Event Knowledge Graph Narrative Planning for Faithful and Coherent Multi-Document Summarization

Yukang Huang[1], Tongshu Liu[2], Wei Liu[*1], Weimin Li[1],

[1]School of Computer Engineering and Science, Shanghai University, Shanghai, China
[2]College of Arts and Sciences, The Ohio State University, Columbus, OH, USA

***Abstract*—Multi-Document Summarization (MDS) remains challenging for Large Language Models (LLMs) due to difficulties in maintaining factual consistency, temporal coherence, and event-level faithfulness across sources. Existing approaches often rely on graph linearization, requiring LLMs to implicitly reconstruct structure and leading to suboptimal narrative quality.**

We propose Event Knowledge Graph Narrative Planning (EKGNP), a neuro-symbolic framework that separates event-level planning from surface realization. EKGNP constructs an event-centric graph and formulates summarization as a planning problem, where a neural-guided MCTS planner generates coherent event sequences. We further introduce a Sinkhorn-based metric to measure structural faithfulness.

Experiments on Multi-News, WCEP, and EventSum show that EKGNP improves factual consistency, temporal accuracy, and structural alignment over strong baselines. Results demonstrate that explicit event-level planning provides an effective solution for faithful and coherent multi-document summarization.

***Index Terms*—multi-document summarization, event knowledge graphs, narrative planning, neuro-symbolic methods, hallucination reduction**

I. Introduction

Multi-Document Summarization (MDS) is widely used for aggregating information across multiple sources such as news aggregation, intelligence analysis, and decision support. Recent large language models (LLMs), such as GPT-4 and LLaMA-3, have shown strong performance in abstractive summarization. However, their limitations become evident in multi-document settings. Prior work reports systematic failures such as contextual amnesia, temporal collapse, and event conflation [1], where distinct events are merged, actions are misattributed across entities, and factually unsupported causal relations are introduced. Several studies report that these errors persist even under high-quality retrieval settings, suggesting that retrieval alone may be insufficient to address underlying architectural limitations. Autoregressive LLMs generate text as linear token sequences and lack explicit mechanisms for reasoning over long-horizon event structures [2].

Most existing approaches attempt to address these limitations through Retrieval-Augmented Generation (RAG) and expanded context windows. While effective for information coverage, such strategies remain limited for narrative synthesis. Standard RAG pipelines retrieve fragmented passages that preserve lexical relevance but fail to encode temporal or causal continuity [3]. More recent GraphRAG-style methods [4] introduce graph representations to impose structure, yet in most cases the graph is subsequently linearized into a textual prompt. As graph complexity increases, linearization obscures the original topology and requires LLMs to implicitly reconstruct structural relations from sequential text. This often leads to degraded generation fidelity, a phenomenon also observed in prior studies on structured-to-text linearization [5].This limitation is not specific to a single method, but is common across graph-enhanced summarization pipelines, where narrative-level decisions are still made implicitly during token-level decoding without explicit global planning.

To address this limitation, in this work, we introduce Event Knowledge Graph Narrative Planning (EKGNP), which separates global event-level planning from surface realization. EKGNP formulates summarization as a planning problem over Event Knowledge Graphs, using a neural-guided Monte Carlo Tree Search (MCTS) planner to construct a globally coherent narrative plan prior to text generation. The plan is optimized with a contrastive novelty-coherence objective that balances informational coverage and logical consistency, allowing event-level structure to be explicitly modeled before text generation. The workflow of the planner is illustrated in Figure 1.

By separating what to say from how to say it, EKGNP addresses the lack of global planning in autoregressive generation. Text is generated by conditioning on a verified narrative plan rather than on a linearized graph prompt, reducing hallucinations and improving temporal and causal consistency across documents.

We evaluate EKGNP on three widely used benchmarks: Multi-News, WCEP, and EventSum. Results show that EKGNP consistently outperforms strong LLM-based and graph-enhanced baselines in factual consistency, narrative coherence, and structural faithfulness, with substantial gains over GraphRAG on FActScore and temporal accuracy.

Our contributions are summarized as follows:

A Neuro-Symbolic Framework for MDS: We propose an

[1*]Corresponding author: liuw@shu.edu.cn

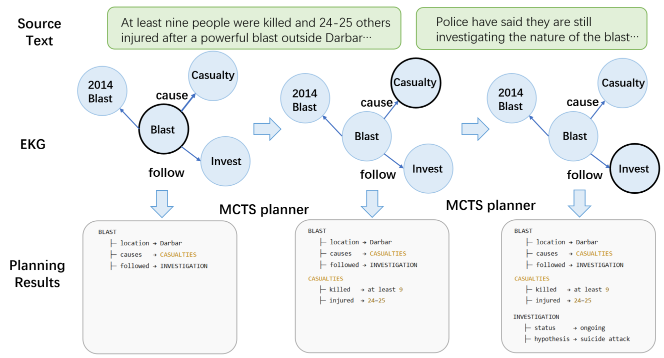

Fig. 1. Illustration of event-level narrative planning with neural-guided Monte Carlo Tree Search (MCTS). Starting from an initial state, the planner explores sequences of event nodes under temporal, causal, and coreferential constraints. Each transition is scored using a log-linear reward that combines coherence, novelty, and graph centrality. The selected path defines a globally coherent narrative plan.

end-to-end framework that integrates event knowledge graph construction with explicit narrative planning.

