MA-RWG: A Multi-Agent Framework for Thematically Structuring and Generation of Related Work

Zhuang Liu[8] , **Jian Liu**[1] , **Chun Kang**[1] , **Chenbin Zhang**[2] , **Rui Li**[3] , **Fanhu Zeng**[4] , **Yong Dai**[5] and **Lei Sha**[1,6,7]
[1]School of Artificial Intelligence, Beihang University
[2]University of Electronic Science and Technology of China
[3]Peking University
[4]University of Chinese Academy of Sciences
[5]x-humanoid
[6]OxTium AI Research
[7]Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing
[8]School of Foreign Languages, Beihang University
{lz20230806, aleczhang13, challengezengfh}@gmail.com, {liujian2024, kangchun, shalei}@buaa.edu.cn, o_l1ru1@stu.pku.edu.cn, daiyongya@outlook.com

Abstract

AI-driven survey generation has advanced rapidly, yet related work generation (RWG) remains relatively underexplored. Unlike surveys that provide broad literature overviews, RWG synthesizes prior studies for a single focal paper, requiring contextual fit, cross-paper comparison, and accurate attribution. To address this gap, we propose MA-RWG, a fully automated multi-agent framework that generates polished related work sections from only a title and abstract. MA-RWG first retrieves high-quality candidate papers through semantic retrieval, optionally enhanced with a diversity-aware term. It then coordinates four specialized agents for summarization, organization, integration, and fact checking, enabling DAG-based taxonomy construction, feedback-guided refinement, and dual-model verification. For evaluation, we introduce a dedicated benchmark for paper-specific related work generation, covering generation quality, citation quality, and claim-level semantic similarity. Experimental results show that MA-RWG outperforms RAG-based baselines and survey-oriented agentic methods on the RWG task. Further ablation and cross-domain experiments demonstrate the soundness and robustness of the proposed framework.

1 Introduction

The related work section serves as a critical foundation for academic papers, contextualizing the current study within the broader research landscape. Its effectiveness hinges not merely on listing prior studies, but on synthesizing existing knowledge through logically structured organization and deep comparative analysis across references. Recent progress in automated academic writing has pursued two related but distinct directions: survey generation and RWG. Survey generation systems (*e.g.*, AutoSurvey [Wang *et al.*, 2024]) aim to produce panoramic overviews of entire research domains, emphasizing coverage, classification, and trend analysis. In contrast, RWG targets the synthesis of literature specific to a single focal paper, requiring contextual fit, side-by-side comparison, and explicit positioning of novelty. While survey-style methods are valuable for domain-level knowledge curation, they do not address the paper-specific needs of RWG, where comparative analysis and precise attribution are crucial.

However, existing methods for automating RWG still fall short of core scholarly standards, especially in coherence, completeness, and fine-grained comparative synthesis across references. The seminal work of [Hoang and Kan, 2010] first formalized the task of automatic RWG and introduced ReWoS (Related Work Summarizer). ReWoS follows a hierarchical, keyword-driven paradigm that combines general content extraction (GCSum) with specific content extraction (SCSum), selecting and concatenating salient sentences from cited papers. Although effective as an early prototype, this extractive "stitching" strategy is inherently limited: it often omits important contextual details, provides only shallow modeling of inter-paper relations, and produces text that lacks fluency and in-depth cross-paper comparison.

With the rapid progress of large language models (LLMs), RWG has gradually shifted from extractive pipelines to fully generative frameworks [Brown *et al.*, 2020; Achiam *et al.*, 2023; Guo *et al.*, 2025]. Early LLM-based studies typically improve input representations by incorporating user-provided or automatically extracted keywords, thereby reducing hallucinations and enhancing factual fidelity. For example, [Li *et al.*, 2024] proposed a human-in-the-loop keyword retrieval mechanism that substantially improves the reliability and precision of generated literature summaries.

More recent studies increasingly adopt agent-based architectures to decompose RWG into specialized subtasks, leading to notable improvements in both coherence and citation accuracy. For instance, [Agarwal *et al.*, 2024b] introduced a two-stage review process that integrates keyword- and embedding-based retrieval with plan-guided generation, enhancing the consistency and depth of automated literature surveys. Extending this approach, [Liu *et al.*, 2025] proposed a multi-agent, graph-aware pipeline that enforces logical structuring and precise relationship mapping among cited works, thereby significantly improving narrative flow and citation correctness.

However, current research on related work generation still faces two notable limitations. First, existing methods typically produce discrete summaries of individual papers, lacking the categorization of similar studies and the organized structure expected in a well-written related work section. Second, a high-quality related work section requires not just a summary, but an accurate, in-depth comparative analysis. This includes methodological contrasts, choices of datasets, experimental designs, and discussions of strengths and weaknesses, all of which serve to highlight a new paper's unique contributions. However, most automated approaches omit this level of side-by-side discussion.

