RETELL: RST-Centric Related Work Generation with LLM

Anonymous ACL submission

Abstract

Related work cherishes structural relationships. While existing automated methods, including those leveraging Large Language Models (LLMs), have advanced content summarization capabilities, they often struggle to replicate the crucial argumentative flow and explicit inter-paper relational structures found in human-written related work, despite many commendable efforts in recent years. To address this concern, we propose a novel approach centered on Rhetorical Structure Theory (RST). We introduce a structure-aware Related Work Generation (RWG) pipeline where LLM agents are guided by an RST-derived structural plan to generate related work with iteratively improved structural relationships. Finally, we craft two structure-specific metrics, Width Profile Similarity (WPS) and Edge Coverage Ratio (ECR), to evaluate the coherence of the generated related work. Through extensive experiments, we demonstrate that our RST-centric generation method significantly enhances the structural and overall quality of RWG.

1 Introduction

Context shapes research. Related work is of paramount value for situating new research within the broader academic landscape. It provides context, highlights new contributions, and helps identify limitations in prior knowledge (Martin-Boyle et al., 2024; Li and Ouyang, 2025). Additionally, engaging with prior studies also convinces the readers that this new research builds upon a solid foundation, avoids redundancy, and contributes meaningfully to ongoing research findings. However, manually compiling related work is becoming increasingly difficult due to the rapid growth of scientific publications (Martin-Boyle et al., 2024; Li and Ouyang, 2025). Consequently, generating a related work section as an initial draft, commonly known as RWG (Hoang and Kan, 2010; Li and Ouyang, 2022, 2024), emerges as a significant and active research area within NLP.

Substantial research has been dedicated to RWG. Initial approaches often employed *extractive summarization techniques* (Hu and Wan, 2014; Wang et al., 2018; Chen and Zhuge, 2019), which identify and assemble key sentences from cited papers. With the advent of Transformer architectures (Vaswani et al., 2017), *abstractive methods* gained prominence (Chen et al., 2021, 2022; Liu et al., 2023a), enabling models to summarize and paraphrase content from source documents. More recently, the remarkable capabilities of LLMs (Zhao et al., 2025) have been increasingly applied to RWG, demonstrating potential for advancing abstractive related work generation (Martin-Boyle et al., 2024; Li and Ouyang, 2025; Zhang et al., 2025).

Considering that related work must clearly articulate the relationships among the target and cited papers, recent endeavors begin to pursue *structural relationships in RWG* but fall short in various aspects: Wang et al. (2024a) proposed a P-S-E-D framework emphasizing the establishment of connections between target and cited papers, but the P-S-E-D framework could be too narrow to capture all essential rhetorical structures in related work. One problem could be: it neglects the relations between the cited papers, which are necessary for RWG. Martin-Boyle et al. (2024) observed that straightforward prompting of LLMs often fails to adequately interconnect cited papers in a manner akin to human-written related work, but did not propose an effective method to solve it. Li and Ouyang (2025) highlighted that information about citation is beneficial and necessary; however, such information is often hard to derive. Zhang et al. (2025) employed knowledge graphs to identify conceptual relations for RWG, yet this approach only excels at concept-level relational structures. Whereas RWG requires paper-level relations.

This paper establishes the foundation for a structure-centric RWG. We introduce a structure-aware RWG pipeline with two quantitative structural assessment metrics. Inspired by RST (Mann and Thompson, 1988), a robust linguistic framework renowned for its ability to analyze text organization by identifying the functional relations between discourse segments, our pipeline employs an LLM guided by an RST-derived structural plan, aiming to generate related work that is not only informative but also exhibits strong adherence to a desired rhetorical structure. Furthermore, we craft two novel evaluation metrics specifically designed to assess the structural coherence of a related work by focusing on the logical flow and explicit relations between cited papers.

Our primary contributions are as follows:

- We develop a structure-aware RWG pipeline, called RETELL, that leverages the RST-derived relational structure to guide an LLM in iteratively writing the related work, promoting adherence to the desired organization during generation.
- We design two novel structure evaluation metrics, Width Profile Similarity (WPS) and Edge Coverage Ratio (ECR), specifically to assess the logical flow and the inter-paper relational structure in related work.
- Through extensive experiments, we demonstrate that our structure-aware generation method significantly improves the structural fidelity and overall quality of generated related work compared to baseline methods.

2 Preliminary & Related Works

This section first formally defines the task of RWG. It then introduces RST as the foundational linguistic framework for our approach. Finally, it reviews prior research in RWG, its evaluation, and the applications of RST in NLP, highlighting the gaps our work aims to address.

2.1 Problem Definition for RWG

Let T denote the target paper for which a related work needs to be generated. We assume access to relevant textual content $\boldsymbol{t}$ from the target paper, typically its abstract. Let $C = \{C_1, C_2, \ldots, C_N\}$ be a predefined set of N cited papers that are required to be discussed in the generated related work. For each cited paper $C_i \in C$, we assume access to its corresponding textual content $\boldsymbol{c_i}$. The objective of RWG is to learn a mapping function f such that, given the target paper's relevant textual content $\boldsymbol{t}$ and the set of cited papers' relevant texutal contents $\{\boldsymbol{c_i}\}_{i=1}^{N}$, it outputs a generated related work $\hat{\boldsymbol{r}} = f(\boldsymbol{t}, \{\boldsymbol{c_i}\}_{i=1}^{N})$. Again, the primary goal is to generate $\hat{\boldsymbol{r}}$ that not only accurately summarizes the content of the cited papers in the context of the target paper but also, crucially, reproduces the logical flow and relational structure observed in a human-written ground truth related work $\boldsymbol{r}$. Capturing and evaluating the logical flow and relational structure is the central focus of our work.

2.2 Rhetorical Structure Theory (RST)

Rhetorical Structure Theory, proposed by Mann and Thompson (1988), is a descriptive linguistic theory focused on text organization and coherence. It analyzes texts by identifying rhetorical relations between non-overlapping and contiguous text segments, known as Elementary Discourse Units (EDUs). According to RST, adjacent text spans, which can be individual EDUs or larger spans composed of multiple contiguous EDUs, are connected by rhetorical relations drawn from a predefined set. Most relations exhibit an asymmetry between a central span, termed the nucleus, and a supporting span, called the satellite. The nucleus is considered more essential to the writer's communicative purpose, while the satellite provides supplementary information that supports the nucleus.