Neural-Guided Event-Level Narrative Planning: We introduce a learnable Neural MCTS planner operating directly on event graphs to optimize global coherence and informational novelty.

Topological Evaluation: We design a differentiable optimal-transport-based metric for quantifying structural faithfulness beyond surface-level lexical overlap.

II. Related Work

A. LLM-Based Summarization and Hallucination

While LLMs have largely superseded sequence-to-sequence models in zero-shot summarization, prior work has shown that they introduce substantial hallucination risks in multi-document settings, particularly in maintaining temporal consistency and source attribution [6]. Belem et al. [7] categorize these failures as "source confusion," where models incorrectly merge distinct events or timelines across documents. Such errors persist even with extended context windows, reflecting structural limitations of autoregressive decoding in long-horizon reasoning [1].

B. Event Knowledge Graph Construction

Event-centric representations have been widely explored to support structured understanding and aggregation of information across multiple documents [8]. Event Knowledge Graphs (EKGs) explicitly encode entities, events, and their relations, providing a natural foundation for multi-document summarization [9] that requires preserving temporal, causal, and coreferential structure. A key prerequisite is accurate event extraction and cross-document coreference resolution.

In practice, many summarization systems treat EKGs as static intermediate artifacts that are subsequently linearized into textual prompts for LLMs [4], [10]. While this strategy preserves graph-derived information, it converts topological structure into sequential text, requiring models to implicitly recover relational dependencies during generation. As graph size and relational complexity increase, this reliance on implicit reconstruction can limit the effectiveness of static graph prompting for maintaining global narrative coherence.

C. Neuro-Symbolic Planning

Neuro-symbolic approaches seek to integrate neural language models with explicit reasoning or planning mechanisms. In language generation, planning-based methods have been explored to impose global control over content selection and ordering, including early content planning and text planning frameworks [11]. Recent frameworks such as Tree of Thoughts [12] formulate generation as a search process over intermediate reasoning states, improving performance on multi-step reasoning tasks.However, these methods operate at the token or text level, offering limited support for reasoning over structured event graphs, especially in multi-document settings.

III. Methodology

Motivated by the computational and structural limitations observed in prior multi-document summarization approaches, we design the Event Knowledge Graph Narrative Planning (EKGNP) architecture to explicitly model narrative structure during generation. The framework operates in four phases: Event Knowledge Graph Construction, Neural Narrative Planning, Graph Prompt, and Structural Verification. The framework is illustrated in Figure 2.

A. Problem Formalization

Let $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ be a cluster of related documents describing a complex, evolving event. Our objective is to generate a summary $\mathcal{S}$ that covers salient events ($e \in \mathcal{E}$) while respecting the temporal and causal constraints inherent in the underlying event structure. We formulate this not as a direct translation $\mathcal{D} \to \mathcal{S}$, but as a graph-search-then-realize process: $\mathcal{D} \to G \to \pi^* \to \mathcal{S}$.

The transformation $\mathcal{D} \to G$ critically depends on accurate cross-document coreference resolution to ensure that event mentions referring to the same real-world occurrence are merged into single nodes, preventing graph fragmentation and redundant planning.

B. Phase 1: Event Knowledge Graph Construction

The goal of this phase is to construct a unified, event-centric graph that preserves cross-document identity, temporal order, and causal structure. The objective is to transform $\mathcal{D}$ into a structured Event Knowledge Graph $G = (V, E)$. Given the scale and heterogeneity of document clusters such as those in WCEP, we adopt the neural xCoRe framework to perform cross-document coreference resolution in a robust and scalable manner.

