ± 0.01**
			Text	80.16 ± 1.15	80.51 ± 1.97	35.89 ± 5.71	47.38 ± 4.84	76.92 ± 38.69	**99.98 ± 0.03**
			Text+Image	78.98 ± 2.24	81.42 ± 1.08	39.52 ± 2.17	54.02 ± 5.78	99.91 ± 0.16	**100.00 ± 0.00**
in.	GDELT	Hits@1	Random	22.36 ± 1.82	25.65 ± 1.08	24.20 ± 0.27	21.78 ± 0.68	21.26 ± 6.60	**30.10 ± 0.68**
			Text	29.34 ± 0.16	28.62 ± 0.81	24.31 ± 0.54	28.34 ± 0.19	**36.87 ± 0.30**	35.76 ± 0.12
			Text+Image	29.06 ± 0.38	28.29 ± 1.02	23.51 ± 0.84	28.64 ± 0.53	33.50 ± 3.64	**35.91 ± 0.48**
		Hits@3	Random	42.45 ± 2.79	41.29 ± 1.18	44.64 ± 0.34	40.61 ± 0.82	26.74 ± 8.93	**48.22 ± 1.68**
			Text	51.21 ± 0.20	46.62 ± 1.29	44.85 ± 0.27	49.44 ± 0.30	57.43 ± 0.60	**57.99 ± 0.19**
			Text+Image	50.70 ± 0.72	46.67 ± 1.69	44.47 ± 0.45	49.36 ± 0.47	49.57 ± 7.72	**58.16 ± 0.25**
		Hits@10	Random	68.83 ± 1.13	65.36 ± 0.99	70.53 ± 0.30	65.29 ± 2.58	41.62 ± 7.13	**74.38 ± 2.24**
			Text	76.65 ± 0.28	71.80 ± 1.43	69.97 ± 0.50	74.15 ± 0.37	79.90 ± 0.38	**80.89 ± 0.02**
			Text+Image	76.65 ± 0.59	72.28 ± 1.70	70.35 ± 0.59	74.33 ± 0.52	68.74 ± 10.16	**80.81 ± 0.03**
	GMCT	Hits@1	Random	3.10 ± 0.35	2.65 ± 0.23	2.99 ± 2.20	2.55 ± 0.03	**13.91 ± 1.25**	8.61 ± 3.15
			Text	19.99 ± 2.34	20.42 ± 3.00	1.94 ± 0.47	6.49 ± 3.19	**31.75 ± 34.51**	21.93 ± 13.40
			Text+Image	18.30 ± 1.17	20.89 ± 2.14	3.08 ± 1.19	9.00 ± 2.12	11.29 ± 3.11	**29.34 ± 14.23**
		Hits@3	Random	7.52 ± 0.51	6.59 ± 0.21	7.09 ± 3.48	6.60 ± 0.09	**55.79 ± 3.14**	26.60 ± 1.74
			Text	39.89 ± 3.02	39.66 ± 4.73	5.60 ± 0.73	13.56 ± 5.81	**53.33 ± 41.31**	45.05 ± 18.51
			Text+Image	37.65 ± 2.32	40.60 ± 3.04	7.45 ± 2.10	18.05 ± 4.02	53.98 ± 6.45	**54.72 ± 23.40**
		Hits@10	Random	18.67 ± 0.81	20.30 ± 0.07	19.13 ± 3.42	20.25 ± 0.07	**99.48 ± 0.54**	91.87 ± 0.45
			Text	65.56 ± 1.66	64.85 ± 3.60	19.08 ± 1.39	28.94 ± 7.31	69.99 ± 47.38	**97.36 ± 1.88**
			Text+Image	63.75 ± 2.14	65.73 ± 1.92	20.05 ± 1.14	35.41 ± 6.52	**99.76 ± 0.24**	97.69 ± 1.91

Table 17: Average Precision (%) for link prediction under different missing-modality filling strategies. "*tr.*" and "*in.*" refer to the transductive and inductive settings, respectively. The best results are highlighted in bold and the second-best results are underlined.

	Datasets	Feature	TGAT	CAWN	TCL	GraphMixer	DyGFormer	TST-MF
tr.	GDELT	Zero Filling	90.24 ± 0.64	93.08 ± 0.18	91.67 ± 0.14	91.18 ± 0.32	92.33 ± 0.17	**93.93 ± 0.11**
		Mean Filling	90.32 ± 0.42	92.94 ± 0.09	91.69 ± 0.05	91.27 ± 0.32	92.26 ± 0.32	**94.07 ± 0.01**
		Learned Embedding	89.96 ± 0.63	92.94 ± 0.48	91.71 ± 0.20	91.39 ± 0.21	92.30 ± 0.18	**94.01 ± 0.16**
		Duplication	90.76 ± 0.34	93.08 ± 0.25	91.47 ± 0.19	91.38 ± 0.20	92.46 ± 0.26	**93.98 ± 0.05**
	GMCT	Zero Filling	82.47 ± 3.32	79.58 ± 0.08	78.63 ± 0.10	79.57 ± 0.38	**95.60 ± 0.27**	87.59 ± 0.99
		Mean Filling	84.19 ± 2.11	79.59 ± 0.03	78.11 ± 0.23	79.54 ± 0.22	**96.16 ± 0.32**	88.55 ± 1.57
		Learned Embedding	81.47 ± 2.19	79.32 ± 0.18	78.55 ± 0.53	79.15 ± 0.13	**96.29 ± 0.41**	89.26 ± 1.57
		Duplication	81.62 ± 2.33	79.65 ± 0.05	78.50 ± 0.22	79.62 ± 0.48	**95.54 ± 0.10**	89.08 ± 0.72
	Amazon - Industrial	Zero Filling	92.73 ± 0.73	92.29 ± 0.23	92.76 ± 0.49	93.17 ± 1.38	92.80 ± 0.03	**93.86 ± 0.35**
		Mean Filling	93.82 ± 0.51	92.43 ± 0.23	92.74 ± 0.45	92.61 ± 2.71	92.91 ± 0.58	**94.07 ± 0.09**
		Learned Embedding	93.88 ± 0.44	92.36 ± 0.19	92.73 ± 0.52	**93.97 ± 0.03**	91.48 ± 0.97	93.94 ± 0.24
		Duplication	93.97 ± 0.18	92.21 ± 0.18	93.11 ± 0.36	93.96 ± 0.37	92.56 ± 0.24	**94.10 ± 0.12**
	Amazon - Office	Zero Filling	91.77 ± 0.82	90.55 ± 0.77	90.87 ± 1.46	**92.33 ± 1.04**	89.41 ± 0.91	92.07 ± 0.48
		Mean Filling	**92.20 ± 0.58**	91.24 ± 1.25	91.65 ± 0.30	91.18 ± 2.62	89.00 ± 2.07	91.94 ± 0.50
		Learned Embedding	92.03 ± 0.89	90.58 ± 1.03	91.50 ± 0.48	**92.39 ± 1.50**	89.25 ± 0.65	91.93 ± 0.21
		Duplication	91.23 ± 0.23	91.64 ± 0.29	91.73 ± 0.19	91.06 ± 2.56	89.00 ± 0.23	**92.67 ± 0.10**
	Yelp	Zero Filling	92.26 ± 0.32	92.06 ± 0.13	87.45 ± 5.84	86.96 ± 0.13	88.43 ± 5.53	**92.47 ± 0.02**
		Mean Filling	91.95 ± 0.45	91.96 ± 0.17	82.61 ± 3.02	87.00 ± 0.07	88.07 ± 5.31	**92.48 ± 0.04**
		Learned Embedding	92.03 ± 0.37	91.97 ± 0.19	87.83 ± 5.89	86.92 ± 0.02	82.04 ± 0.09	**92.47 ± 0.04**
		Duplication	92.22 ± 0.42	91.94 ± 0.13	87.44 ± 5.74	87.26 ± 0.19	88.55 ± 5.51	**92.65 ± 0.04**
in.	GDELT	Zero Filling	77.86 ± 0.36	80.99 ± 0.27	78.39 ± 0.19	78.34 ± 0.62	82.45 ± 0.79	**85.40 ± 0.17**
		Mean Filling	77.60 ± 0.36	80.71 ± 0.30	78.45 ± 0.13	78.42 ± 0.35	82.11 ± 0.65	**85.37 ± 0.13**
		Learned Embedding	77.65 ± 0.37	80.88 ± 0.71	78.55 ± 0.08	78.57 ± 0.29	82.48 ± 0.64	**85.20 ± 0.21**
		Duplication	78.38 ± 0.67	81.02 ± 0.41	78.33 ± 0.29	78.58 ± 0.34	83.05 ± 0.28	**85.44 ± 0.14**
	GMCT	Zero Filling	68.11 ± 5.19	62.31 ± 0.41	61.31 ± 0.23	62.62 ± 0.74	**82.44 ± 2.53**	62.35 ± 0.69
		Mean Filling	70.94 ± 3.99	62.42 ± 0.05	61.03 ± 0.29	62.79 ± 0.78	**82.01 ± 0.74**	63.07 ± 2.06
		Learned Embedding	65.97 ± 4.21	61.87 ± 0.12	61.30 ± 0.27	61.87 ± 0.16	**83.76 ± 1.99**	64.54 ± 1.98
		Duplication	66.02 ± 4.08	62.30 ± 0.18	60.90 ± 0.23	62.83 ± 0.89	**83.05 ± 3.36**	65.49 ± 1.00
	Amazon - Industrial	Zero Filling	78.31 ± 2.09	77.08 ± 0.83	78.66 ± 1.76	80.39 ± 5.07	77.24 ± 0.60	**81.62 ± 0.92**
		Mean Filling	81.66 ± 1.82	77.42 ± 0.94	78.47 ± 1.42	78.24 ± 8.89	77.24 ± 2.01	**82.14 ± 0.22**
		Learned Embedding	81.89 ± 1.39	77.17 ± 0.86	78.22 ± 2.11	**82.90 ± 0.32**	75.11 ± 0.99	81.79 ± 0.36
		Duplication	82.29 ± 0.70	76.74 ± 0.86	79.70 ± 1.09	**82.75 ± 1.15**	76.33 ± 0.58	81.95 ± 0.54
	Amazon - Office	Zero Filling	76.81 ± 2.47	73.25 ± 2.43	73.73 ± 5.01	**78.50 ± 2.73**	70.80 ± 3.64	77.21 ± 1.57
		Mean Filling	**77.92 ± 1.72**	75.13 ± 3.59	76.39 ± 1.24	75.30 ± 7.14	69.62 ± 6.07	77.07 ± 1.60
		Learned Embedding	77.43 ± 2.63	73.27 ± 3.03	75.89 ± 1.48	**78.66 ± 4.27**	73.10 ± 2.44	77.10 ± 0.57
		Duplication	74.85 ± 0.80	76.70 ± 0.84	76.61 ± 0.54	75.02 ± 7.04	68.64 ± 1.02	**78.16 ± 0.31**
	Yelp	Zero Filling	88.93 ± 0.39	88.51 ± 0.18	83.84 ± 5.89	83.08 ± 0.17	84.82 ± 5.68	**89.14 ± 0.06**
		Mean Filling	88.52 ± 0.55	88.38 ± 0.21	78.97 ± 3.03	83.17 ± 0.10	84.48 ± 5.47	**89.13 ± 0.13**
		Learned Embedding	88.60 ± 0.47	88.40 ± 0.27	84.28 ± 5.91	83.07 ± 0.06	78.30 ± 0.15	**89.12 ± 0.05**
		Duplication	88.93 ± 0.53	88.38 ± 0.20	83.91 ± 5.83	83.45 ± 0.16	85.05 ± 5.64	**89.39 ± 0.05**