Based on the RST, a text $\boldsymbol{r}$ can be analyzed and built into a hierarchical, tree-like structure, commonly referred to as Rhetorical structure tree (RS-tree), denoted $\mathcal{T}(\boldsymbol{r})$. From a bottom-up perspective on the RS-tree, the leaf nodes correspond to the individual EDUs of the text. Internal nodes represent larger text spans formed by the recursive application of rhetorical relations that connect adjacent sub-spans. The root node represents the entire text. This hierarchical representation explicitly models the discourse structure and the underlying logic of the text's organization.

2.3 Related Work

Related Work Generation. The automated generation of related work has evolved from extractive methods that select salient sentences (Hu and Wan, 2014; Wang et al., 2018; Chen and Zhuge, 2019), to abstractive approaches that synthesize novel text (Chen et al., 2021, 2022; Wang et al., 2022). This shift towards abstraction has naturally

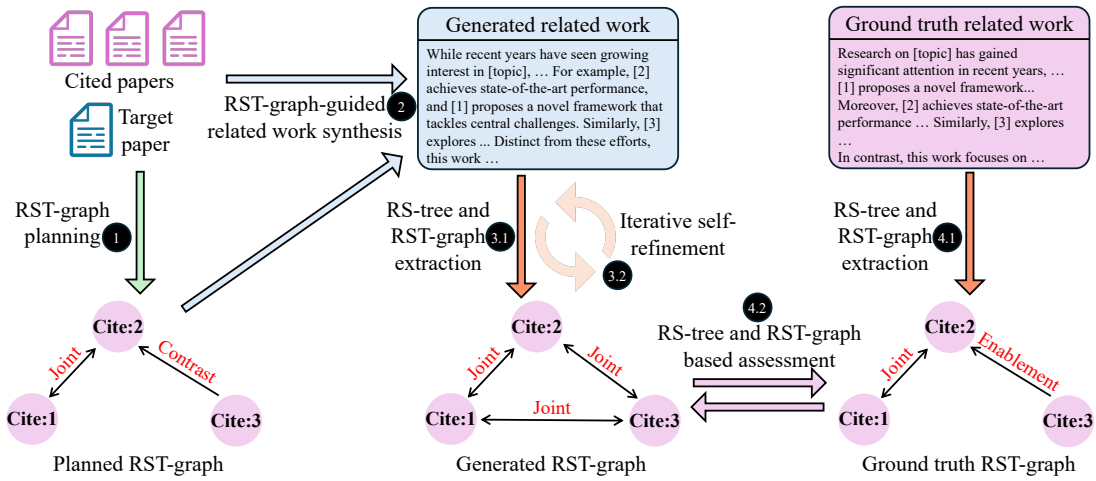

Figure 1: Illustration on RST-centric related work generation and assessment.

increased focus on the structural quality of generated text (Liu et al., 2023a; Wang et al., 2024a). With the advent of LLMs (Zhao et al., 2025), recent RWG research has further explored organizational aspects, for instance, through human-AI collaboration (Martin-Boyle et al., 2024), analyzing input information's impact on coherence (Li and Ouyang, 2025), or using knowledge graphs to guide generation (Zhang et al., 2025). Despite these advances, ensuring human-like coherence and logical flow remains a key challenge in RWG, which is comprehensively addressed by RETELL (See Figure 1).

Evaluation of RWG. A significant hurdle in advancing RWG is the evaluation of structural quality. Traditional n-gram metrics like ROUGE (Lin, 2004) primarily assess lexical overlap and are insufficient for structural evaluation. While semantic metrics (Zhao et al., 2019; Yuan et al., 2021) and LLM-based judgments (Liu et al., 2023b; Wang et al., 2024a) offer deeper content assessment, they do not provide a direct, comprehensive measure of logical flow and relational structure. Some studies explore specific structural checks like novelty statement detection (Nishimura et al., 2024) or citation grouping (Martin-Boyle et al., 2024; Nishimura et al., 2024), but these address limited facets. Relying on human evaluation for coherence (Li and Ouyang, 2025) is expensive and impractical at scale. This persistent gap in evaluation capabilities makes it difficult to reliably measure progress in generating structurally sound related work and to guide models effectively. Therefore, to directly address this limitation, we propose two novel graph-based evaluation metrics grounded in RST.

RST in NLP. Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) is a well-established linguistic framework for analyzing text organization and coherence by identifying functional relations between text segments. Its applications in NLP are diverse (Hou et al., 2020), notably in RST parsing for automatically deriving discourse structures (Sagae, 2009; Li et al., 2014; Liu et al., 2021) and in RST-guided text generation to produce more coherent and organized text (Adewoyin et al., 2022; Liu and Demberg, 2024; Kim et al., 2025). Beyond generation, RST has also been applied to other tasks such as distinguishing between human-written and machine-generated text by analyzing structural features (Kim et al., 2024). The proven ability of RST to model and enhance textual structure underpins its suitability for addressing the challenges in RWG.

3 RST-centric Related Work Generation and Assessment

3.1 Overview

Figure 1 illustrates our RST-centric design, i.e., RETELL. Initially, the LLM-based agent, given the target paper and a set of cited papers, constructs a planned RST-graph (1). This graph explicitly defines the intended rhetorical relationships between the cited papers and the target paper. Subsequently, the LLM agent uses this planned RST-graph as a structural blueprint to synthesize the initial draft of the related work (2). After that, the agent iteratively refines the related work until the predicted RST-graph converges or a certain iteration threshold is met (3.1 and 3.2). Finally, we extract the RS-tree and RST-graph for the ground truth related work (4.1). We compare the RS-tree and

RST-graph of our predicted related work and the ground truth related work to quantitatively assess their structural similarity (4.2).

In the following subsections, we first detail our LLM-based workflow for RST-based RWG (Section 3.2). We then explain the process of RST-graph extraction (Section 3.3), which is integral to both the generation's refinement loop and the final evaluation. Finally, we introduce our novel RST-based evaluation metrics (Section 3.4).

3.2 RST-based Related Work Generation

This section introduces our LLM-powered agent, which aims to emulate a structured human writing process: first understanding inter-paper relations, leveraging this understanding to construct a well-organized related work, and then keep refining it by re-analyzing the structure. The generation workflow comprises three key phases:

RST-graph planning 1. The input for this phase consists of information from the target paper and the set of cited papers. Given the input, the LLM is prompted to determine the most salient rhetorical relationships between the target and cited papers. To ensure global structural coherence and avoid the potential loss of context, we let the LLM generate the overall graph structure with all papers provided in the context, thanks to long context support in recent LLM models (Beltagy et al., 2020; Su et al., 2024; Wang et al., 2024b).