To address these limitations and enhance structural coherence, analytical depth, and factual precision, we introduce **MA-RWG**, an automated multi-agent pipeline for generating thematically organized related-work sections. Unlike survey-oriented systems, MA-RWG is explicitly designed for RWG, prioritizing contextual relevance and critical comparison over domain-wide coverage.

MA-RWG operationalizes RWG through four collaborative agents: (1) *Candidate Construction*, which combines semantic retrieval with an optional diversity term ranking to assemble representative candidate sets; (2) *Contextual Summarization*, where the Summarization Agent produces query-conditioned summaries that preserve essential scholarly contributions; (3) *Taxonomy Organization*, where the Organization Agent organizes the retrieved literature into a multi-dimensional directed acyclic graph (DAG) spanning Tasks, Methods, Datasets, and Evaluation under the guidance of a large language model (LLM), and derives an outline at an appropriate granularity; and (4) *Drafting, Refinement, and Verification*, where the Integration Agent drafts and iteratively refines the related-work section for clarity and critical synthesis, while the Fact-Checking Agent applies dual-model verification to detect and correct unsupported or misattributed statements. Through this integrated process, MA-RWG produces related-work sections that are coherent, analytically rigorous, and factually precise.

To evaluate our approach, we introduce a benchmark consisting of 200 research topics and a retrieval corpus of 8,000 related papers. We assess the quality of generated related work from three perspectives: citation quality, text quality, and claim-level semantic similarity. Experimental results show that our method exhibits unique advantages over RAG-based approaches and agent-based methods designed for survey generation. Meanwhile, the experimental results indicate that survey-oriented generation methods such as SurveyForge exhibit lower citation utilization and produce more diffuse descriptions. This clearly highlights the fundamental differences between survey generation and the RWG task. Further ablation and domain generalization experiments provide strong evidence for the soundness and robustness of our approach.

Our contributions are summarized as follows:

- We propose **MA-RWG**, an end-to-end agent-based system specifically designed for related-work generation, which employs four coordinated components to produce related work that is structurally rigorous, analytically comparative, and linguistically concise.
- We introduce a dedicated benchmark for paper-specific related-work generation, consisting of 200 research topics and a retrieval corpus of 8,000 papers, along with three categories of evaluation metrics covering citation quality, text quality, and claim-level semantic similarity.
- Extensive experiments show that MA-RWG exhibits unique advantages on the RWG task, achieving the best performance among the compared baselines on most evaluation metrics. Further ablation and domain-generalization studies confirm the soundness and robustness of our approach.

Our implementation is publicly available [Liu *et al.*, 2026].

2 Related Work

2.1 Automated Generation and Summarization of Related Work

The automated synthesis of related work sections faces three core challenges: thematic coherence, comparative analysis, and factual grounding. Early approaches mainly focused on content extraction and summarization. [Shah and Barzilay, 2021] proposed a tree-based content planning model, while [Wang *et al.*, 2019] developed a neural summarizer with joint attention mechanisms. However, as noted by [Mandal *et al.*, 2024], these methods often produce generic summaries lacking contextual relevance, frequently overlooking specific citation contexts. This limitation has driven the development of more sophisticated retrieval and grounding techniques.

Recent attributed generation methods significantly improve factual accuracy. [Li *et al.*, 2023] demonstrated that grounding improves when conditioning on cited text spans, while datasets like CORWA [Li *et al.*, 2022] and OARelatedWork [Docekal *et al.*, 2024] enable better training. However, multi-paper synthesis and nuanced comparative analysis remain challenging. [Liu *et al.*, 2023] introduced causal intervention to enhance coherence, and [Anand *et al.*, 2023] utilized knowledge graphs for citation generation. Despite these advances, [Martin-Boyle *et al.*, 2024] found that even state-of-the-art LLMs struggle with detailed synthesis without human intervention, highlighting the need for more robust automated solutions.

2.2 Multi-Agent Frameworks for Scientific Text Synthesis

Multi-agent systems overcome the limits of monolithic designs by dividing complex tasks into specialized components.

[Li *et al.*, 2025] introduced ChatCite, an LLM agent for comparative literature summarization, and [Kang *et al.*, 2023] developed SYNERGI for scholarly synthesis. These studies illustrate the promise of agent-based approaches but each targets only a narrow subtask. Likewise, [Torres *et al.*, 2024] automated systematic literature reviews (SLRs). However, its sequential design lacks the dynamic interaction that characterizes true multi-agent systems.