1) Step 1: Open-Schema Event Argument Identification: We employ GLiNER [13], a generalist sequence labeling model, to identify event triggers and associated arguments under an open-schema event representation. GLiNER is configured to recognize event triggers and core event arguments, including

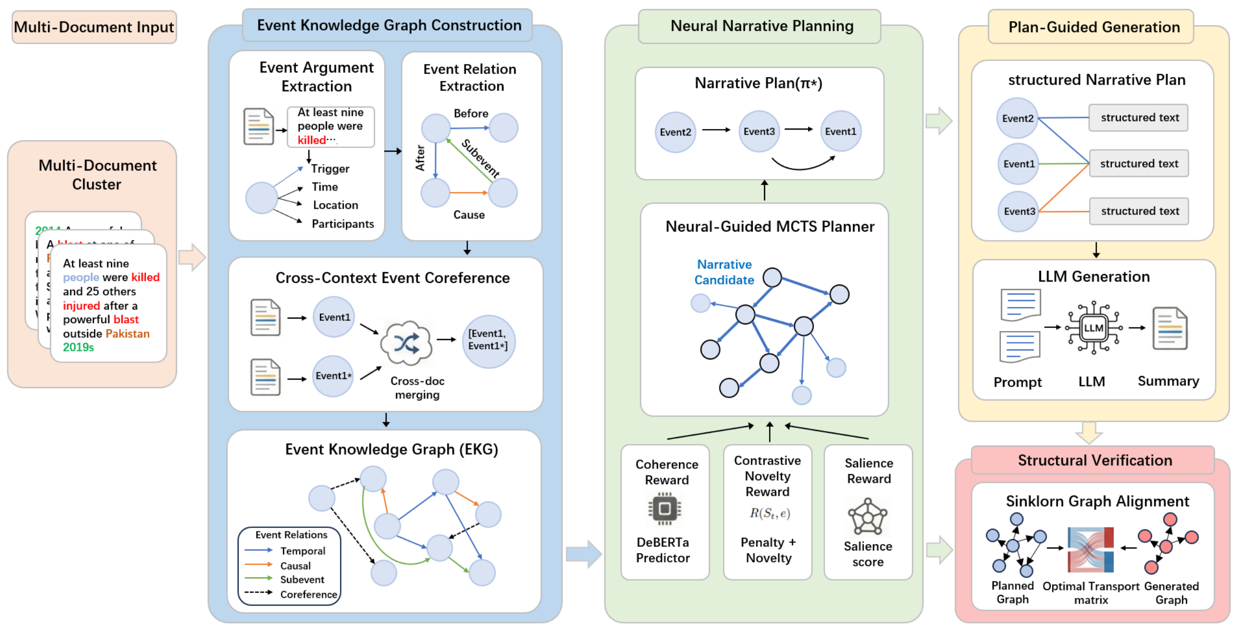

Fig. 2. Overview of the Event Knowledge Graph Narrative Planning (EKGNP) framework.

⟨EventTrigger⟩, ⟨Time⟩, ⟨Location⟩, ⟨Participant⟩. For each sentence s_{ij}, we extract the set of triggers $\mathcal{T}_{ij} = \text{GLiNER}(s_{ij})$.

2) Step 2: Event-Event Relations: To construct a structured Event Knowledge Graph from raw multi-document inputs, We adopt a pretrained MAVEN-ERE [14] relation extractor to identify event relations, including coreference, temporal, causal, and subevent relations.

3) Step 3: Cross-Context Coreference Resolution: We adopt xCoRe [15] to perform cross-document coreference resolution via mention extraction, clustering, and cross-context merging.

Specifically, for each document $d_i \in \mathcal{D}$ treated as a separate context, xCoRe first extracts mentions using a start-to-end span prediction approach with Transformer-based encodings. Local clusters are then formed using a multi-expert mention-pair scorer that assigns specialized scorers based on linguistic types (e.g., pronoun-pronoun, entity-pronoun, exact match). Finally, cluster representations are computed via a single-layer Transformer, and cross-document clusters are merged by predicting pairwise coreference probabilities $p_{cm}(W^a_{c_i}, W^b_{c_j})$ between cluster pairs from different contexts using a bilinear classifier.

C. Phase 2: Neural-Guided MCTS Narrative Planner

We model the selection of key events as a sequential decision process solved via neural-guided Monte Carlo Tree Search (MCTS). Starting from an empty narrative state, MCTS incrementally constructs a narrative plan by expanding candidate event nodes from the Event Knowledge Graph. At each step, the current state corresponds to a partial event sequence, and each action selects a candidate next event to append to the sequence. The search follows the standard MCTS procedure of selection, expansion, simulation, and backpropagation, while using neural and graph-based signals to evaluate candidate transitions. After repeated simulations, the planner identifies a high-reward event path $\pi^* = [v_1, v_2, \ldots, v_k]$ through the graph, which is then used as the narrative plan for generation. To guide this search process, we define a transition-level reward function as follows.