Table 18: Precision (%) for edge classification under different missing-modality filling strategies. The best results are highlighted in bold and the second-best results are underlined.

Datasets	Feature	TGAT	CAWN	TCL	GraphMixer	DyGFormer	TST-MF
GDELT	Zero Filling	47.83 ± 0.31	**51.86 ± 3.13**	40.77 ± 0.00	48.19 ± 0.50	47.54 ± 0.16	49.39 ± 1.96
	Mean Filling	47.93 ± 0.17	50.12 ± 3.53	42.44 ± 2.89	47.85 ± 0.30	47.54 ± 0.08	**51.56 ± 0.47**
	Learned Embedding	47.96 ± 0.11	**50.77 ± 2.57**	44.51 ± 6.47	48.27 ± 0.23	47.46 ± 0.14	50.54 ± 1.90
	Duplication	48.15 ± 0.09	51.62 ± 5.99	40.77 ± 0.00	48.08 ± 0.31	47.54 ± 0.10	**53.56 ± 1.70**
GMCT	Zero Filling	55.39 ± 0.40	56.55 ± 2.10	54.90 ± 0.51	58.51 ± 0.02	**59.22 ± 0.31**	53.83 ± 1.21
	Mean Filling	55.16 ± 0.28	55.46 ± 0.09	54.28 ± 0.28	58.35 ± 0.25	**58.56 ± 2.63**	52.31 ± 0.91
	Learned Embedding	54.89 ± 0.14	55.43 ± 0.03	54.37 ± 0.03	**58.57 ± 0.61**	57.60 ± 3.35	50.79 ± 4.92
	Duplication	55.45 ± 0.29	56.64 ± 2.07	54.40 ± 0.19	**58.58 ± 0.76**	56.77 ± 1.72	53.45 ± 1.57
Amazon - Industrial	Zero Filling	48.68 ± 0.35	55.31 ± 5.62	47.77 ± 0.59	56.90 ± 1.47	**58.62 ± 3.79**	50.51 ± 1.83
	Mean Filling	48.82 ± 0.02	55.48 ± 0.40	47.57 ± 0.31	**56.20 ± 1.83**	55.31 ± 1.11	52.17 ± 0.33
	Learned Embedding	48.92 ± 0.28	56.03 ± 6.50	47.45 ± 0.36	57.91 ± 0.17	**58.44 ± 2.61**	51.95 ± 1.01
	Duplication	48.50 ± 0.12	**59.83 ± 2.25**	48.05 ± 0.42	58.12 ± 0.51	59.38 ± 1.85	50.32 ± 1.86
Amazon - Office	Zero Filling	55.17 ± 0.55	**57.85 ± 2.55**	47.76 ± 0.46	56.21 ± 0.83	56.37 ± 0.51	54.22 ± 2.12
	Mean Filling	54.39 ± 1.01	**56.81 ± 0.62**	47.05 ± 0.76	56.03 ± 0.80	56.05 ± 0.57	53.61 ± 0.45
	Learned Embedding	54.70 ± 1.40	56.49 ± 0.79	49.03 ± 5.77	**57.17 ± 1.59**	56.58 ± 0.87	55.63 ± 2.45
	Duplication	53.51 ± 0.27	**57.67 ± 0.90**	50.40 ± 5.15	56.58 ± 0.86	56.02 ± 0.50	54.42 ± 0.02
Yelp	Zero Filling	53.10 ± 1.10	52.30 ± 0.47	50.21 ± 3.17	52.91 ± 0.58	**54.35 ± 0.18**	52.51 ± 0.42
	Mean Filling	50.76 ± 1.02	51.65 ± 1.27	44.25 ± 5.38	52.34 ± 0.89	**54.04 ± 0.49**	52.80 ± 0.22
	Learned Embedding	49.88 ± 1.74	52.70 ± 0.22	42.54 ± 4.72	52.59 ± 0.65	**54.15 ± 0.09**	53.00 ± 0.40
	Duplication	53.47 ± 0.13	52.58 ± 0.90	51.97 ± 0.84	52.91 ± 0.55	53.99 ± 0.18	**55.13 ± 0.87**

Table 19: Recall (%) for edge classification under different missing-modality filling strategies. The best results are highlighted in bold and the second-best results are underlined.

Datasets	Feature	TGAT	CAWN	TCL	GraphMixer	DyGFormer	TST-MF
GDELT	Zero Filling	64.01 ± 0.04	64.28 ± 0.09	63.85 ± 0.00	**64.29 ± 0.23**	63.91 ± 0.04	63.89 ± 0.64
	Mean Filling	63.98 ± 0.10	**64.18 ± 0.09**	63.85 ± 0.01	64.16 ± 0.13	63.92 ± 0.05	63.99 ± 0.16
	Learned Embedding	64.01 ± 0.03	64.26 ± 0.12	63.86 ± 0.01	**64.34 ± 0.09**	63.88 ± 0.08	64.13 ± 0.09
	Duplication	64.05 ± 0.03	64.24 ± 0.12	63.85 ± 0.00	**64.26 ± 0.11**	63.93 ± 0.08	63.77 ± 0.07
GMCT	Zero Filling	67.17 ± 0.08	**67.52 ± 0.13**	66.26 ± 1.27	67.37 ± 0.35	**67.52 ± 0.17**	64.27 ± 3.20
	Mean Filling	67.00 ± 0.27	67.45 ± 0.08	66.98 ± 0.19	67.55 ± 0.08	**67.62 ± 0.11**	64.59 ± 1.93
	Learned Embedding	66.81 ± 0.47	**67.54 ± 0.07**	66.50 ± 0.81	67.46 ± 0.19	67.35 ± 0.18	64.92 ± 1.97
	Duplication	67.36 ± 0.15	**67.60 ± 0.23**	66.61 ± 0.54	67.48 ± 0.35	67.47 ± 0.14	67.18 ± 0.00
Amazon - Industrial	Zero Filling	64.45 ± 0.13	65.10 ± 0.08	64.23 ± 0.70	64.43 ± 0.13	**65.40 ± 0.07**	63.99 ± 0.39
	Mean Filling	64.90 ± 0.47	64.90 ± 0.13	64.17 ± 0.44	64.33 ± 0.04	**65.39 ± 0.11**	64.09 ± 0.41
	Learned Embedding	64.56 ± 1.20	65.03 ± 0.18	64.20 ± 0.41	64.20 ± 0.11	**65.48 ± 0.22**	64.12 ± 0.52
	Duplication	64.81 ± 0.44	64.96 ± 0.21	64.60 ± 0.55	64.32 ± 0.06	**65.20 ± 0.11**	**65.20 ± 0.01**
Amazon - Office	Zero Filling	66.53 ± 0.20	66.71 ± 0.02	65.70 ± 0.17	66.02 ± 0.36	**66.95 ± 0.02**	66.27 ± 0.38
	Mean Filling	66.62 ± 0.15	66.71 ± 0.05	65.65 ± 0.08	66.05 ± 0.32	**66.94 ± 0.04**	66.37 ± 0.08
	Learned Embedding	66.52 ± 0.27	66.71 ± 0.03	65.85 ± 0.83	66.12 ± 0.39	**66.87 ± 0.06**	66.07 ± 0.28
	Duplication	66.09 ± 0.17	66.70 ± 0.06	66.02 ± 0.53	66.06 ± 0.48	**66.90 ± 0.03**	66.59 ± 0.01
Yelp	Zero Filling	57.81 ± 0.80	57.57 ± 0.35	56.71 ± 1.44	57.83 ± 0.66	**58.49 ± 0.06**	57.64 ± 0.32
	Mean Filling	56.13 ± 0.07	56.80 ± 1.23	53.84 ± 1.92	57.36 ± 0.52	**58.53 ± 0.14**	57.72 ± 0.11
	Learned Embedding	55.65 ± 0.79	57.79 ± 0.23	53.08 ± 1.72	57.42 ± 0.39	**58.54 ± 0.07**	57.77 ± 0.19
	Duplication	57.93 ± 0.17	57.54 ± 0.48	57.63 ± 0.22	57.72 ± 0.44	58.53 ± 0.19	**58.72 ± 1.11**

Table 20: F1 (%) for edge classification under different missing-modality filling strategies. The best results are highlighted in bold and the second-best results are underlined.