We constrain the LLM to utilize a predefined set of ten key rhetorical relations (the specific choices and their definitions are detailed in Appendix A), where the definitions are provided in the system prompt. Furthermore, the LLM is guided to ensure the resulting graph is a directed acyclic graph (DAG) (Yao et al., 2024; Ma, 2025), thereby avoiding cycles, and to maintain overall coherence. Finally, we get the output of this phase: a DAG where nodes represent the cited papers, and the directed, labeled edges signify the intended rhetorical relations.

RST-graph-guided related work synthesis 2. This phase begins with outline generation, where, before drafting the full text, the LLM formulates a high-level outline based on the planned RST-graph. This outline specifies the overall theme, the sequence for introducing cited papers, strategies for transitioning between them, explicitly reflecting the relationships encoded in the planned RST-graph, and the bridges between the cited papers and the target paper. This step effectively translates the structured graph into planned text discourse. Following the outline generation, the LLM generates an initial draft of the related work, denoted as $\hat{\boldsymbol{r}}_0$.

Iterative self-refinement (3.1, 3.2). This phase aims to iteratively refine the related work by making the LLM re-analyze the inter-paper relations. This phase is necessary as it is hard for LLM to completely understand the rhetorical relations and even harder for it to follow the planned RST graph to write a related work. However, it is well-known that iteratively asking LLM to refine the output would lead to better generated results, see (Madaan et al., 2023; Chen et al., 2024). Below, we discuss our design:

Once an initial draft $\hat{\boldsymbol{r}}_0$ is generated, RETELL extracts its corresponding RST-graph $G_{\hat{\boldsymbol{r}}_0}$, following Section 3.3. Subsequently, the LLM is prompted to analyze each relation within this parsed $G_{\hat{\boldsymbol{r}}_0}$ by referencing back to the contents of the cited and target papers. Particularly, for each edge in $G_{\hat{\boldsymbol{r}}_0}$, the LLM agent verifies its correctness by checking back to the content of the paper. Changing the perspective from the target and cited papers to the newly generated related work and its RST-graph provides a new context for the LLM to reflect and improve the structure. Following this analysis, the LLM generates new action items for improving the related work. Based on the action items, the LLM refines the draft, producing a new version $\hat{\boldsymbol{r}}_1$. Subsequently, an RST-graph $G_{\hat{\boldsymbol{r}}_1}$ is extracted from this new draft $\hat{\boldsymbol{r}}_1$.

The core of the refinement cycle involves comparing the newly extracted graph $G_{\hat{\boldsymbol{r}}_1}$ with the graph from the previous iteration $G_{\hat{\boldsymbol{r}}_0}$. If mismatches are detected between these two graphs, the LLM is prompted to focus on these differing relations and generate new action items for further refinement. This iterative process of analysis, action item generation, and textual revision continues. The cycle terminates when the extracted RST-graphs from two consecutive iterations, $G_{\hat{\boldsymbol{r}}_i}$ and $G_{\hat{\boldsymbol{r}}_{i+1}}$, are identical, or when i reaches a preset maximum number of iterations, where i is the number of iterations. Finally, we get the related work, denoted as $\hat{\boldsymbol{r}}$.

3.3 RS-tree and RST-graph Extraction

This section introduces how to extract the RST-graph $G(\boldsymbol{r})$ from a given related work text $\boldsymbol{r}$, a critical component utilized in steps 3.1 and 4.1. Figure 2 explains how to build an RS-tree for the related work example (on the top). Subsequently,

a. Related Work Example: [Recently, VDSH @Cite:1 proposed to use a VAE]^{1} [to learn the latent representations of documents]^{2} [and then use a separate stage]^{3} [to cast the continuous representations into binary codes.]^{4} [While fairly successful,]^{5} [this generative hashing model requires a two-stage training.]^{6} [NASH @Cite:2 proposed to substitute the Gaussian prior in VDSH with a Bernoulli]^{7} [prior to tackle this problem,]^{8} [by using a straight-through estimator @Cite:3]^{9} [to estimate the gradient of neural network]^{10} [involving the binary variables.]^{11} [This model can be trained in an end-to-end manner.]^{12} [Our models differ from VDSH and NASH]^{13} [in that mixture priors are employed]^{14} [to yield better hashing codes,]^{15} [whereas only the simplest priors are used in both VDSH and NASH.]^{16}

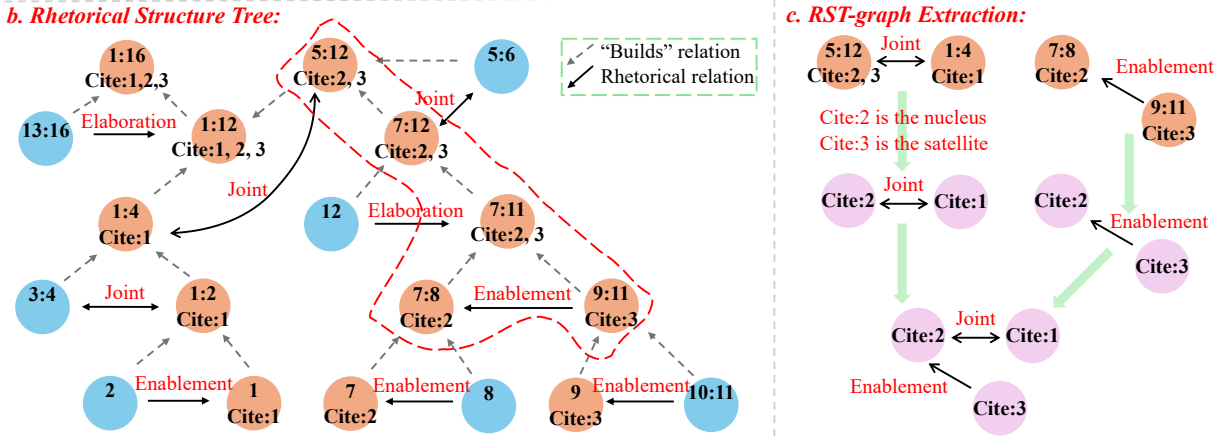

Figure 2: Illustration on RS-tree parsing and RST-graph extraction (nodes building blue nodes are omitted).

we can extract an RST-graph from the RS-tree. Of note, this related work example is an example from the **Multi-XScience** dataset (Lu et al., 2020).