To score candidate transitions during MCTS, we adopt a log-linear reward function rather than a multiplicative formulation, which can be unstable when any individual factor becomes very small. The reward $R(S_t, e)$ for transitioning from the current partial plan S_t to a candidate node e is defined as:

$$R(S_t, e) = \alpha \cdot \log P_{\text{coh}}(e|S_t) + \beta \cdot \mathcal{N}(e, S_t) + \gamma \cdot \mathcal{C}(e)$$

where S_t denotes the current partial event sequence (state) at step t, e is a candidate next event node, and $\alpha, \beta, \gamma \in \mathbb{R}^+$ are weighting coefficients that control the contribution of coherence, novelty, and salience respectively.

Coherence (P_{coh}): A lightweight DeBERTa-based "Next-Event Predictor" [16] estimates the transition probability in milliseconds:

$$P_{\text{coh}}(e|S_t) = \text{softmax}\left(\text{DeBERTa}([S_t; e])\right)$$

Contrastive Novelty ($\mathcal{N}$): We impose a hard penalty on redundancy by minimizing the maximum similarity against all prior nodes in the history:

$$\mathcal{N}(e, S_t) = 1 - \max_{e' \in S_t}\left(\frac{\mathbf{h}_e^\top \mathbf{h}_{e'}}{\|\mathbf{h}_e\|\|\mathbf{h}_{e'}\|}\right)$$

where h_e and $h_{e'}$ denote the embedding representations of events e and e' respectively.

Salience ($\mathcal{C}$): The PageRank centrality of node e within the global graph G.

The MCTS explores the state space using the UCT algorithm, balancing exploration of new narrative arcs with the exploitation of high-coherence paths.

D. Phase 3: Plan-Guided Prompt Construction

1) Plan Serialization Format

The optimal path π^* is serialized into a structured logical form. Unlike standard "triple-based" serialization which lists disjoint facts, our format emphasizes the transition between events to guide the model's discourse flow. Each step in the plan is expanded with its trigger, arguments, and the specific relation type that justified the transition in the graph.

The serialization template is defined as follows:

```
Event 1:
    Time: <Time_1> | Location: <Loc_1>
    Actor: <Actor_1> :: DID <Trigger_1>
    ↓ (Relation_1_2)
Event 2:
    Time: <Time_2> | Location: <Loc_2>
    Actor: <Actor_2> :: DID <Trigger_2>
    ↓ (Relation_2_x)
    ...
```

2) Contextual Grounding with Plan-Relevant Retrieval

To ground the abstract narrative plan in source text, we perform a targeted retrieval step that selects evidence sentences relevant to the planned events. For each event node $e \in \pi^*$, we retrieve the top-k most relevant source sentences from the document cluster $\mathcal{D}$, denoted as $C(\pi^*)$. In practice, retrieval is limited to events included in the plan, which helps reduce interference from unrelated sub-stories, present in the source documents (a common issue in WCEP).

3) Instruction Tuning

The prompt fed to the Large Language Model (Llama-3-8B-Instruct) wraps the serialized plan and retrieved context in a strict instruction set designed to enforce structural faithfulness:

Prompt Template

System: You are an expert investigative journalist. You will be provided with a Narrative Plan consisting of a chronologically ordered sequence of events.
Task: You are given an ordered narrative plan extracted from multiple news reports. Generate a factual and coherent summary strictly following the events below.
Input Plan: $\langle \pi^* \rangle$
Source Context: $[C(\pi^*)]$
Constraints: - Do not merge current and historical events. - Do not invent casualty numbers. - Clearly distinguish confirmed facts from speculation.

Compared to full-graph linearization, this strategy reduces the prompt length from $O(|V| + |E|)$ to $O(|\pi^*|)$, enabling richer contextual grounding while constraining generation to the verified narrative plan.

E. Phase 4: Structural Verification via Sinkhorn Distance

To measure structural faithfulness in a differentiable manner, we introduce the Sinkhorn Graph Alignment Score (SGAS), which is based on entropy-regularized optimal transport.

Let the graph induced by the generated summary be G_{gen} with node embeddings $\mathbf{X}$, and the source plan be G_{src} with embeddings $\mathbf{Y}$. We seek an optimal transport plan $\mathbf{P}^*$ that minimizes the semantic cost matrix $\mathbf{C}_{ij} = 1 - \cos(\mathbf{x}_i, \mathbf{y}_j)$. The objective is:

$$L_{OT}(\mathbf{P}) = \sum_{i,j} \mathbf{P}_{ij}\mathbf{C}_{ij} + \epsilon H(\mathbf{P})$$

where $H(\mathbf{P})$ is the entropic regularization term. We solve for $\mathbf{P}^*$ using Sinkhorn-Knopp iterations:

$$\mathbf{u}^{(k+1)} = \frac{\mathbf{a}}{\mathbf{K}\mathbf{v}^{(k)}}, \quad \mathbf{v}^{(k+1)} = \frac{\mathbf{b}}{\mathbf{K}^\top \mathbf{u}^{(k+1)}}$$

with $\mathbf{K} = \exp(-\mathbf{C}/\epsilon)$. The final SGAS metric is defined as:

$$\text{SGAS} = \exp\left(-\frac{1}{\eta}\langle \mathbf{P}^*, \mathbf{C}\rangle_F\right)$$

where η is a temperature scaling factor, and $\langle \cdot, \cdot \rangle_F$ denotes the Frobenius inner product.