Datasets	Feature	TGAT	CAWN	TCL	GraphMixer	DyGFormer	TST-MF
GDELT	Zero Filling	50.86 ± 0.21	**53.40 ± 0.11**	49.77 ± 0.00	52.62 ± 0.66	52.85 ± 0.22	53.27 ± 0.34
	Mean Filling	50.52 ± 0.59	**53.38 ± 0.09**	49.80 ± 0.05	52.24 ± 0.30	52.87 ± 0.03	53.22 ± 0.74
	Learned Embedding	50.69 ± 0.08	**53.26 ± 0.16**	49.79 ± 0.04	52.95 ± 0.19	52.75 ± 0.14	53.09 ± 0.55
	Duplication	50.81 ± 0.16	53.45 ± 0.10	49.77 ± 0.00	52.66 ± 0.86	52.81 ± 0.07	**53.64 ± 0.03**
GMCT	Zero Filling	59.53 ± 0.58	59.19 ± 0.09	59.05 ± 0.64	**60.02 ± 0.33**	59.75 ± 0.10	55.43 ± 1.24
	Mean Filling	59.25 ± 0.50	59.42 ± 0.07	58.13 ± 0.49	**59.78 ± 0.12**	59.60 ± 0.26	55.67 ± 0.54
	Learned Embedding	58.93 ± 0.57	59.18 ± 0.06	58.50 ± 0.36	**59.84 ± 0.18**	59.78 ± 0.21	55.55 ± 1.37
	Duplication	59.41 ± 0.31	59.33 ± 0.05	58.48 ± 0.65	59.88 ± 0.17	**59.89 ± 0.06**	56.32 ± 0.21
Amazon - Industrial	Zero Filling	55.38 ± 0.29	55.61 ± 0.04	53.42 ± 1.71	**56.56 ± 0.43**	55.98 ± 0.27	55.27 ± 0.09
	Mean Filling	55.56 ± 0.07	55.70 ± 0.12	54.32 ± 0.28	**56.59 ± 0.29**	55.38 ± 0.22	55.35 ± 0.12
	Learned Embedding	55.40 ± 0.22	55.62 ± 0.10	54.16 ± 0.33	**56.68 ± 0.18**	55.59 ± 0.53	55.39 ± 0.10
	Duplication	55.24 ± 0.10	55.64 ± 0.09	54.58 ± 0.51	**56.78 ± 0.28**	55.72 ± 0.44	54.89 ± 0.12
Amazon - Office	Zero Filling	55.43 ± 0.66	57.26 ± 0.25	54.20 ± 1.24	**58.27 ± 0.36**	57.28 ± 0.25	55.82 ± 0.20
	Mean Filling	56.05 ± 0.35	57.55 ± 0.21	53.29 ± 1.28	**58.18 ± 0.29**	57.35 ± 0.39	55.88 ± 0.10
	Learned Embedding	56.04 ± 0.57	57.45 ± 0.32	53.95 ± 3.05	**58.14 ± 0.36**	57.49 ± 0.50	55.88 ± 0.17
	Duplication	55.44 ± 0.46	57.73 ± 0.17	53.99 ± 1.90	**58.11 ± 0.43**	57.40 ± 0.34	56.97 ± 0.02
Yelp	Zero Filling	50.57 ± 1.30	51.13 ± 0.59	49.13 ± 2.58	**52.42 ± 0.30**	52.00 ± 0.06	50.43 ± 0.71
	Mean Filling	48.57 ± 0.71	50.35 ± 1.84	43.83 ± 4.66	**51.87 ± 0.41**	51.86 ± 0.21	50.51 ± 0.18
	Learned Embedding	47.10 ± 1.56	51.35 ± 0.27	44.02 ± 2.65	**52.41 ± 0.12**	52.14 ± 0.25	50.69 ± 0.21
	Duplication	50.98 ± 0.34	51.31 ± 0.22	50.93 ± 0.75	52.37 ± 0.62	52.34 ± 0.66	**52.43 ± 2.95**

Table 21: Hits@1 (%) for node retrieval under different missing-modality filling strategies. "*tr.*" and "*in.*" refer to the transductive and inductive settings, respectively. The best results are highlighted in bold and the second-best results are underlined.

	Datasets	Feature	TGAT	CAWN	TCL	GraphMixer	DyGFormer	TST-MF
tr.	GDELT	Zero Filling	31.11 ± 1.23	33.56 ± 0.81	31.62 ± 0.04	30.92 ± 0.34	30.98 ± 3.35	**38.32 ± 0.36**
		Mean Filling	29.89 ± 0.96	33.26 ± 0.78	31.71 ± 0.00	30.97 ± 0.07	31.05 ± 1.39	**38.36 ± 0.16**
		Learned Embedding	30.05 ± 0.48	33.46 ± 1.22	31.61 ± 0.01	31.02 ± 0.25	33.92 ± 3.22	**38.09 ± 1.07**
		Duplication	30.42 ± 1.64	33.65 ± 1.09	31.53 ± 0.31	31.11 ± 0.64	34.19 ± 4.03	**38.31 ± 0.74**
	GMCT	Zero Filling	14.73 ± 4.35	11.54 ± 0.16	11.29 ± 0.25	11.62 ± 0.25	**91.12 ± 7.18**	53.84 ± 11.14
		Mean Filling	17.04 ± 1.98	11.58 ± 0.16	11.36 ± 0.23	11.53 ± 0.07	**93.53 ± 1.50**	65.77 ± 2.01
		Learned Embedding	13.80 ± 3.21	11.07 ± 0.27	11.57 ± 0.33	10.50 ± 2.46	**91.94 ± 1.93**	60.24 ± 14.12
		Duplication	13.63 ± 2.32	11.64 ± 0.20	11.58 ± 0.40	10.51 ± 1.45	**93.29 ± 1.56**	73.27 ± 3.06
	Amazon - Industrial	Zero Filling	30.69 ± 3.30	28.00 ± 1.02	32.88 ± 2.42	34.11 ± 7.43	**61.11 ± 6.32**	60.44 ± 7.31
		Mean Filling	35.79 ± 3.66	28.49 ± 1.87	32.25 ± 1.31	31.24 ± 12.57	64.38 ± 4.66	**65.01 ± 5.78**
		Learned Embedding	36.19 ± 2.32	27.96 ± 1.75	31.23 ± 4.83	38.51 ± 0.45	58.00 ± 1.47	**62.51 ± 2.87**
		Duplication	37.95 ± 0.71	27.24 ± 1.35	33.27 ± 2.11	38.32 ± 1.45	62.78 ± 4.46	**65.49 ± 6.91**
	Amazon - Office	Zero Filling	29.34 ± 2.70	22.68 ± 4.53	24.06 ± 8.72	30.72 ± 4.39	**69.47 ± 2.85**	61.60 ± 1.63
		Mean Filling	29.13 ± 1.96	24.85 ± 4.67	29.65 ± 0.12	26.00 ± 10.63	**64.92 ± 2.86**	58.27 ± 1.08
		Learned Embedding	29.54 ± 4.94	22.29 ± 4.26	26.83 ± 2.98	30.51 ± 6.06	**67.86 ± 4.96**	60.14 ± 2.74
		Duplication	24.51 ± 1.88	27.54 ± 1.14	28.33 ± 0.64	26.39 ± 8.54	69.52 ± 1.77	**69.83 ± 1.73**
	Yelp	Zero Filling	28.34 ± 0.87	27.22 ± 0.38	20.09 ± 6.79	19.27 ± 0.24	19.10 ± 7.11	**30.26 ± 0.07**
		Mean Filling	27.32 ± 1.39	26.79 ± 0.43	13.93 ± 2.15	19.25 ± 0.30	22.59 ± 7.78	**29.74 ± 0.23**
		Learned Embedding	27.44 ± 1.17	26.77 ± 0.46	20.83 ± 7.19	19.11 ± 0.10	15.29 ± 0.81	**30.17 ± 0.21**
		Duplication	28.30 ± 1.40	26.96 ± 0.49	20.09 ± 6.56	19.61 ± 0.32	24.07 ± 7.01	**29.94 ± 0.47**
in.	GDELT	Zero Filling	11.20 ± 0.58	17.27 ± 0.56	10.80 ± 0.28	10.84 ± 0.32	19.64 ± 1.37	**23.58 ± 1.05**
		Mean Filling	11.47 ± 0.45	16.80 ± 0.30	11.06 ± 0.28	10.87 ± 0.54	19.47 ± 0.87	**23.34 ± 0.31**
		Learned Embedding	11.07 ± 0.60	16.96 ± 0.51	10.88 ± 0.07	10.69 ± 0.12	20.80 ± 0.93	**23.09 ± 0.39**
		Duplication	12.92 ± 1.65	16.94 ± 0.51	10.87 ± 0.11	11.16 ± 0.30	21.12 ± 1.80	**23.80 ± 0.73**
	GMCT	Zero Filling	4.98 ± 2.27	3.42 ± 0.22	3.34 ± 0.20	3.40 ± 0.62	**62.08 ± 20.77**	11.74 ± 5.45
		Mean Filling	6.24 ± 1.18	3.37 ± 0.04	3.66 ± 0.11	3.10 ± 0.48	**65.42 ± 5.53**	18.47 ± 1.65
		Learned Embedding	4.31 ± 1.76	3.89 ± 0.44	3.65 ± 0.08	3.42 ± 0.92	**59.43 ± 10.14**	17.48 ± 6.40
		Duplication	4.34 ± 1.21	3.33 ± 0.11	3.71 ± 0.24	3.20 ± 0.80	**65.20 ± 7.17**	32.64 ± 2.59
	Amazon - Industrial	Zero Filling	16.54 ± 2.15	14.12 ± 0.84	16.04 ± 0.96	18.09 ± 3.17	23.30 ± 5.65	**33.85 ± 7.18**
		Mean Filling	19.51 ± 1.59	14.47 ± 0.97	15.12 ± 2.15	16.06 ± 6.87	27.70 ± 6.81	**36.58 ± 7.01**
		Learned Embedding	19.70 ± 0.81	14.01 ± 0.98	15.54 ± 2.80	20.22 ± 0.50	19.42 ± 3.98	**31.70 ± 2.55**
		Duplication	20.36 ± 0.56	13.17 ± 0.95	16.47 ± 2.09	20.54 ± 0.52	23.93 ± 3.99	**38.92 ± 1.99**
	Amazon - Office	Zero Filling	16.50 ± 0.71	10.88 ± 3.14	10.39 ± 6.09	16.26 ± 2.37	**44.61 ± 3.76**	37.54 ± 2.02
		Mean Filling	15.65 ± 0.19	13.07 ± 2.58	13.76 ± 0.91	14.10 ± 5.51	**36.78 ± 6.18**	32.75 ± 3.28
		Learned Embedding	15.53 ± 2.55	11.46 ± 2.39	10.59 ± 2.91	16.32 ± 3.20	**41.05 ± 6.52**	37.87 ± 4.11
		Duplication	13.76 ± 1.27	13.68 ± 0.50	12.43 ± 0.36	14.83 ± 4.81	44.47 ± 4.14	**48.45 ± 2.43**
	Yelp	Zero Filling	23.12 ± 0.69	21.74 ± 0.41	16.94 ± 5.34	16.18 ± 0.21	16.58 ± 5.28	**24.91 ± 0.14**
		Mean Filling	22.36 ± 1.10	21.23 ± 0.40	12.17 ± 1.40	16.24 ± 0.31	18.87 ± 5.65	**24.56 ± 0.16**
		Learned Embedding	22.42 ± 0.86	21.23 ± 0.43	17.40 ± 5.55	16.03 ± 0.20	13.77 ± 0.72	**24.75 ± 0.22**
		Duplication	23.12 ± 1.12	21.43 ± 0.40	16.88 ± 4.98	16.56 ± 0.28	20.26 ± 5.04	**24.58 ± 0.45**