It begins with RST parsing, where the input text $\boldsymbol{r}$, i.e., related work example, is processed by an RST parser. As shown in Figure 2a, we use "[", "]" pair to split this related work into 16 EDUs. We put the EDU number as the superscript of "[]". Figure 2b depicts the corresponding RS-tree. Each circle indicates a text span, where we use the number pair $i:j$ to indicate the text span ranges from i-th EDU to j-th EDU of $\boldsymbol{r}$. For example in the bottom left of Figure 2b, the span node(*1:2*) is formed by an *Enablement* relation, where node(*1*) is the nucleus and node(*2*) is the satellite.

Following parsing, the next step involves identifying cited papers within the RS-Tree. Each node in the RS-tree is analyzed to identify any cited papers mentioned within its textual content, which can be done using regular expressions. In Figure 2b, text spans containing citations are color-coded as orange nodes, blue nodes otherwise.

The final step is to extract the RST-graph $G(\boldsymbol{r})$ from the RS-tree $\mathcal{T}(\boldsymbol{r})$, as shown in Figure 2c. In an RST-graph, each node represents a unique cited paper. Firstly, edges that connect nodes both containing citations in the RS-tree with rhetorical relationships are identified. If each node only contains one citation, we link the two paper-nodes with the rhetorical relation on the edge. For example, we connect the paper node(*Cite:2*) with (*Cite:3*) with the *Enablement* relation between span node(*7:8*) and (*9:11*).

If a selected text span node contains more than one citation, we will perform two attempts: (i) We will identify the main citation across these citations, guided by the nucleus/satellite distinction within the RS-tree. For example, the span node (*5:12*) has two citations, i.e., Cite 2, 3, where we need to identify which citation of the span node (*5:12*) is the core for this *Joint* relationship. We achieve this via traversing the RS-tree as follows: Within the RS-tree of Figure 2b, the text span node (*7:8*, *Cite:2*) is identified as a nucleus relative to the text span node (*9:11*, *Cite:3*). Therefore, Cite:2 is the core citation when compared to Cite:3. This indicates that Cite:2 results in a *Joint* edge between paper-node (*Cite:3*) and paper-node (*Cite:1*) in the RST-graph. We thus only keep this *Joint* edge. (ii) Chances are that multiple distinct citations in a single text span might not have any relationship in (i). In this case, we simply regard all these citations in the single text span as the main citations, resulting in multiple rhetorical relationships in the RST-graph. Further, we will also establish a *Joint* relationship between these citations.

3.4 RST-based Assessment

To evaluate the structural quality of a generated related $\hat{\boldsymbol{r}}$, we compare its RS-tree $\mathcal{T}(\hat{\boldsymbol{r}})$ and RST-graph $G(\hat{\boldsymbol{r}})$ against those of the ground truth related work $\boldsymbol{r}$ ($\mathcal{T}(\boldsymbol{r})$ and $G(\boldsymbol{r})$). We propose two

graph-based metrics (4.2) for this purpose:

Width Profile Similarity (WPS) This metric assesses the similarity in the overall hierarchical shape of the RS-trees $\mathcal{T}(\boldsymbol{r})$ and $\mathcal{T}(\hat{\boldsymbol{r}})$. Firstly, each RS-tree is converted into a "width profile" vector $P = [w_0, w_1, \ldots, w_k]$, where w_j represents the number of nodes at depth j in the tree. For WPS, we consider the tree formed by only "builds" relations. The shape of this tree can be indicative of the discourse strategy; for instance, a wider tree structure might suggest a breadth-first discussion covering multiple points with similar emphasis, whereas a deeper, narrower tree could imply a focus on a primary argument with extensive elaboration. The WPS is then calculated as the cosine similarity between the width profile vectors of the ground truth and generated RS-trees:

$$\text{WPS}(\boldsymbol{r}, \hat{\boldsymbol{r}}) = \cos(P(\mathcal{T}(\boldsymbol{r})), P(\mathcal{T}(\hat{\boldsymbol{r}})) \quad (1)$$

A higher WPS score indicates that the generated related work exhibits a hierarchical organization, in terms of layer-wise node distribution, that is more similar to that of the ground truth.

Edge Coverage Ratio (ECR) While WPS captures the general tree shape, ECR focuses on the specific relational connections between cited papers as represented in their respective RST-graphs $G(\boldsymbol{r})$ and $G(\hat{\boldsymbol{r}})$. This metric measures the proportion of correctly identified edges from the ground truth graph that are present in the generated graph. ECR is calculated as:

$$\text{ECR}(\boldsymbol{r}, \hat{\boldsymbol{r}}) = \frac{|E \cap \hat{E}|}{|E|} \quad (2)$$

where E denotes the set of edges in the ground truth RST-graph $G(\boldsymbol{r})$, and $\hat{E}$ represents the set of edges in the generated RST-graph $G(\hat{\boldsymbol{r}})$.

We opt for ECR over the Jaccard index based on the natural assumption that all the rhetorical relationships in the ground truth are correct (i.e., rigorously verified by the expert writer). However, human-crafted related work might be subject to missing some rhetorical relationships that were captured by RETELL.

ECR assesses how well the generated RWS covers the essential relationships present in the ground truth. ECR directly measures recall of these ground truth edges, which we deem more critical for structural fidelity in this context than penalizing relationships that were absent in ground truth but captured by automatically generated related work. Of note, the Jaccard index would emphasize that penalization in the denominator.

4 Experiments

4.1 Dataset

Following common practice in related work generation research (Chen et al., 2022; Liu et al., 2023a; Zhang et al., 2025), we evaluate our proposed method on three publicly available datasets: **Multi-XScience**, **TAS2**, and **TAD**. **Multi-XScience** (Lu et al., 2020) is constructed by integrating data from arXiv (Ginsparg, 1991) and the Microsoft Academic Graph (MAG) (Sinha et al., 2015). **TAS2** (Chen et al., 2021) is derived from the S2ORC dataset (Lo et al., 2020), encompassing multiple scientific domains such as physics and mathematics. **TAD** (Chen et al., 2021) consists of related work sections from computer science articles, sourced from the Delve dataset (Akujuobi and Zhang, 2017). For evaluation, we randomly selected 500 samples of related work from the test splits of each dataset. Each selected sample included at least four cited papers. In all datasets, the input consists of the abstracts from the cited papers, while the ground truth is the related work section from the target paper. The **Multi-XScience** dataset additionally provides the abstract of the target paper. Detailed statistics for these datasets are available in Appendix B.