IV. Experimental Setup

A. Datasets and Baseline Systems

We evaluate our proposed EKGNP framework on three widely adopted multi-document summarization benchmarks: Multi-News [17], WCEP [18], and EventSum [19]. Multi-News consists of 56,216 news clusters, each containing approximately 2-3 related articles, paired with professionally written abstractive summaries of around 260 words. WCEP contains 22,778 document clusters with significantly higher redundancy, specifically challenging the scalability of graph-based approaches under high-density conditions. EventSum is a Chinese multi-document summarization dataset comprising approximately 1,200 event clusters. We use an English version of EventSum obtained via LLM-based machine translation.

To comprehensively assess performance, we compare EKGNP against representative baseline systems across four paradigms: (1) Fine-tuned pretrained models including PRIMERA [20] and PEGASUS [21], which leverage hierarchical attention and gap-sentence pre-training respectively; (2) LLM-based approaches encompassing LLaMA-3-8B under zero-shot and 3-shot settings, alongside GPT-4-Turbo with optimized prompting; (3) Graph-enhanced models comprising CoKG [10](iterative knowledge graph refinement), and GraphRAG [3] (entity-centric retrieval augmentation).

B. Evaluation Metrics

We evaluate system performance from multiple complementary perspectives. Surface-level summary quality is assessed using ROUGE-1/2/L [22]. Faithfulness is measured with FActScore [23], which evaluates factual consistency by

TABLE I
MAIN RESULTS ON MULTI-NEWS AND WCEP DATASETS.

Model	Multi-News				WCEP			
	R-1	R-2	R-L	FActScore	R-1	R-2	R-L	FActScore
Fine-tuned Pretrained Models								
PRIMERA	32.8	10.2	16.1	52.3	22.1	8.4	13.7	48.7
PEGASUS	31.2	10.1	17.8	50.8	20.5	7.9	13.4	47.2
LLM-based Approaches								
LLaMA-3 (Zero-shot)	41.7	12.8	21.1	72.6	36.2	12.1	23.1	68.3
LLaMA-3 (Few-shot)	43.9	13.9	22.9	74.2	37.8	13.5	24.4	70.8
GPT-4-Turbo	42.5	13.5	22.3	77.9	36.7	12.9	23.8	74.1
Graph-Enhanced Models								
CoKG	43.8	14.2	22.4	81.3	39.2	14.1	22.1	75.7
GraphRAG	43.1	14.3	22.2	82.5	39.8	14.7	22.9	76.3
EKGNP (Ours)	**44.2**	**14.5**	**23.1**	**84.1**	**40.5**	**15.2**	**24.3**	**77.8**

decomposing summaries into atomic facts, along with entity hallucination rate to quantify unsupported entity mentions.

To assess structural fidelity, we adopt three event-centric metrics: Event Recall, measuring the proportion of reference events covered by the generated summary; Temporal Accuracy, evaluating the correctness of event ordering; and the proposed Sinkhorn Graph Alignment Score (SGAS), which quantifies structural alignment between the generated summary and the source narrative plan.

C. Implementation Details

Our framework uses Llama-3-8B-Instruct as the base language model for summary generation. Event extraction is performed using GLiNER-large-v2 with a custom event-centric schema, and cross-document coreference resolution is handled by xCoRe with a DeBERTa-v3 encoder. Narrative planning is implemented using Monte Carlo Tree Search (MCTS), with a DeBERTa-based coherence predictor fine-tuned on event transition data from Multi-News.

The MCTS planner is configured with 200 simulations per step and a maximum plan length of 15 events. Reward weights are set to $\alpha = 1.0$ for coherence, $\beta = 2.5$ for novelty, and $\gamma = 0.5$ for salience. Sinkhorn regularization is set to $\epsilon = 0.1$. All reported results are conducted on NVIDIA RTX 5090 GPUs (32GB memory), and averaged over three runs with different random seeds.

V. EXPERIMENT RESULTS

A. Overall Performance Analysis

Table I summarizes the results on the Multi-News and WCEP benchmarks. Across both datasets, EKGNP consistently outperforms all compared baselines on both ROUGE and faithfulness metrics.

On Multi-News, EKGNP attains 44.2 ROUGE-1, 14.5 ROUGE-2, and 23.1 ROUGE-L, outperforming the strongest baseline GraphRAG by 1.1, 0.2, and 0.9 points respectively.