Table 22: Hits@3 (%) for node retrieval under different missing-modality filling strategies. "*tr.*" and "*in.*" refer to the transductive and inductive settings, respectively. The best results are highlighted in bold and the second-best results are underlined.

	Datasets	Feature	TGAT	CAWN	TCL	GraphMixer	DyGFormer	TST-MF
tr.	GDELT	Zero Filling	52.29 ± 0.85	56.05 ± 0.66	53.79 ± 0.13	51.47 ± 1.10	45.49 ± 7.59	**59.28 ± 0.39**
		Mean Filling	50.96 ± 1.07	55.35 ± 0.45	53.63 ± 0.38	51.93 ± 0.55	46.53 ± 3.51	**59.50 ± 0.02**
		Learned Embedding	50.86 ± 1.06	55.71 ± 1.37	53.61 ± 0.16	52.29 ± 0.38	51.30 ± 6.33	**59.07 ± 1.45**
		Duplication	51.56 ± 2.13	55.65 ± 1.06	53.46 ± 0.21	51.71 ± 2.01	52.06 ± 6.41	**59.05 ± 0.36**
	GMCT	Zero Filling	30.79 ± 6.07	25.57 ± 0.12	25.54 ± 0.38	25.81 ± 0.74	**99.98 ± 0.03**	97.40 ± 1.97
		Mean Filling	33.83 ± 3.37	25.51 ± 0.14	24.72 ± 0.23	26.26 ± 0.98	**100.00 ± 0.00**	99.31 ± 0.08
		Learned Embedding	29.24 ± 4.29	24.97 ± 0.06	25.25 ± 0.31	24.20 ± 2.36	**100.00 ± 0.00**	97.12 ± 4.11
		Duplication	29.04 ± 4.09	25.53 ± 0.18	25.24 ± 0.17	25.07 ± 1.56	**100.00 ± 0.01**	99.69 ± 0.69
	Amazon - Industrial	Zero Filling	48.80 ± 3.59	47.07 ± 0.75	52.22 ± 2.67	53.16 ± 9.28	**96.09 ± 1.45**	88.71 ± 7.04
		Mean Filling	55.19 ± 4.06	47.60 ± 1.78	51.96 ± 2.95	50.42 ± 14.84	**96.99 ± 0.82**	93.35 ± 4.18
		Learned Embedding	56.00 ± 3.01	47.20 ± 1.56	50.16 ± 4.18	58.36 ± 0.26	**95.60 ± 0.68**	93.42 ± 3.09
		Duplication	57.01 ± 0.78	46.79 ± 1.57	52.72 ± 2.19	57.74 ± 1.93	**97.08 ± 0.70**	95.71 ± 2.92
	Amazon - Office	Zero Filling	45.70 ± 4.63	39.67 ± 3.80	43.31 ± 7.01	49.18 ± 6.06	**94.45 ± 1.17**	89.28 ± 1.19
		Mean Filling	47.04 ± 3.62	42.28 ± 5.93	48.96 ± 0.89	43.10 ± 13.78	**92.40 ± 1.03**	87.84 ± 0.96
		Learned Embedding	47.25 ± 6.40	39.06 ± 5.38	49.41 ± 1.49	48.94 ± 8.33	**93.01 ± 2.33**	85.72 ± 2.05
		Duplication	40.87 ± 1.54	45.82 ± 2.08	48.84 ± 1.91	43.01 ± 10.75	94.75 ± 0.75	**97.02 ± 1.93**
	Yelp	Zero Filling	50.67 ± 1.15	48.96 ± 0.47	38.15 ± 11.37	36.37 ± 0.43	34.49 ± 11.73	**53.22 ± 0.28**
		Mean Filling	49.35 ± 1.76	48.53 ± 0.67	27.84 ± 4.38	36.48 ± 0.42	41.15 ± 13.46	**52.23 ± 0.41**
		Learned Embedding	49.50 ± 1.40	48.52 ± 0.68	39.18 ± 12.07	36.26 ± 0.12	27.90 ± 1.23	**52.84 ± 0.42**
		Duplication	50.58 ± 1.65	48.54 ± 0.69	38.16 ± 11.46	37.01 ± 0.40	42.84 ± 12.07	**53.91 ± 0.94**
in.	GDELT	Zero Filling	27.21 ± 0.58	28.89 ± 0.63	27.17 ± 0.50	27.50 ± 0.52	30.64 ± 3.50	**39.49 ± 0.67**
		Mean Filling	26.45 ± 0.25	28.79 ± 0.65	27.64 ± 0.32	27.50 ± 0.44	30.87 ± 1.42	**39.13 ± 0.72**
		Learned Embedding	26.20 ± 0.51	28.99 ± 0.84	27.20 ± 0.38	27.27 ± 0.27	32.99 ± 2.97	**38.52 ± 0.45**
		Duplication	26.96 ± 1.22	29.43 ± 0.89	27.23 ± 0.12	27.70 ± 0.35	34.12 ± 2.69	**39.08 ± 0.49**
	GMCT	Zero Filling	12.97 ± 4.49	9.45 ± 0.10	8.95 ± 0.45	9.17 ± 1.24	**95.42 ± 5.08**	49.28 ± 13.14
		Mean Filling	15.13 ± 2.31	9.53 ± 0.03	9.38 ± 0.30	8.64 ± 0.97	**98.04 ± 1.05**	64.78 ± 1.94
		Learned Embedding	11.57 ± 3.49	9.29 ± 0.17	9.60 ± 0.10	8.76 ± 2.50	**97.42 ± 1.07**	58.44 ± 18.84
		Duplication	11.71 ± 2.41	9.42 ± 0.10	9.70 ± 0.19	8.30 ± 1.87	**97.81 ± 1.18**	79.29 ± 3.29
	Amazon - Industrial	Zero Filling	28.84 ± 3.76	25.05 ± 1.26	30.05 ± 2.49	31.78 ± 7.07	**57.56 ± 10.16**	57.21 ± 7.74
		Mean Filling	34.15 ± 3.62	25.83 ± 2.28	29.69 ± 1.26	28.53 ± 12.56	**63.59 ± 7.70**	62.31 ± 6.83
		Learned Embedding	34.71 ± 2.23	25.07 ± 1.90	28.73 ± 5.34	35.92 ± 0.82	52.85 ± 3.17	**58.87 ± 3.88**
		Duplication	36.56 ± 0.92	24.05 ± 1.52	31.24 ± 2.42	36.06 ± 0.97	60.52 ± 6.87	**62.75 ± 8.84**
	Amazon - Office	Zero Filling	29.99 ± 2.16	21.43 ± 5.55	22.40 ± 11.18	30.69 ± 4.43	**70.64 ± 3.02**	61.55 ± 1.61
		Mean Filling	28.99 ± 1.57	23.92 ± 4.96	28.98 ± 0.53	25.46 ± 10.88	**65.65 ± 3.45**	58.40 ± 2.02
		Learned Embedding	29.12 ± 5.21	21.17 ± 4.61	23.76 ± 4.36	30.43 ± 6.12	**69.59 ± 5.93**	61.27 ± 2.32
		Duplication	24.08 ± 2.53	26.80 ± 0.86	26.38 ± 0.63	26.45 ± 9.14	70.02 ± 1.85	**73.66 ± 3.00**
	Yelp	Zero Filling	41.97 ± 1.05	40.08 ± 0.46	32.10 ± 8.86	30.72 ± 0.40	29.85 ± 9.10	**43.89 ± 0.14**
		Mean Filling	40.84 ± 1.46	39.63 ± 0.61	24.17 ± 2.85	30.83 ± 0.34	34.42 ± 10.47	**43.34 ± 0.32**
		Learned Embedding	41.02 ± 1.11	39.59 ± 0.67	32.90 ± 9.22	30.67 ± 0.17	24.94 ± 1.50	**43.68 ± 0.39**
		Duplication	41.90 ± 1.38	39.65 ± 0.58	31.92 ± 8.53	31.26 ± 0.36	36.34 ± 8.74	**44.44 ± 0.65**