In the more challenging task of long-form survey generation, [Wang *et al.*, 2024] proposed AutoSurvey, a systematic pipeline featuring retrieval, outline generation, and multi-stage refinement. While efficient and well-structured, its reliance on static outlines constrains adaptability to complex knowledge structures. Similarly, [Yan *et al.*, 2025] developed SurveyForge, which employs logical outline analysis and a memory-augmented navigation agent. Despite its organizational strengths, the multi-agent architecture suffers from information loss during inter-agent communication and lacks posterior fact-checking, limiting output reliability despite relying on memory storage.

MA-RWG advances this line of work through orchestration rather than by claiming each module is individually new. Compared with citation-aware RWG systems such as ChatCite and graph-aware synthesis methods such as SYNERGI, MA-RWG targets the complete related-work pipeline: candidate construction, contextual summarization, multidimensional DAG organization, iterative integration, and dual-model fact verification. Compared with survey-oriented agents such as LitLLM and SurveyForge, it is optimized for paper-specific synthesis, attribution, and novelty positioning rather than broad domain coverage.

3 Multi-Agent Collaborative Generation

We propose an end-to-end framework for automated related work generation, where four collaborative agents collectively produce academically rigorous text through **precision content retrieval, structured knowledge organization**, and **dual-model factual verification**. Given the input title a_t and abstract a_i, the system generates a related work section R that achieves an optimal balance between topic relevance and the depth and breadth of related research. Figure 1 demonstrates the entire process of MA-RWG.

3.1 Data Preprocessing

This study's data preprocessing aims to efficiently identify highly relevant articles from a large literature database. We follow [Beger and Henneking, 2025]'s hybrid retrieval method and use their established database as our foundation. The process is divided into three main stages: initial candidate identification, full-text content extraction and refinement, and final selection. Data preprocessing begins with semantic retrieval. The system encodes the input title a_t and abstract a_i into a d-dimensional vector $\mathbf{v}_s \in \mathbb{R}^d$ (using the `all-mpnet-base-v2` Sentence Transformer [Reimers and Gurevych, 2019]) and simultaneously extracts topic identifiers $\mathcal{T}$ (via a fine-tuned BERTopic version [Grootendorst, 2022]). In the Milvus vector database [Wang *et al.*, 2021], initial candidate paper sets $\mathcal{C}$ are retrieved through cosine similarity. To balance relevance and diversity, we employ an iterative selection mechanism over $\mathcal{P}$ (the full set of candidate papers derived from $\mathcal{C}$):

$$i^* = \arg\max_{i \notin \mathcal{S}} \big[(1-\lambda)\cdot s_i + \lambda \cdot \big(1 - \min_{p_j \in \mathcal{S}} \mathrm{sim}(e_i, e_j)\big)\big] \quad (1)$$

Here, $\mathcal{P} = \{p_1, p_2, \cdots, p_n\}$ denotes all candidate papers (expanded from the initial $\mathcal{C}$), and $\mathcal{S} \subseteq \mathcal{P}$ is the subset of papers already selected in the iteration. Each paper p_i is associated with an abstract embedding $e_i \in \mathbb{R}^d$ (generated using the same `all-mpnet-base-v2` model). $\mathrm{sim}(e_i, e_j)$ represents the cosine similarity between embeddings of p_i and p_j; s_i quantifies the relevance of p_i to the input (specifically, the average cosine similarity between e_i and $\mathbf{v}_s$, the joint embedding of a_t and a_i).

The diversity weight $\lambda \in [0, 1]$ modulates the trade-off: $\lambda = 0$ prioritizes pure relevance (selecting papers most similar to the input), while $\lambda > 0$ rewards papers that are dissimilar to at least one already selected paper (via $1 - \min_{p_j \in \mathcal{S}} \mathrm{sim}(e_i, e_j)$, which increases as p_i becomes less similar to the "least similar" paper in $\mathcal{S}$). By iteratively applying this criterion, we construct a longlist $\mathcal{L}$ containing $3 * B$ papers (where B is the breadth parameter, defining the size of the final refined set), enabling flexible control over the balance between input relevance and perspective diversity.

Next, each paper in the longlist undergoes full-text processing, filtering out non-critical sections. To precisely extract content, we leverage the aforementioned balancing mechanism. Specifically, each page is treated as a candidate unit, and the principle of the above formula is reapplied. Based on its similarity to the input abstract (s_i) and its diversity relative to already selected pages, the k most relevant pages are chosen, given a depth parameter D.

Finally, the key pages from the longlist papers are integrated with their abstracts to form an aggregated content representation for each paper. Through further average similarity filtering of these aggregated contents, a refined shortlist S_t containing B papers is determined. These papers in the shortlist constitute all relevant sources for subsequent content synthesis and analysis. Although our experiments use a fixed benchmark corpus for reproducibility, the retrieval stage is index-based: newly ingested papers can be encoded with the same sentence encoder and inserted into the vector database without retraining the downstream agents; taxonomy nodes are then rebuilt for the retrieved shortlist at generation time.