TABLE II
STRUCTURAL FAITHFULNESS METRICS ON EVENTSUM (ENGLISH SUBSET). ER = EVENT RECALL, AR = ARGUMENT RECALL, CR = CAUSAL RECALL, TR = TEMPORAL RECALL, TA = TEMPORAL ACCURACY, SGAS = SINKHORN GRAPH ALIGNMENT SCORE.

Model	ER	AR	CR	TR	TA	SGAS
LLaMA-3	15.1	26.9	24.4	20.2	42.7	0.521
GPT-4-Turbo	21.7	46.2	56.1	40.0	58.3	0.647
CoKG	23.5	43.7	57.9	42.3	61.2	0.658
GraphRAG	19.8	41.3	52.6	38.7	56.8	0.612
EKGNP (Ours)	**25.7**	**50.3**	**67.3**	**56.8**	**68.5**	**0.701**

In terms of factual consistency, EKGNP reaches a FActScore of 84.1, improving upon GraphRAG by 1.6 points and vanilla LLaMA-3 by 6.2 points. The improvements are particularly pronounced on WCEP, where document clusters exhibit higher redundancy and temporal complexity. EKGNP obtains 40.5 ROUGE-1 and 77.8 FActScore, establishing +0.7 and +1.5 improvements over GraphRAG.

Comparing against LLM-based approaches, EKGNP consistently outperforms both zero-shot and few-shot LLaMA-3 configurations despite using the same base model for generation. This performance gap—ranging from 2.5 to 11.5 FActScore points. Overall, the results indicate that explicitly modeling event structure can be more effective than relying solely on prompt design. Even GPT-4-Turbo, despite its substantially larger parameter count, underperforms EKGNP by 6.2 FActScore on Multi-News, suggesting that model scale cannot compensate for the absence of explicit structural reasoning.

B. Structural Faithfulness Evaluation

Table II evaluates structural fidelity on EventSum using fine-grained event-centric metrics. EKGNP achieves 25.7% Event Recall, 50.3% Argument Recall, and 67.3% Causal Recall, substantially outperforming all baselines. Compared to CoKG, EKGNP improves Causal Recall by 9.4 points, indicating a

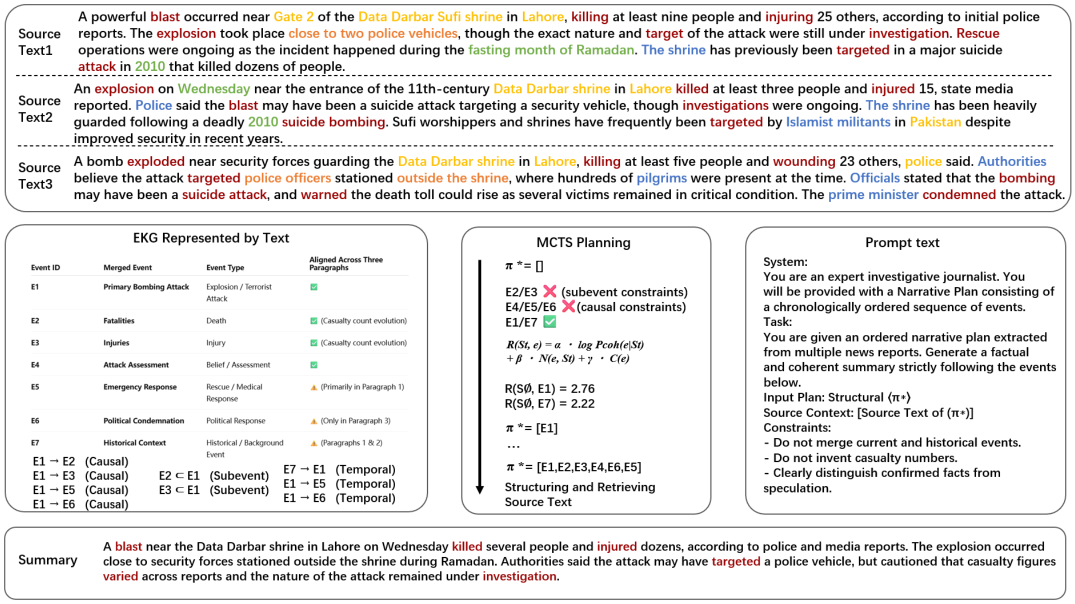

Fig. 3. An Example of Event Knowledge Graph Narrative Planning under Conflicting News Reports

stronger alignment with causal relations encoded in the event structure.