Table 23: Hits@10 (%) for node retrieval under different missing-modality filling strategies. "*tr.*" and "*in.*" refer to the transductive and inductive settings, respectively. The best results are highlighted in bold and the second-best results are underlined.

	Datasets	Feature	TGAT	CAWN	TCL	GraphMixer	DyGFormer	TST-MF
tr.	GDELT	Zero Filling	70.55 ± 2.09	79.20 ± 0.13	75.11 ± 0.31	70.79 ± 1.22	63.87 ± 10.27	**81.83 ± 0.38**
		Mean Filling	70.47 ± 0.99	79.05 ± 0.27	74.92 ± 0.59	71.28 ± 1.57	64.12 ± 5.79	**82.01 ± 0.66**
		Learned Embedding	69.57 ± 2.06	79.41 ± 0.70	74.84 ± 0.22	71.74 ± 1.53	70.45 ± 8.88	**81.67 ± 0.72**
		Duplication	70.79 ± 1.04	79.08 ± 0.35	74.66 ± 0.43	72.41 ± 1.16	71.34 ± 7.82	**81.62 ± 0.33**
	GMCT	Zero Filling	54.67 ± 4.52	50.77 ± 0.04	50.20 ± 0.20	50.84 ± 0.43	**100.00 ± 0.00**	**100.00 ± 0.00**
		Mean Filling	57.29 ± 3.80	50.72 ± 0.11	49.83 ± 0.45	50.91 ± 0.33	**100.00 ± 0.00**	**100.00 ± 0.00**
		Learned Embedding	52.68 ± 2.56	50.53 ± 0.08	49.78 ± 0.41	50.34 ± 0.35	**100.00 ± 0.00**	99.97 ± 0.04
		Duplication	53.02 ± 2.88	50.73 ± 0.02	50.27 ± 0.24	50.72 ± 0.69	**100.00 ± 0.00**	**100.00 ± 0.00**
	Amazon - Industrial	Zero Filling	77.50 ± 1.36	77.23 ± 0.67	78.22 ± 0.43	78.21 ± 2.64	**100.00 ± 0.01**	99.43 ± 0.88
		Mean Filling	79.61 ± 1.07	77.42 ± 0.27	77.88 ± 0.91	78.85 ± 3.58	**100.00 ± 0.01**	99.96 ± 0.07
		Learned Embedding	79.61 ± 0.82	77.29 ± 0.28	78.02 ± 0.57	79.93 ± 0.93	99.94 ± 0.10	**99.99 ± 0.02**
		Duplication	79.92 ± 0.13	77.07 ± 0.57	78.61 ± 1.00	79.72 ± 1.43	**100.00 ± 0.00**	**100.00 ± 0.00**
	Amazon - Office	Zero Filling	73.44 ± 1.70	71.26 ± 2.07	71.69 ± 1.20	75.07 ± 2.25	**99.94 ± 0.08**	99.88 ± 0.11
		Mean Filling	75.00 ± 1.51	73.11 ± 2.22	72.82 ± 0.71	73.89 ± 4.72	98.77 ± 1.11	**99.93 ± 0.05**
		Learned Embedding	74.58 ± 1.58	71.59 ± 2.45	73.68 ± 0.80	75.21 ± 2.73	97.78 ± 1.05	**99.63 ± 0.47**
		Duplication	73.38 ± 0.47	72.72 ± 1.24	73.01 ± 0.94	73.42 ± 3.39	**99.99 ± 0.01**	**99.99 ± 0.01**
	Yelp	Zero Filling	74.25 ± 0.82	73.20 ± 0.35	63.00 ± 13.50	61.09 ± 0.35	57.35 ± 13.90	**77.24 ± 0.25**
		Mean Filling	73.34 ± 1.10	72.98 ± 0.47	51.82 ± 7.19	61.16 ± 0.23	65.03 ± 15.52	**76.31 ± 0.18**
		Learned Embedding	73.53 ± 1.05	72.92 ± 0.49	63.73 ± 13.73	60.96 ± 0.09	49.57 ± 1.42	**76.86 ± 0.33**
		Duplication	74.19 ± 1.22	72.94 ± 0.44	63.01 ± 13.26	61.72 ± 0.29	67.54 ± 14.55	**79.11 ± 0.49**
in.	GDELT	Zero Filling	47.70 ± 0.76	49.61 ± 0.21	48.39 ± 0.16	47.74 ± 0.85	44.27 ± 9.39	**59.23 ± 0.52**
		Mean Filling	46.80 ± 0.60	49.23 ± 1.03	48.43 ± 0.36	47.87 ± 0.38	46.38 ± 4.35	**58.97 ± 0.54**
		Learned Embedding	47.81 ± 0.72	49.83 ± 1.08	48.53 ± 0.18	47.94 ± 0.38	49.95 ± 6.63	**58.58 ± 0.46**
		Duplication	46.86 ± 1.00	49.69 ± 0.94	48.55 ± 0.27	47.72 ± 1.15	52.50 ± 3.86	**59.23 ± 0.58**
	GMCT	Zero Filling	29.68 ± 6.25	23.32 ± 0.35	22.78 ± 0.29	23.32 ± 0.18	**100.00 ± 0.00**	99.91 ± 0.11
		Mean Filling	33.24 ± 4.40	23.40 ± 0.16	22.39 ± 0.48	23.39 ± 0.42	**100.00 ± 0.00**	99.97 ± 0.04
		Learned Embedding	27.60 ± 4.42	22.59 ± 0.38	22.86 ± 0.28	22.26 ± 2.15	**100.00 ± 0.00**	98.50 ± 2.45
		Duplication	27.40 ± 4.31	23.23 ± 0.12	22.75 ± 0.33	23.15 ± 0.75	**100.00 ± 0.01**	**100.00 ± 0.00**
	Amazon - Industrial	Zero Filling	44.25 ± 4.52	41.65 ± 1.19	47.74 ± 3.54	49.12 ± 10.35	**98.88 ± 0.63**	88.47 ± 9.13
		Mean Filling	51.51 ± 4.62	42.41 ± 2.28	47.90 ± 3.30	45.77 ± 16.63	**99.35 ± 0.33**	94.36 ± 5.86
		Learned Embedding	52.36 ± 3.61	41.76 ± 1.98	45.20 ± 6.09	54.78 ± 0.44	**98.86 ± 0.90**	94.83 ± 4.80
		Duplication	53.98 ± 1.23	41.28 ± 1.89	48.54 ± 2.34	54.10 ± 1.97	99.51 ± 0.19	**99.94 ± 0.99**
	Amazon - Office	Zero Filling	43.90 ± 4.75	36.64 ± 4.58	39.73 ± 10.53	47.37 ± 6.14	**97.51 ± 0.70**	91.51 ± 1.75
		Mean Filling	44.77 ± 3.81	39.11 ± 6.39	47.14 ± 0.77	40.31 ± 14.60	**95.44 ± 1.37**	89.11 ± 2.15
		Learned Embedding	44.91 ± 6.89	35.73 ± 5.63	47.51 ± 1.66	46.87 ± 8.71	**96.23 ± 1.59**	86.05 ± 3.09
		Duplication	38.04 ± 1.99	43.72 ± 2.65	46.78 ± 1.55	40.61 ± 11.80	98.10 ± 0.61	**98.99 ± 0.31**
	Yelp	Zero Filling	65.24 ± 0.91	63.79 ± 0.39	54.95 ± 11.01	53.12 ± 0.38	50.75 ± 11.42	**67.92 ± 0.20**
		Mean Filling	64.25 ± 1.17	63.59 ± 0.50	45.40 ± 5.04	53.15 ± 0.23	57.24 ± 13.49	**67.05 ± 0.22**
		Learned Embedding	64.40 ± 1.09	63.49 ± 0.57	55.70 ± 11.16	53.01 ± 0.14	44.30 ± 1.66	**67.63 ± 0.46**
		Duplication	65.24 ± 1.17	63.49 ± 0.58	54.84 ± 10.96	53.71 ± 0.20	59.48 ± 12.06	**67.84 ± 0.48**

Table 24: Hits@10 (%) for recurrence-aware node retrieval with historical negative sampling. "*tr.*" and "*in.*" refer to the transductive and inductive settings, respectively. The best results are highlighted in bold and the second-best results are underlined.