4.2 Settings

We compare our RST-based approach against three alternative LLM-based methods for related work generation:

- **Group-based** (Martin-Boyle et al., 2024): This method first employs an LLM to organize citations into coherent groups based on their topical similarity and relevance to the target paper. The LLM then generates the related work using these pre-defined groups.

- **Feature-based** (Li and Ouyang, 2025): This method involves prompting an LLM to create a faceted summary (including object, method, findings, contribution, and keywords) for each cited paper. It also generates a main idea for the target related work based on these summaries. All this information is subsequently fed to the LLM to produce the related work.

Model	Method	Mutli-XScience			TAS2			TAD		
		ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
Llama	Group-based	29.60	5.21	14.40	24.45	3.31	13.64	27.51	3.70	13.93
	Feature-based	27.65	5.32	13.25	23.50	3.66	12.71	29.37	4.32	13.99
	MiniGraph	26.58	5.47	12.63	23.09	3.74	12.24	29.64	**4.54**	13.76
	RETELL(Ours)	**32.18**	**5.63**	**14.92**	**26.43**	**3.75**	**13.94**	**30.18**	4.42	**14.48**
Qwen	Group-based	29.50	4.76	14.64	25.21	3.43	13.84	28.29	3.62	14.14
	Feature-based	27.43	4.94	12.87	23.55	3.49	12.41	29.13	4.10	13.68
	MiniGraph	26.28	4.88	12.09	23.42	3.64	11.97	29.18	**4.16**	13.14
	RETELL(Ours)	**31.79**	**5.04**	**14.67**	**26.73**	**3.65**	**14.01**	**29.82**	4.11	**14.29**

Table 1: Performance with respect to the ROUGE metrics.

- **MiniGraph** (Zhang et al., 2025): This method first constructs a knowledge graph, called minigraph, from entities and relations extracted from the text of a subset of cited papers. The LLM is then prompted to generate a summary for each minigraph. These chunk summaries are finally combined to form the related work.

For the RST parsing component, we used the parser developed by Liu et al. (2021). The self-refinement loop in our method was constrained to a maximum of 5 iterations. Our experiments were conducted using two LLM backbones: **Llama-3.1-8B-Instruct** (Llama Team, 2024) and **Qwen2.5-14B-Instruct** (Qwen Team, 2024). To evaluate structural quality, we employed our proposed RST-graph-based metrics, **WPS** and **ECR**, as defined in Section 3.4. Additionally, we utilized standard ROUGE F1 scores (**ROUGE-1**, **ROUGE-2**, and **ROUGE-L**) (Lin, 2004) to assess generation quality. All experiments were performed in a single run, with results averaged across the entire dataset.

4.3 Structural Quality

Model	Method	Mutli-XScience		TAS2		TAD	
		WPS	ECR	WPS	ECR	WPS	ECR
Llama	Group-based	62.48	29.64	70.06	27.50	72.65	29.24
	Feature-based	54.08	32.56	59.59	33.91	64.96	41.11
	MiniGraph	60.50	31.49	68.29	33.20	79.76	37.39
	RETELL(Ours)	**75.40**	**39.92**	**75.75**	**34.71**	**82.30**	**41.77**
Qwen	Group-based	76.59	16.32	76.13	12.96	78.53	13.52
	Feature-based	67.53	21.23	65.96	34.23	77.93	34.93
	MiniGraph	60.42	36.22	66.55	29.68	78.92	43.41
	RETELL(Ours)	**77.70**	**44.43**	**76.94**	**39.98**	**81.92**	**44.01**

Table 2: Structure quality.

Table 2 displays the results of the structural quality assessment using our proposed metrics. Our method consistently achieves the highest scores for both WPS and ECR across all models. This indicates that our approach generates related work that has a similar overall structure to the ground truth related work, and also captures the relationships among cited papers well (a specific example is provided in Appendix C). Regarding the other methods, their performance for the second-best WPS scores varies across different datasets and models. Generally, group-based methods perform the poorest on the ECR metric, suggesting that merely grouping cited papers is insufficient for uncovering the intricate relations among them. The Feature-based and MiniGraph methods show better ECR performance, implying that information about the target paper's main idea and conceptual relationships can, to some extent, aid in establishing connections between cited papers.

4.4 Overall Performance

Table 1 showcases the overall performance of the methods based on the standard ROUGE-1, ROUGE-2, and ROUGE-L metrics. For example, on the Multi-XScience dataset, when averaging results from the Llama and Qwen models, our method demonstrates improvements over the second-best performing baseline by 2.43 points in ROUGE-1, 0.13 in ROUGE-2, and 0.27 in ROUGE-L. Similarly, for the TAS2 dataset, our approach yields average gains of 1.75 in ROUGE-1, 0.01 in ROUGE-2, and 0.23 in ROUGE-L. On the TAD dataset, our method shows improvements of 0.59 in ROUGE-1 and 0.31 in ROUGE-L, although a slight decrease is observed in the ROUGE-2 score compared to the top baseline.

This strong performance suggests that the enhancement in the structure positively influences the overall generation quality. By guiding the LLM to produce better-organized content that accurately reflects inter-paper relationships, our method likely facilitates more focused and relevant text generation for each segment of the related work.

4.5 Ablation Study

To assess the contribution of the iterative self-refinement phase in our method, we performed an ablation study. We compared our complete generation pipeline (referred to as "Ours") with a variant that omits the self-refinement module (referred to

as "Ours-Ref."). In this ablated version, the generation process concludes after the initial synthesis phase, treating the draft related work as the final output, without subsequent revisions. This comparison was conducted using the Llama backbone on the Multi-XScience, TAS2, and TAD datasets.

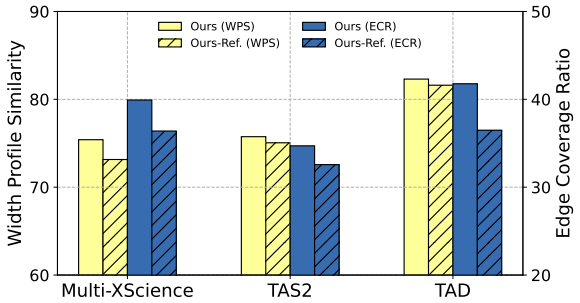

Figure 3: Ablation study of the iterative self-refinement phase on structural quality.