Temporal Accuracy—measuring the proportion of correctly ordered event pairs—reaches 68.5% for EKGNP, a +7.3 improvement over CoKG and +10.2 over GPT-4-Turbo. The comparison results indicate that enforcing a global event order prior to generation plays a critical role. Temporal inconsistencies often arise during purely autoregressive decoding, and pre-enforcing such a global event order can effectively mitigate such temporal inconsistencies. EKGNP also achieves the highest SGAS score (0.701), exceeding CoKG by 0.043, which reflects closer alignment between the generated summaries and the planned event structures.

Notably, GraphRAG underperforms CoKG on structural metrics despite competitive ROUGE scores, revealing a critical limitation: entity-centric retrieval fragments event narratives into disconnected entity subgraphs, degrading temporal and causal coherence. EKGNP's event-centric formulation explicitly addresses this by treating events as first-class citizens in both graph construction and planning.

C. Ablation Study

Table III quantifies the contribution of each architectural component. Removing MCTS planning and reverting to static graph prompting causes a severe -4.8 FActScore and -0.066 SGAS degradation, confirming that linearization of graph structure forces LLMs to implicitly reconstruct topology—a process prone to failure under complex narrative arcs.

Ablating the Contrastive Novelty reward ($\mathcal{N}$) and retaining only coherence-based selection reduces FActScore by -2.4 points, demonstrating that greedy coherence maximization leads to redundant event chains. The novelty term actively penalizes semantic overlap, ensuring informational diversity across the narrative plan. Interestingly, ROUGE scores remain relatively stable (-0.3 to -0.7), suggesting that surface n-gram metrics are insensitive to redundancy—a known limitation of lexical overlap measures.

TABLE III
ABLATION STUDY ON MULTI-NEWS TEST SET. EACH ROW REMOVES ONE COMPONENT FROM THE FULL EKGNP SYSTEM. Δ SHOWS ABSOLUTE CHANGE FROM FULL SYSTEM.

Model Variant	R-1	R-2	FActScore	SGAS
EKGNP (Full)	**44.2**	**14.5**	**84.1**	**0.687**
w/o MCTS Planning ↪ (Static Graph Prompt)	42.8 (-1.4)	13.9 (-0.6)	79.3 (-4.8)	0.621 (-0.066)
w/o Novelty Reward ↪ (Coherence Only)	43.5 (-0.7)	14.2 (-0.3)	81.7 (-2.4)	0.659 (-0.028)
w/o EKG ↪ (RAG Baseline)	41.3 (-2.9)	13.1 (-1.4)	76.5 (-7.6)	0.583 (-0.104)

The most severe degradation occurs when removing the Event Knowledge Graph entirely, reducing the system to a standard RAG pipeline with dense retrieval. Removing the event graph leads to substantial drops in both FActScore and SGAS, indicating that explicit graph structure plays a critical role in maintaining factual and structural consistency. Without explicit event nodes and relations, the model lacks the structural scaffolding necessary to maintain temporal and causal consistency.

D. Case Study

Figure 3 shows a case study involving three news reports describing an explosion near the Data Darbar shrine in Lahore.

Event extraction and cross-document alignment yield seven event nodes. A single core event, E_1(Primary Bombing Attack), is aligned across all sources, while E_2 (Fatalities) and E_3 (Injuries) are modeled as subevents with evolving counts. Speculative assessments (E_4) and reactions (E_5-E_6) are causally downstream of E_1, and the 2010 suicide bombing is abstracted as a separate historical event (E_7).

During neural MCTS planning, subevents and reactions are initially pruned due to structural constraints, leading the planner to prioritize E1 over historical context. The final narrative plan $\pi^* = [\mathrm{E}_1, \mathrm{E}_2, \mathrm{E}_3, \mathrm{E}_4, \mathrm{E}_6, \mathrm{E}_5]$ emphasizes the current attack and its verified consequences while excluding historical background. The generated summary correctly reflects uncertainty, preserves temporal and causal structure, and avoids conflating past and present events.

VI. Limitations

Despite improvements in structural faithfulness and hallucination reduction, EKGNP has several limitations.

Error Propagation from Event Graph Construction: The quality of generated summaries depends on the accuracy of the underlying event knowledge graph. Errors in event extraction or relation construction may propagate to final outputs. While the planning module can downweight unreliable nodes, it cannot fully eliminate upstream inaccuracies.

Computational Trade-offs in MCTS Planning: The cost of Monte Carlo Tree Search (MCTS) increases with graph density, leading to longer planning time in high-redundancy datasets such as WCEP. This highlights scalability challenges for explicit search-based planning.

Dataset Limitations: EKGNP is designed for event-centric summarization and evaluated on event-heavy benchmarks. While effective for news domains, its generalization to domains with less explicit event structure (e.g., legal or medical documents) remains uncertain.

VII. Conclusion

In this paper, we propose EKGNP, a planning-based framework for multi-document summarization that separates global narrative planning from surface generation. By constructing an event-centric knowledge graph and applying Monte Carlo Tree Search over event sequences, EKGNP mitigates the limitations of linearized graph prompting and retrieval-augmented methods.