	Datasets	Feature	TGAT	CAWN	TCL	GraphMixer	DyGFormer	TST-MF
tr.	GDELT	Random	20.78 ± 0.26	26.18 ± 0.80	20.98 ± 0.16	20.70 ± 0.25	**33.32 ± 0.57**	27.27 ± 0.78
		Text	20.77 ± 1.49	27.81 ± 1.13	20.65 ± 0.09	21.58 ± 1.52	**35.39 ± 0.49**	34.81 ± 1.46
		Text+Image	22.70 ± 1.50	27.73 ± 0.91	20.75 ± 0.11	20.82 ± 0.25	35.56 ± 0.34	**36.03 ± 0.78**
	GMCT	Random	41.08 ± 3.52	44.08 ± 0.09	44.04 ± 0.14	44.06 ± 0.05	**99.91 ± 0.12**	47.54 ± 30.15
		Text	47.63 ± 4.12	44.15 ± 0.03	43.91 ± 0.22	25.25 ± 16.46	**99.96 ± 0.06**	95.85 ± 3.10
		Text+Image	45.95 ± 1.68	44.19 ± 0.06	43.99 ± 0.18	34.15 ± 11.72	**99.99 ± 0.02**	81.52 ± 3.04
	Amazon - Industrial	Random	**45.17 ± 3.35**	41.86 ± 1.79	23.97 ± 5.51	34.07 ± 2.97	23.54 ± 4.73	31.44 ± 0.75
		Text	52.44 ± 2.15	**55.54 ± 2.87**	40.98 ± 5.95	41.04 ± 1.94	29.04 ± 2.83	43.61 ± 0.70
		Text+Image	52.17 ± 1.83	**52.21 ± 7.99**	40.07 ± 3.37	43.77 ± 1.48	24.55 ± 4.41	45.84 ± 5.78
	Amazon - Office	Random	47.59 ± 5.13	45.72 ± 20.31	35.80 ± 2.62	**52.06 ± 2.19**	39.27 ± 3.67	35.09 ± 8.59
		Text	**53.21 ± 5.69**	49.56 ± 9.19	37.34 ± 5.72	53.08 ± 4.72	49.95 ± 7.10	46.09 ± 7.49
		Text+Image	**56.72 ± 1.79**	47.05 ± 4.99	33.21 ± 1.62	52.76 ± 3.65	55.60 ± 2.67	53.29 ± 3.91
	Yelp	Random	13.64 ± 0.43	14.32 ± 1.22	12.63 ± 0.06	12.94 ± 0.26	14.16 ± 0.31	**15.53 ± 0.32**
		Text	**49.28 ± 0.09**	49.06 ± 0.43	36.46 ± 9.07	30.33 ± 1.28	37.04 ± 19.78	48.27 ± 0.33
		Text+Image	**49.05 ± 1.37**	47.95 ± 0.93	32.38 ± 17.37	31.13 ± 0.54	35.52 ± 17.70	47.95 ± 0.11
in.	GDELT	Random	15.14 ± 0.22	19.68 ± 0.46	15.63 ± 0.03	15.28 ± 0.16	**26.70 ± 1.56**	23.09 ± 1.12
		Text	15.71 ± 0.84	22.93 ± 1.09	15.66 ± 0.09	15.99 ± 1.16	29.85 ± 0.50	**32.81 ± 0.99**
		Text+Image	16.96 ± 1.32	22.44 ± 0.84	15.63 ± 0.16	15.47 ± 0.31	30.20 ± 0.43	**34.40 ± 0.40**
	GMCT	Random	43.20 ± 2.32	46.85 ± 0.28	45.30 ± 0.45	47.03 ± 0.04	**99.54 ± 0.77**	52.09 ± 31.43
		Text	49.46 ± 3.29	47.35 ± 0.18	45.76 ± 0.22	28.25 ± 16.53	**99.80 ± 0.35**	96.05 ± 2.64
		Text+Image	47.80 ± 1.74	47.46 ± 0.06	45.24 ± 0.57	37.05 ± 12.35	**99.95 ± 0.06**	82.43 ± 3.03
	Amazon - Industrial	Random	**20.58 ± 1.97**	12.94 ± 2.22	12.13 ± 3.14	13.26 ± 0.78	9.91 ± 5.15	13.28 ± 0.63
		Text	30.27 ± 1.41	23.75 ± 1.51	19.79 ± 4.09	22.80 ± 0.87	26.07 ± 2.74	**38.69 ± 1.65**
		Text+Image	28.74 ± 0.80	20.38 ± 1.34	18.48 ± 3.36	23.30 ± 1.55	23.37 ± 3.70	**42.72 ± 7.43**
	Amazon - Office	Random	22.22 ± 0.99	18.16 ± 5.32	14.64 ± 0.48	**31.52 ± 1.38**	22.84 ± 10.73	13.21 ± 3.12
		Text	31.11 ± 3.51	17.22 ± 3.17	15.03 ± 4.65	31.05 ± 6.94	**46.87 ± 8.53**	33.54 ± 4.26
		Text+Image	32.55 ± 1.31	16.10 ± 3.79	13.14 ± 0.17	27.78 ± 6.13	**55.07 ± 3.69**	40.31 ± 4.99
	Yelp	Random	11.33 ± 0.60	**12.91 ± 0.89**	10.11 ± 0.11	10.46 ± 0.14	8.80 ± 1.25	12.60 ± 0.21
		Text	40.67 ± 0.05	**41.62 ± 0.47**	28.73 ± 7.57	24.16 ± 1.17	30.96 ± 16.80	40.32 ± 0.31
		Text+Image	40.52 ± 1.26	**40.56 ± 0.96**	25.91 ± 13.64	24.71 ± 0.55	29.61 ± 14.47	40.05 ± 0.23

Table 25: Repeated Hits@10 (%) for recurrence-aware node retrieval with historical negative sampling. "*tr.*" and "*in.*" refer to the transductive and inductive settings, respectively. The best results are highlighted in bold and the second-best results are underlined.

	Datasets	Feature	TGAT	CAWN	TCL	GraphMixer	DyGFormer	TST-MF
tr.	GDELT	Random	27.26 ± 0.32	35.93 ± 1.14	27.38 ± 0.22	27.11 ± 0.33	**45.43 ± 0.76**	37.46 ± 1.11
		Text	27.07 ± 2.30	38.30 ± 1.62	26.88 ± 0.12	28.35 ± 2.12	**47.80 ± 0.67**	46.73 ± 1.94
		Text+Image	29.62 ± 1.98	38.03 ± 1.32	27.06 ± 0.12	27.36 ± 0.32	**48.13 ± 0.46**	48.08 ± 0.95
	GMCT	Random	33.76 ± 2.96	37.38 ± 0.28	38.49 ± 0.04	38.27 ± 0.20	**99.03 ± 0.10**	45.75 ± 28.91
		Text	42.04 ± 4.45	37.58 ± 0.20	37.74 ± 0.32	24.11 ± 12.57	**99.05 ± 0.06**	94.89 ± 3.06
		Text+Image	40.61 ± 2.61	37.54 ± 0.22	38.32 ± 0.13	30.79 ± 8.96	**99.09 ± 0.00**	81.00 ± 2.87
	Amazon - Industrial	Random	**60.92 ± 3.03**	58.68 ± 1.15	38.04 ± 14.49	55.23 ± 1.60	42.04 ± 4.60	51.49 ± 0.96
		Text	65.81 ± 1.39	**67.74 ± 1.30**	58.35 ± 7.13	60.53 ± 0.74	45.40 ± 3.04	61.11 ± 0.82
		Text+Image	**65.38 ± 1.38**	63.45 ± 5.87	55.23 ± 3.02	62.23 ± 0.70	40.09 ± 4.90	62.19 ± 4.69
	Amazon - Office	Random	47.59 ± 3.36	48.03 ± 14.41	39.07 ± 1.77	**52.22 ± 1.30**	41.28 ± 3.53	38.04 ± 6.93
		Text	52.39 ± 4.84	52.04 ± 6.60	40.46 ± 3.81	**53.22 ± 3.41**	50.11 ± 6.03	47.66 ± 5.42
		Text+Image	**55.79 ± 1.85**	50.14 ± 3.75	37.74 ± 1.80	52.39 ± 2.87	54.39 ± 1.50	53.12 ± 2.81
	Yelp	Random	10.37 ± 1.03	79.68 ± 11.06	9.15 ± 0.10	9.21 ± 0.21	94.08 ± 5.13	**96.89 ± 0.18**
		Text	86.48 ± 0.30	96.81 ± 0.47	57.72 ± 19.21	47.76 ± 2.82	94.56 ± 6.60	**98.60 ± 0.98**
		Text+Image	85.45 ± 3.16	96.91 ± 0.38	51.20 ± 36.87	49.65 ± 1.64	93.48 ± 6.45	**98.46 ± 1.08**
in.	GDELT	Random	21.93 ± 0.41	29.73 ± 1.07	22.99 ± 0.07	22.30 ± 0.30	**42.00 ± 1.88**	35.64 ± 2.03
		Text	22.85 ± 2.02	33.89 ± 1.10	22.77 ± 0.19	23.48 ± 1.46	45.26 ± 0.57	**49.23 ± 1.48**
		Text+Image	24.69 ± 1.98	33.42 ± 1.44	22.80 ± 0.18	22.79 ± 0.39	46.12 ± 0.44	**50.36 ± 0.87**
	GMCT	Random	35.59 ± 1.96	39.93 ± 0.61	38.45 ± 0.92	39.87 ± 0.41	**98.80 ± 0.32**	49.32 ± 31.01
		Text	43.50 ± 4.09	40.58 ± 0.20	38.96 ± 0.82	25.38 ± 12.79	**98.87 ± 0.20**	95.72 ± 2.46
		Text+Image	42.62 ± 2.75	40.99 ± 0.14	38.39 ± 0.37	33.04 ± 8.62	**98.95 ± 0.06**	82.62 ± 2.87
	Amazon - Industrial	Random	**39.38 ± 1.84**	29.62 ± 1.96	24.29 ± 9.53	35.94 ± 1.24	21.30 ± 10.44	33.71 ± 1.30
		Text	44.40 ± 3.89	41.22 ± 2.76	38.01 ± 9.36	42.69 ± 2.57	40.12 ± 2.36	**55.93 ± 2.53**
		Text+Image	41.97 ± 0.86	33.97 ± 1.93	27.70 ± 6.33	42.58 ± 1.43	36.73 ± 3.83	**59.02 ± 5.69**
	Amazon - Office	Random	25.75 ± 1.29	23.19 ± 3.44	21.80 ± 0.74	**34.77 ± 1.18**	26.97 ± 9.24	19.84 ± 3.58
		Text	33.89 ± 1.79	23.76 ± 1.68	20.84 ± 1.81	33.78 ± 5.29	**46.78 ± 5.88**	36.67 ± 2.67
		Text+Image	35.06 ± 1.96	24.92 ± 1.90	19.22 ± 1.09	30.96 ± 3.66	**52.69 ± 2.63**	42.85 ± 2.72
	Yelp	Random	10.17 ± 0.73	76.62 ± 7.13	8.39 ± 0.05	8.15 ± 0.39	86.67 ± 2.08	**88.10 ± 0.31**
		Text	78.01 ± 0.84	86.47 ± 0.26	49.78 ± 16.68	36.28 ± 1.99	**87.94 ± 0.16**	87.59 ± 0.88
		Text+Image	76.84 ± 3.31	86.88 ± 0.21	44.14 ± 31.43	37.99 ± 1.27	86.98 ± 0.46	**87.84 ± 0.54**