First, we examine the impact of the iterative self-refinement phase on structural quality, as shown in Figure 3. The results clearly demonstrate the value of self-refinement. For both WPS and ECR metrics, which assess the overall hierarchical structure and relation among the cited papers, the full method significantly outperforms the ablated version across all three datasets. This indicates that the refinement loop is effective in correcting structural errors and aligning the generated text more closely with the planned rhetorical structure.

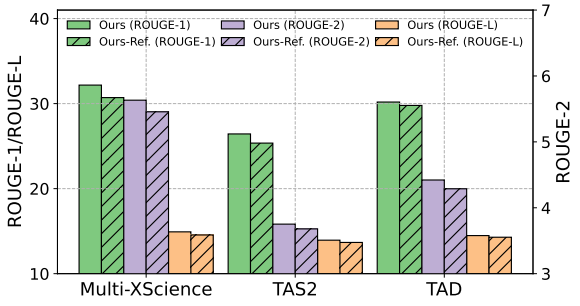

Figure 4: Ablation study of the iterative self-refinement phase on generation quality.

Next, Figure 4 illustrates the impact of the interactive self-refinement module on overall performance as measured by ROUGE scores. Consistent with the improvements in structural quality, the full method generally achieves slightly higher ROUGE-1, ROUGE-2, and ROUGE-L scores compared to the version without refinement across all three datasets. This again suggests that the enhanced structural integrity due to the refinement process can positively contribute to the content relevance captured by ROUGE metrics.

4.6 Impact of Number of Cited Papers

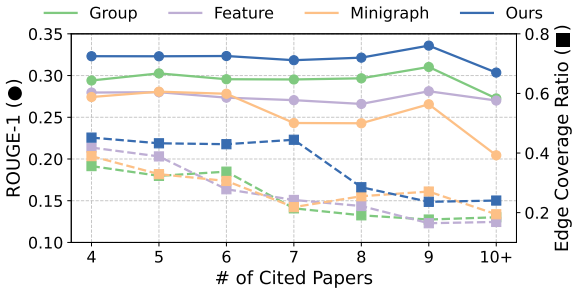

Figure 5: Study on the impact of number of cited paper.

We investigate the impact of varying the number of cited papers on both structural quality and generation performance. We conduct this sensitivity analysis on the Multi-XScience dataset with the Llama model as the LLM backbone. Our findings are illustrated in Figure 5. The results indicate that the ROUGE-1 score remains largely stable across all methods, irrespective of the number of cited papers. In contrast, the Edge Coverage Ratio demonstrates significant sensitivity to this variable. Notably, as the number of cited papers increases, the ECR for baseline methods tends to decline rapidly. Our proposed RST-based method, however, maintains a more consistent ECR in these scenarios, indicating a better preservation of structural relations. Despite this relative robustness, a noticeable performance drop in ECR is observed for all approaches, including ours, when the number of cited papers reaches eight or more. This suggests that generating a structurally coherent related work section becomes substantially more challenging as the density of inter-paper relations increases, potentially indicating an inherent difficulty in the task at higher citation counts.

5 Conclusion

In this paper, we aim to address the critical challenge of generating structurally coherent related work and the inadequacy of existing metrics. Inspired by RST, we introduce RETELL, a novel LLM-based pipeline for generating structure-aware related work. Besides, we introduce two new graph-based metrics for assessing the logical flow and inter-paper relational structure. Our experiments demonstrate our method significantly enhances the structural fidelity and overall quality of the output.

6 Limitations

Due to resource constraints, our evaluations only explored the open-source LLM backbones and did not include a comprehensive comparison against the latest large-scale commercial models. Consequently, our reported results primarily serve to demonstrate the relative efficacy of our approach. Moreover, our experiments were conducted on three standard datasets where, consistent with common practice in many existing works, only abstracts were used as the textual input for cited papers. We would like to explore full-text documents that contain richer contextual and relational information.

7 Ethical Considerations

The development of automated academic writing tools, such as our proposed related work generation system, brings some ethical concerns. Key risks include misuse, such as plagiarism through unmodified use of the generated text, and the potential for factual inaccuracies or biases inherent in LLMs. Although our RST-guided approach enhances coherence, it does not ensure factual accuracy. We emphasize the necessity of responsible use, rigorous content verification, and strict adherence to academic integrity to mitigate these risks.

References

Rilwan Adewoyin, Ritabrata Dutta, and Yulan He. 2022. RSTGen: Imbuing fine-grained interpretable control into long-FormText generators. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1822–1835, Seattle, United States. Association for Computational Linguistics.

Uchenna Akujuobi and Xiangliang Zhang. 2017. Delve: A dataset-driven scholarly search and analysis system. *SIGKDD Explor. Newsl.*, 19(2):36–46.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *Preprint*, arXiv:2004.05150.

Lynn Carlson and Daniel Marcu. 2001. Discourse tagging reference manual. Technical Report ISI-TR-545, University of Southern California Information Sciences Institute.

Jingqiang Chen and Hai Zhuge. 2019. Automatic generation of related work through summarizing citations. *Concurrency and Computation: Practice and Experience*, 31(3):e4261. E4261 CPE-16-0462.R2.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2024. Iterative translation refinement with large language models. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 181–190, Sheffield, UK. European Association for Machine Translation (EAMT).

Xiuying Chen, Hind Alamro, Mingzhe Li, Shen Gao, Rui Yan, Xin Gao, and Xiangliang Zhang. 2022. Target-aware abstractive related work generation with contrastive learning. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 373–383, New York, NY, USA. Association for Computing Machinery.

Xiuying Chen, Hind Alamro, Mingzhe Li, Shen Gao, Xiangliang Zhang, Dongyan Zhao, and Rui Yan. 2021. Capturing relations between scientific papers: An abstractive model for related work section generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6068–6077, Online. Association for Computational Linguistics.

Paul Ginsparg. 1991. arxiv.org e-print archive. `https://arxiv.org`. Accessed: 2025-04-30.

Cong Duy Vu Hoang and Min-Yen Kan. 2010. Towards automated related work summarization. In *Coling 2010: Posters*, pages 427–435, Beijing, China. Coling 2010 Organizing Committee.

Shengluan Hou, Shuhan Zhang, and Chaoqun Fei. 2020. Rhetorical structure theory: A comprehensive review of theory, parsing methods and applications. *Expert Systems with Applications*, 157:113421.

Yue Hu and Xiaojun Wan. 2014. Automatic generation of related work sections in scientific papers: An optimization approach. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1624–1633, Doha, Qatar. Association for Computational Linguistics.