Experiments on Multi-News, WCEP, and EventSum show that EKGNP consistently improves factual consistency and structural faithfulness.

Overall, our results highlight the value of explicit structural planning for improving coherence in long-form generation, suggesting that purely autoregressive decoding is insufficient for multi-document synthesis. EKGNP demonstrates a promising direction for integrating structured reasoning with neural text generation.

References

[1] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the middle: How language models use long contexts," *Transactions of the Association for Computational Linguistics*, pp. 157–173, 2024.

[2] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," in *International Conference on Learning Representations (ICLR)*, 2020.

[3] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, and J. Larson, "From local to global: A graph rag approach to query-focused summarization," *arXiv preprint arXiv:2404.16130*, 2024.

[4] Y. Lei and R. Huang, "Multi-document summarization through multi-document event relation graph reasoning in llms: A case study in framing bias mitigation," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025, pp. 26 603–26 619.

[5] R. Koncel-Kedziorski, D. Bekal, Y. Luan, M. Lapata, and H. Hajishirzi, "Text generation from knowledge graphs with graph transformers," in *Proceedings of NAACL 2019*, 2019, pp. 2284–2293.

[6] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On faithfulness and factuality in abstractive summarization," in *Proceedings of ACL 2020*, 2020, pp. 1906–1919.

[7] C. G. Belém, P. Pezeshkpour, H. Iso, S. Maekawa, N. Bhutani, and E. Hruschka, "From single to multi: How llms hallucinate in multi-document summarization," in *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025, pp. 5276–5309.

[8] S. Guan, X. Cheng, L. Bai, F. Zhang, Z. Li, Y. Zeng, X. Jin, and J. Guo, "What is event knowledge graph: A survey," *IEEE Transactions on Knowledge and Data Engineering*, pp. 7569–7589, 2023.

[9] N. Chambers and D. Jurafsky, "Unsupervised learning of narrative event chains," in *Proceedings of ACL-08*, 2008, pp. 789–797.

[10] K. Lee, J. Jang, Y. Lim, and M. Shin, "Chain of knowledge graph: Information-preserving multi-document summarization for noisy documents," in *Proceedings of COLING 2025 Workshop*, 2025, pp. 1–5.

[11] A. Moryossef, Y. Goldberg, and I. Dagan, "Step-by-step: Separating planning from realization in neural data-to-text generation," in *Proceedings of NAACL 2019*, 2019, pp. 2267–2277.

[12] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," in *Advances in Neural Information Processing Systems*, 2023, pp. 11 809–11 822.

[13] U. Zaratiana, N. Tomeh, P. Holat, and T. Charnois, "Gliner: Generalist model for named entity recognition using bidirectional transformer," in *Proceedings of NAACL 2024*, 2024, pp. 5364–5376.

[14] X. Wang, Y. Chen, N. Ding, H. Peng, Z. Wang, Y. Lin, X. Han, L. Hou, J. Li, Z. Liu, P. Li, and J. Zhou, "Maven-ere: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction," in *Proceedings of EMNLP*, 2022.

[15] G. Martinelli, B. Gatti, and R. Navigli, "xcore: Cross-context coreference resolution," in *Proceedings of EMNLP 2025*, 2025, pp. 34 264–34 278.

[16] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced bert with disentangled attention," in *International Conference on Learning Representations*, 2021.

[17] A. Fabbri, I. Li, T. She, S. Li, and D. Radev, "Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model," in *Proceedings of ACL 2019*, 2019, pp. 1074–1084.

[18] D. Gholipour Ghalandari, C. Hokamp, N. T. Pham, J. Glover, and G. Ifrim, "A large-scale multi-document summarization dataset from the wikipedia current events portal," in *Proceedings of ACL 2020*, 2020, pp. 1302–1308.

[19] M. Zhu, K. Zeng, M. Wang, K. Xiao, L. Hou, H. Huang, and J. Li, "Eventsum: A large-scale event-centric summarization dataset for chinese multi-news documents," in *Proceedings of AAAI*, 2025, pp. 26 138–26 147.

[20] W. Xiao, I. Beltagy, G. Carenini, and A. Cohan, "Primera: Pyramid-based masked sentence pre-training for multi-document summarization," in *Proceedings of ACL 2022*, 2022, pp. 5245–5263.

[21] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," in *Proceedings of ICML*, 2020, pp. 11 328–11 339.

[22] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, 2004, pp. 74–81.

[23] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, "Factscore: Fine-grained atomic evaluation of factual precision in long form text generation," in *Proceedings of EMNLP 2023*, 2023, pp. 12 076–12 100.