Table 26: Novel Hits@10 (%) for recurrence-aware node retrieval with historical negative sampling. "*tr.*" and "*in.*" refer to the transductive and inductive settings, respectively. The best results are highlighted in bold and the second-best results are underlined.

	Datasets	Feature	TGAT	CAWN	TCL	GraphMixer	DyGFormer	TST-MF
tr.	GDELT	Random	8.77 ± 0.16	7.83 ± 0.20	9.04 ± 0.08	8.81 ± 0.11	**10.42 ± 0.23**	8.27 ± 0.15
		Text	9.09 ± 0.04	8.06 ± 0.28	9.04 ± 0.03	8.98 ± 0.36	11.96 ± 0.55	**12.44 ± 0.55**
		Text+Image	9.79 ± 0.66	8.34 ± 0.64	9.03 ± 0.07	8.69 ± 0.21	11.81 ± 0.26	**13.40 ± 0.60**
	GMCT	Random	41.22 ± 3.52	44.20 ± 0.09	44.15 ± 0.15	44.17 ± 0.04	**99.91 ± 0.12**	47.41 ± 30.28
		Text	47.73 ± 4.12	44.27 ± 0.02	44.03 ± 0.23	25.27 ± 16.54	**99.96 ± 0.06**	95.89 ± 3.10
		Text+Image	46.03 ± 1.66	44.30 ± 0.06	44.10 ± 0.17	34.22 ± 11.77	**99.99 ± 0.02**	81.45 ± 3.09
	Amazon - Industrial	Random	**44.79 ± 3.36**	41.48 ± 1.81	23.67 ± 5.40	33.60 ± 3.00	23.14 ± 4.72	30.84 ± 0.78
		Text	52.11 ± 2.17	**55.25 ± 2.91**	40.58 ± 5.92	40.59 ± 1.97	28.67 ± 2.83	43.03 ± 0.74
		Text+Image	51.85 ± 1.83	**51.95 ± 8.03**	39.72 ± 3.37	43.34 ± 1.51	24.21 ± 4.42	45.29 ± 5.80
	Amazon - Office	Random	47.55 ± 5.16	45.65 ± 20.38	35.72 ± 2.63	**52.01 ± 2.19**	39.21 ± 3.67	35.02 ± 8.60
		Text	**53.18 ± 5.70**	49.50 ± 9.22	37.27 ± 5.75	53.04 ± 4.74	49.91 ± 7.10	46.04 ± 7.51
		Text+Image	**56.70 ± 1.80**	46.98 ± 5.01	33.12 ± 1.63	52.73 ± 3.66	55.57 ± 2.68	53.26 ± 3.92
	Yelp	Random	**13.75 ± 0.41**	12.13 ± 1.05	12.74 ± 0.06	13.06 ± 0.26	11.48 ± 0.23	12.74 ± 0.34
		Text	**48.05 ± 0.10**	47.47 ± 0.44	35.75 ± 8.73	29.75 ± 1.23	35.11 ± 20.22	46.54 ± 0.33
		Text+Image	**47.84 ± 1.32**	46.31 ± 0.95	31.75 ± 16.72	30.51 ± 0.52	33.58 ± 18.08	46.21 ± 0.08
in.	GDELT	Random	10.24 ± 0.10	12.32 ± 0.37	10.33 ± 0.04	10.28 ± 0.07	**15.44 ± 1.35**	14.02 ± 0.57
		Text	10.64 ± 0.14	14.78 ± 1.16	10.60 ± 0.03	10.62 ± 0.88	18.45 ± 0.47	**20.88 ± 0.73**
		Text+Image	11.34 ± 0.80	14.27 ± 0.56	10.51 ± 0.09	10.23 ± 0.31	18.44 ± 0.46	**22.78 ± 0.53**
	GMCT	Random	43.34 ± 2.32	46.97 ± 0.28	45.42 ± 0.44	47.14 ± 0.04	**99.54 ± 0.77**	52.13 ± 31.45
		Text	49.56 ± 3.29	47.47 ± 0.18	45.88 ± 0.21	28.29 ± 16.60	**99.80 ± 0.35**	96.05 ± 2.64
		Text+Image	47.89 ± 1.73	47.57 ± 0.07	45.35 ± 0.57	37.12 ± 12.41	**99.95 ± 0.06**	82.43 ± 3.03
	Amazon - Industrial	Random	**20.17 ± 1.99**	12.58 ± 2.26	11.87 ± 3.02	12.79 ± 0.77	9.68 ± 5.05	12.87 ± 0.62
		Text	29.97 ± 1.34	23.37 ± 1.47	19.39 ± 3.96	22.36 ± 0.95	25.78 ± 2.77	**38.29 ± 1.65**
		Text+Image	28.48 ± 0.79	20.11 ± 1.36	18.30 ± 3.29	22.89 ± 1.55	23.11 ± 3.73	**42.34 ± 7.47**
	Amazon - Office	Random	22.16 ± 1.01	18.08 ± 5.34	14.54 ± 0.48	**31.46 ± 1.39**	22.77 ± 10.74	13.12 ± 3.12
		Text	31.04 ± 3.53	17.12 ± 3.20	14.95 ± 4.68	31.00 ± 6.95	**46.83 ± 8.55**	33.47 ± 4.28
		Text+Image	32.50 ± 1.30	15.97 ± 3.81	13.04 ± 0.19	27.73 ± 6.15	**55.06 ± 3.71**	40.25 ± 5.01
	Yelp	Random	11.32 ± 0.60	**12.12 ± 0.83**	10.11 ± 0.11	10.47 ± 0.14	7.85 ± 1.25	11.67 ± 0.22
		Text	40.17 ± 0.06	**41.02 ± 0.48**	28.42 ± 7.46	23.96 ± 1.16	30.22 ± 16.98	39.70 ± 0.30
		Text+Image	**40.03 ± 1.23**	39.94 ± 0.98	25.64 ± 13.40	24.51 ± 0.56	28.87 ± 14.62	39.42 ± 0.23

Table 27: Instance-level utility of each modality for TST-MF under the Text+Image setting. For a test instance i and modality m, $\Delta_{i,m} = \ell(f_{\setminus m}(i), y_i) - \ell(f(i), y_i)$, where $f_{\setminus m}$ zeroes modality m in every pathway of the model and ℓ is the negative log-likelihood of the true interaction (link prediction) or of the true class (edge classification); $\Delta_{i,m} > 0$ means that the modality helps the prediction.