Zae Myung Kim, Kwang Lee, Preston Zhu, Vipul Raheja, and Dongyeop Kang. 2024. Threads of subtlety: Detecting machine-generated texts through discourse motifs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5449–5474, Bangkok, Thailand. Association for Computational Linguistics.

Zae Myung Kim, Anand Ramachandran, Farideh Tavazoee, Joo-Kyung Kim, Oleg Rokhlenko, and Dongyeop Kang. 2025. Align to structure: Aligning large language models with structural information. *Preprint*, arXiv:2504.03622.

Sujian Li, Liang Wang, Ziqiang Cao, and Wenjie Li. 2014. Text-level discourse dependency parsing. In

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 25–35, Baltimore, Maryland. Association for Computational Linguistics.

Xiangci Li and Jessica Ouyang. 2022. Automatic related work generation: A meta study. *Preprint*, arXiv:2201.01880.

Xiangci Li and Jessica Ouyang. 2024. Related work and citation text generation: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13846–13864, Miami, Florida, USA. Association for Computational Linguistics.

Xiangci Li and Jessica Ouyang. 2025. Explaining relationships among research papers. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1080–1105, Abu Dhabi, UAE. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Dongqi Liu and Vera Demberg. 2024. RST-LoRA: A discourse-aware low-rank adaptation for long document abstractive summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2200–2220, Mexico City, Mexico. Association for Computational Linguistics.

Jiachang Liu, Qi Zhang, Chongyang Shi, Usman Naseem, Shoujin Wang, Liang Hu, and Ivor Tsang. 2023a. Causal intervention for abstractive related work generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2148–2159, Singapore. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021. DMRST: A joint framework for document-level multilingual RST discourse segmentation and parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 154–164, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

AI @ Meta Llama Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-XScience: A large-scale dataset for extreme multi-document summarization of scientific articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8068–8074, Online. Association for Computational Linguistics.

Jing Ma. 2025. Causal inference with large language model: A survey. *Preprint*, arXiv:2409.09822.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Anna Martin-Boyle, Aahan Tyagi, Marti A. Hearst, and Dongyeop Kang. 2024. Shallow synthesis of knowledge in gpt-generated texts: A case study in automatic related work composition. *Preprint*, arXiv:2402.12255.

Kazuya Nishimura, Kuniaki Saito, Tosho Hirasawa, and Yoshitaka Ushiku. 2024. Toward structured related work generation with novelty statements. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 38–57, Bangkok, Thailand. Association for Computational Linguistics.

Alibaba Cloud Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Kenji Sagae. 2009. Analysis of discourse structure with syntactic dependencies and data-driven shift-reduce parsing. In *Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09)*, pages 81–84, Paris, France. Association for Computational Linguistics.

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, page 243–246, New York, NY, USA. Association for Computing Machinery.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomput.*, 568(C).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Pancheng Wang, Shasha Li, Kunyuan Pang, Liangliang He, Dong Li, Jintao Tang, and Ting Wang. 2022. Multi-document scientific summarization from a knowledge graph-centric view. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6222–6233, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Pancheng Wang, Shasha Li, Jintao Tang, and Ting Wang. 2024a. What can rhetoric bring us? incorporating rhetorical structure into neural related work generation. *Expert Systems with Applications*, 251:123781.

Xindi Wang, Mahsa Salmani, Parsa Omidi, Xiangyu Ren, Mehdi Rezagholizadeh, and Armaghan Eshaghi. 2024b. Beyond the limits: a survey of techniques to extend the context length in large language models. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24.

Yongzhen Wang, Xiaozhong Liu, and Zheng Gao. 2018. Neural related work summarization with a joint context-driven attention mechanism. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1776–1786, Brussels, Belgium. Association for Computational Linguistics.

Yang Yao, Xin Wang, Zeyang Zhang, Yijian Qin, Ziwei Zhang, Xu Chu, Yuekui Yang, Wenwu Zhu, and Hong Mei. 2024. Exploring the potential of large language models in graph generation. *Preprint*, arXiv:2403.14358.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Zhi Zhang, Yan Liu, Sheng hua Zhong, Gong Chen, Yu Yang, and Jiannong Cao. 2025. Mixture of knowledge minigraph agents for literature review generation. *Preprint*, arXiv:2411.06159.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2025. A survey of large language models. *Preprint*, arXiv:2303.18223.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

A Rhetorical Relation Set

We adopt the design of eighteen relations proposed by Carlson and Marcu (2001). To better align with the task of related work generation, we selected ten relations and defined them as follows in the system prompts:

- **Joint:** A Joint relation between two papers means the two papers exhibit some sort of parallel structure between two papers, but are not in contrast.
- **Elaboration:** An Elaboration relation from one paper to another means the first paper adds detail or explanation to the main topic of the second paper.
- **Explanation:** An Explanation relation from one paper to another means the first paper provides evidence or justification for the situation presented in the second paper.
- **Contrast:** A Contrast relation between two papers means the two papers are in contrast with each other along some dimension.
- **Temporal:** A Temporal relation from one paper to another means the situation presented in the first paper occurs before or leads up to the situation in the second paper.
- **Background:** A Background relation from one paper to another means the first paper establishes the context or grounds with respect to which the second paper is to be interpreted.
- **Manner-Means:** A Manner-Means relation from one paper to another means the first paper explains the way in which something in the second paper is done.
- **Enablement:** An Enablement relation from one paper to another means the first paper enables the second paper to happen.
- **Evaluation:** An Evaluation relation from one paper to another means the first paper assesses the second paper and recognizes the value it assigns.

Dataset	# of Samples	Textual Content of Target Paper	Textual Content of Cited Paper	Avg. # of Cited Papers	Avg. # of Words in Cited Papers	Avg. # of Relations
Multi-XScience	500	Abstract	Abstract	6.05	156.27	11.99
TAS2	500	None	Abstract	4.45	223.27	4.91
TAD	500	None	Abstract	4.19	176.29	4.41

Table 3: Dataset statistics.

- **Cause:** A Cause relation from one paper to another means the first paper is the cause and the second paper is the effect.

B Dataset Statistics

The statistics of the evaluated dataset are presented in Table 3.

C Generated Related Work Example

We present an example from the Multi-XScience dataset to demonstrate the different outputs generated by different methods, as shown in the grey boxes on the following pages.