Datasets	Task	Mean Δ [95% CI] Text	Mean Δ [95% CI] Image	Share of $\Delta > 0$ (%) Text	Share of $\Delta > 0$ (%) Image
GDELT	LP (*tr.*)	-0.096 [-0.102, -0.092]	-0.103 [-0.130, -0.084]	42.1	36.6
	LP (*in.*)	-0.124 [-0.130, -0.117]	-0.131 [-0.158, -0.108]	37.7	33.3
	EC	-0.004 [-0.014, +0.003]	+0.000 [-0.006, +0.006]	46.8	48.1
GMCT	LP (*tr.*)	+0.149 [+0.003, +0.406]	+0.282 [+0.006, +0.818]	76.6	77.2
	LP (*in.*)	+0.163 [+0.003, +0.443]	+0.308 [+0.006, +0.889]	78.6	79.3
	EC	+0.011 [+0.010, +0.012]	+0.008 [+0.006, +0.010]	44.1	56.8
Amazon - Industrial	LP (*tr.*)	+0.678 [+0.025, +1.642]	+0.375 [-0.002, +1.107]	74.9	68.1
	LP (*in.*)	+0.707 [+0.024, +1.725]	+0.397 [+0.000, +1.181]	75.3	69.8
	EC	+0.146 [+0.075, +0.221]	+0.005 [+0.002, +0.008]	67.5	46.7
Amazon - Office	LP (*tr.*)	+0.026 [+0.007, +0.037]	+0.009 [-0.020, +0.026]	69.6	59.4
	LP (*in.*)	+0.023 [+0.003, +0.034]	+0.012 [-0.018, +0.031]	70.2	61.8
	EC	+0.120 [+0.060, +0.203]	+0.018 [+0.006, +0.027]	63.9	45.4
Yelp	LP (*tr.*)	+0.204 [+0.123, +0.353]	+0.042 [+0.023, +0.063]	81.6	77.8
	LP (*in.*)	+0.186 [+0.116, +0.311]	+0.045 [+0.029, +0.062]	81.2	77.6
	EC	+0.085 [+0.073, +0.100]	+0.057 [+0.045, +0.070]	63.8	63.2

E Ablation Study

We conduct ablation studies on our proposed **TST-MF** method to verify the effectiveness of each design choice. The result is given in Table 28 and 27.

Table 28: Ablation Study over Link Prediction Task. Four variants are experimented. "w/o target" removes target guidance from the context cross-attention layer; "w/o hier." replaces hierarchical fusion with a single-stage alternative; "w/o spatial" removes the structural-context branch and "w/o historical" removes the interaction-history branch. Best results are **bolded**.

	Datasets	Metric	TST-MF	w/o target	w/o hier.	w/o spatial	w/o historical
tr.	GDELT	Precision	**93.98 ± 0.05**	92.08 ± 0.21	91.90 ± 0.15	92.06 ± 0.22	93.95 ± 0.06
		AUC-ROC	**93.71 ± 0.08**	92.78 ± 0.18	91.54 ± 0.20	92.74 ± 0.29	91.78 ± 0.09
	GMCT	Precision	**89.08 ± 0.72**	88.83 ± 1.24	82.28 ± 3.99	88.46 ± 0.66	89.03 ± 0.34
		AUC-ROC	**92.49 ± 0.66**	91.67 ± 0.89	82.28 ± 4.06	91.33 ± 0.88	91.87 ± 0.04
	Amazon - Industrial	Precision	**94.10 ± 0.12**	94.01 ± 0.28	93.56 ± 0.50	94.01 ± 0.23	93.99 ± 0.05
		AUC-ROC	**94.26 ± 0.16**	93.23 ± 0.27	93.94 ± 0.22	93.25 ± 0.19	93.38 ± 0.11
	Amazon - Office	Precision	**92.67 ± 0.10**	91.86 ± 0.11	91.94 ± 0.22	92.01 ± 0.29	91.40 ± 0.38
		AUC-ROC	**93.23 ± 0.43**	92.97 ± 0.35	92.72 ± 0.15	92.87 ± 0.15	92.62 ± 0.21
	Yelp	Precision	**92.65 ± 0.04**	92.51 ± 0.41	92.31 ± 0.07	92.32 ± 0.09	92.28 ± 0.10
		AUC-ROC	**93.83 ± 0.02**	92.86 ± 0.26	92.64 ± 0.07	92.69 ± 0.10	92.61 ± 0.11
in.	GDELT	Precision	85.44 ± 0.14	84.69 ± 0.52	**85.50 ± 0.27**	84.44 ± 0.22	84.32 ± 0.19
		AUC-ROC	**84.51 ± 0.11**	84.46 ± 0.48	84.48 ± 0.34	84.42 ± 0.13	84.41 ± 0.06
	GMCT	Precision	**65.49 ± 1.00**	64.92 ± 2.72	64.80 ± 6.82	64.57 ± 1.03	64.78 ± 0.44
		AUC-ROC	**67.77 ± 1.48**	66.30 ± 2.30	67.30 ± 6.50	66.62 ± 1.86	67.56 ± 0.19
	Amazon - Industrial	Precision	**81.95 ± 0.54**	81.74 ± 0.93	80.72 ± 1.98	81.70 ± 0.89	81.34 ± 0.48
		AUC-ROC	**80.00 ± 0.60**	79.64 ± 1.26	79.03 ± 1.08	79.52 ± 0.99	79.10 ± 0.17
	Amazon - Office	Precision	**78.16 ± 0.31**	76.88 ± 1.49	77.27 ± 0.71	77.11 ± 1.00	75.05 ± 1.03
		AUC-ROC	**75.38 ± 1.60**	74.98 ± 1.33	74.59 ± 0.45	74.73 ± 0.78	73.28 ± 1.05
	Yelp	Precision	**89.39 ± 0.05**	89.24 ± 0.37	88.99 ± 0.11	88.98 ± 0.10	88.92 ± 0.01
		AUC-ROC	**89.10 ± 0.05**	89.28 ± 0.28	88.92 ± 0.12	89.01 ± 0.10	88.89 ± 0.09

Effect of target guidance. When we replace the context cross-attention layer with a direct self-attention layer for the concatenated historical and structural embeddings, performance drops consistently across datasets and tasks. This proves that multimodal utility in dynamic graphs is target-dependent. The same modality can be helpful for some prediction targets, but distracting for others. Without target-guided querying, the fusion module tends to produce a target-agnostic mixture, which amplifies irrelevant signals and weakens the model's ability to focus on evidence that is most informative for the queried target.

Effect of the hierarchical architecture. Replacing the hierarchical two-stage fusion with a single-stage attention (querying the concatenated structural, historical, and multimodal features with the prediction targets) consistently degrades performance. This suggests that one-shot fusion is more susceptible to modality noise and misalignment, whereas the two-stage design first forms a target-aware spatiotemporal context and then uses it to regulate cross-modal interplay, yielding more stable fusion under dynamic graphs.

Effect of spatial vs. historical context. Finally, removing either the spatial neighborhood context or the historical temporal context harms performance. These results confirm that multimodal fusion in dynamic graphs should be conditioned on both *where* the target lies in the evolving graph and *when* it is observed along its interaction history. Relying on only one source of context provides incomplete cues for deciding which modality to trust.

Effect of modality. The effect of a modality is neither consistent in direction nor uniform across instances. For example, for link prediction, the same modalities are beneficial on average on Yelp but harmful on average on GDELT, and even where they help on average, they degrade a substantial share of individual predictions. A fusion rule that treats each modality as uniformly informative is therefore ill-suited to multimodal dynamic graphs, which motivates conditioning the fusion on the prediction target, as in **TST-MF**.

F Broader Impacts

This work aims to support research on multimodal dynamic graph learning by providing datasets, tasks, baselines, and a target-aware fusion framework. MM-DyGraph can help researchers evaluate when multimodal information is useful in dynamic relational systems, rather than assuming that adding more modalities always improves performance. By revealing target-level inconsistency, it encourages more careful model design and evaluation for applications such as recommendation, urban computing, and event analysis.

However, multimodal dynamic graph learning may also raise societal risks if used without safeguards. Dynamic graphs can encode evolving relationships among users, organizations, locations, and events, while text and image attributes may make learned representations more semantically informative. As a result, models trained on such data could potentially amplify biases, support user profiling, or produce misleading predictions when relying on noisy or spurious multimodal cues.

To mitigate these risks, MM-DyGraph is intended for research evaluation rather than direct high-stakes deployment. Users should follow data licenses and terms of use, avoid attempts to identify individuals, and refrain from using the benchmark for surveillance or automated profiling.

G Limitations

Although MM-DyGraph provides a benchmark for multimodal dynamic graph learning, several limitations remain. First, dataset-specific pruning and VLM-based filtering improve multimodal quality but may bias the data toward clearer, higher-quality reviews or interactions, underrepresenting noisy, sparse, or weakly aligned real-world samples. Second, MM-DyGraph mainly focuses on text and image modalities, leaving other signals such as audio, video, tabular metadata, and sensor data unexplored. Third, this work focuses on dynamic graph structure with multimodal attributes and does not systematically cover successive updates to the raw text or images of the same entity. Datasets with explicit entity-level feature updates are an important direction for future works. Finally, VLM-based filtering is model- and prompt-dependent and cannot guarantee perfect cross-modal alignment.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main contributions, including the MM-DyGraph benchmark, empirical observations on target-level inconsistency, and the TST-MF framework. The claims are supported by the dataset description, evaluation tasks, empirical analysis, and ablation studies.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations in Appendix G, including potential bias introduced by pruning and VLM-based filtering, the focus on text and image modalities, limited domain coverage, and the model-/prompt-dependence of VLM-based quality assessment.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper does not present theoretical results, formal theorems, or proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We describe the dataset construction, evaluation tasks, feature extraction, baselines, and model architecture, with additional implementation details provided in the appendix. And the processed datasets and code are also released.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide anonymized access to the code and processed datasets.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify experimental setting in Section 4.1, and additional experimental details in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report mean performance with standard deviations over multiple runs for the main experiments, as shown in the experimental tables.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the computational resources used for the experiments in Appendix C.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed and followed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss broader impacts in Appendix F.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: we release the benchmark with preprocessed embedding rather than the raw text and images to avoid safety risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original sources of the datasets, pretrained encoders, and baseline methods used in this work.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide anonymous link for the code and datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve new crowdsourcing, user studies, or direct interaction with human participants.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve new human-subject experiments or newly collected participant data requiring IRB approval.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We use VLM as a quality assessment tool during dataset construction to score cross-modal relevance and consistency. The model is used for filtering noisy image-text associations, and the prompt and filtering procedure are described in the paper and appendix.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.