Ground Truth:
Powered by large-scale datasets of 3D CAD models (e.g., ShapeNet [1]), deep-learning-based methods have been proposed for 3D reconstruction. Both 3D-R2N2 [2] and LSM [4] use RNNs to infer 3D shape from single or multiple input images and achieve impressive results. However, RNNs are time-consuming and permutation-variant, which produce inconsistent reconstruction results. 3DensiNet [3] uses max pooling to aggregate the features from multiple images. However, max pooling only extracts maximum values from features, which may ignore other valuable features that are useful for 3D reconstruction.

Ours (ROUGE-1: 30.89, ECR: 1.0):
Previous research has established the foundational concept of using deep learning for 3D shape representation [1], which has been further developed in subsequent works [2] and [4] . These studies explore the use of deep learning for 3D reconstruction, with [2] employing recurrent neural networks (RNNs) for generating 3D reconstructions from multiple images, while [4] leverages geometric constraints for more efficient and accurate 3D reconstruction using fewer input images. In contrast, [3] proposes an alternative approach using density heat-maps for 2D-to-3D transformation, addressing the limitations of RNN-based methods by introducing a novel architecture named 3DensiNet. This approach not only enhances the quality of the generated 3D volumes but also integrates adversarial learning for improved accuracy. Our work, Pix2Vox, builds on these advancements by introducing a context-aware fusion module to enhance the consistency and quality of multi-view 3D reconstructions, thereby outperforming existing methods in terms of both accuracy and efficiency.

Group-based (ROUGE-1: 25.20, ECR: 0.167):
This paper delves into the use of recurrent neural networks (RNNs) for 3D reconstruction from multi-view images, a topic that has been explored in several recent works. Notably, [2] discusses the inherent limitations of RNN-based approaches in this context. The paper highlights that when given the same set of input images in different orders, RNNs are unable to produce consistent reconstruction results, a critical issue for multi-view 3D reconstruction. Additionally, due to long-term memory loss, RNNs cannot fully leverage input images to refine reconstruction outcomes, thus limiting their effectiveness in capturing detailed 3D representations [2].
Another relevant study [4] addresses 3D reconstruction through a different lens, emphasizing the importance of leveraging 3D geometry via feature projection and unprojection along viewing rays. Although this approach does not directly use RNNs, it underscores the challenges of handling multi-view images for 3D reconstruction, providing a broader context for the limitations and potential solutions beyond the scope of RNN-based methods.

Feature-based (ROUGE-1: 22.59, ECR: 0.167):
Previous research efforts in the domain of 3D reconstruction from RGB images have primarily focused on leveraging deep learning techniques to enhance the representation and accuracy of reconstructed 3D models. For instance, [1] introduced 3D ShapeNets, a Convolutional Deep Belief Network (CDBN) designed to represent geometric 3D shapes as probability distributions on a 3D voxel grid. This method not only facilitates the recovery of full 3D shapes from 2.5D depth maps but also supports joint object recognition and shape completion, significantly improving performance over existing methods. However, 3D ShapeNets rely on synthetic CAD data and do not directly address the challenges of converting RGB images to 3D volumes.
In contrast, [2] proposed the 3D Recurrent Reconstruction Neural Network (3D-R2N2), which leverages recurrent neural networks (RNNs) to map images to 3D occupancy grids. This approach is particularly effective in handling single-view and multi-view reconstructions without requiring annotations or object class labels. Despite its success, RNN-based methods like 3D-R2N2 suffer from limitations such as inconsistency in reconstruction results when input images are presented in different orders and the inability to fully utilize all input images due to long-term memory loss.
To overcome these limitations, recent works have explored alternative architectures that do not rely on RNNs. For example, [3] introduced 3DensiNet, a neural network that predicts 3D volumetric objects from 2D images using density heat-maps as an intermediate step. This method has been shown to outperform classical 3D autoencoders and other state-of-the-art techniques. Similarly, [4] presents a multi-view stereopsis system that combines learning with geometric constraints to achieve efficient and complete 3D reconstruction. However, these methods do not address the inconsistencies and inefficiencies associated with sequential processing in RNN frameworks.
In this paper, we introduce Pix2Vox, a novel framework that enhances the consistency and efficiency of 3D reconstruction from RGB images. Unlike previous methods, Pix2Vox utilizes an encoder-decoder architecture to generate a coarse 3D volume from each input image and a context-aware fusion module to select high-quality reconstructions. Additionally, a refiner is employed to further refine the fused 3D volume, leading to superior performance and speed compared to existing approaches. Experimental results on the ShapeNet and Pix3D benchmarks demonstrate the effectiveness of Pix2Vox in terms of reconstruction accuracy and generalization to unseen categories.

MiniGraph (ROUGE-1: 24.89, ECR: 0.5):
Recent advancements in 3D shape representation and reconstruction have spurred significant progress in visual understanding [1]. Specifically, the introduction of 3D ShapeNets, which utilize Convolutional Deep Belief Networks to model 3D shapes as probability distributions on voxel grids, has enabled joint object recognition and shape completion from 2.5D depth maps [1]. However, the reliance on 2.5D depth maps limits the applicability of such methods to scenarios with RGB data only. To address this, [2] proposed the 3D Recurrent Reconstruction Neural Network (3D-R2N2), a recurrent neural network capable of reconstructing 3D shapes from multi-view RGB images without the need for annotations or object class labels. Although 3D-R2N2 demonstrates superior performance in single-view reconstruction, it suffers from inherent limitations such as inconsistency in reconstruction results due to the order of input images and long-term memory loss in RNNs, which hampers the effective refinement of reconstructions [2]. In parallel, [3] introduced 3DensiNet, a novel architecture that leverages density heat-maps for 2D-to-3D transformations, showing promising results in generating 3D volumetric objects from single 2D images. Despite these advances, the aforementioned methods either rely on specific input modalities or face challenges in refining and fusing information from multiple views effectively. Recent advancements in 3D reconstruction from multi-view and single-view RGB images have been predominantly driven by deep learning techniques. For instance, [4] introduces a learnt system for multi-view stereopsis that leverages feature projection and unprojection along viewing rays, enabling end-to-end learning for metric 3D reconstruction. This system integrates shape priors and geometric constraints, outperforming both classical approaches and recent learning-based methods on the ShapeNet dataset. However, existing works like 3D-R2N2, which rely on recurrent neural networks (RNNs), suffer from order dependency and long-term memory loss, making it challenging to produce consistent and refined reconstructions from multiple inputs. Our proposed Pix2Vox framework addresses these limitations by employing an encoder-decoder architecture for coarse 3D volume generation and a context-aware fusion module to adaptively select high-quality reconstructions, thereby ensuring superior performance and consistency across different input orders.