3.2 Summarization Agent

The summarization agent synthesizes *contextualized* summaries for shortlisted papers. For each paper $P_i \in \mathcal{S}_t$, it conditions a large language model $\mathrm{LLM}_{\mathrm{sum}}$ on three complementary sources: the original abstract a_i, the paper's own abstract b_i, and the top-D page contents $\{p^*_{i1}, \ldots, p^*_{iD}\}$. The contextualized summary s'_i is then produced by executing the following prompt:

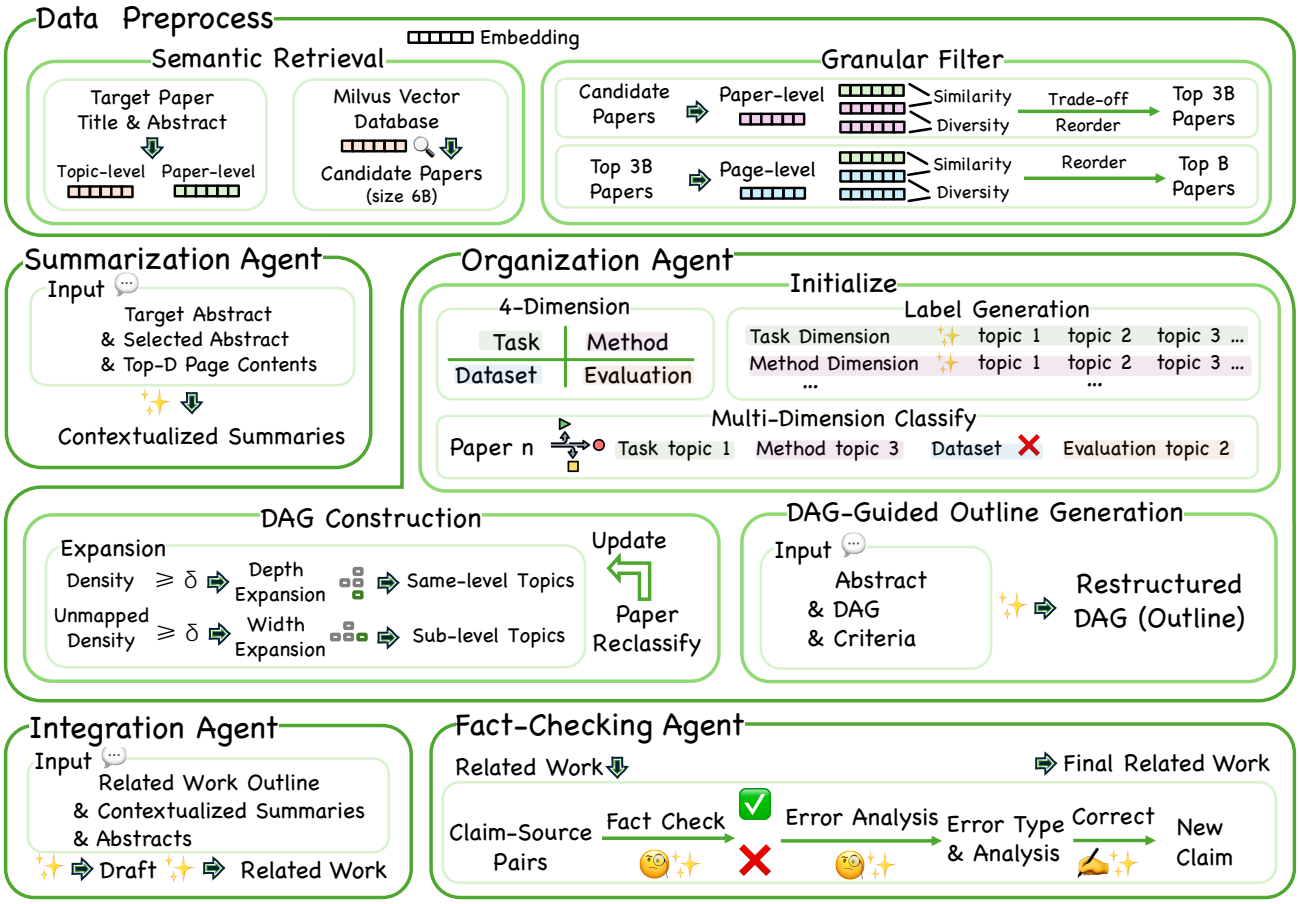

Figure 1: The **MA-RWG** framework is a multi-agent system for related-work synthesis. It starts with **Data Preprocessing** using semantic retrieval and diversity reordering to select candidate papers. The **Summarization Agent** produces *contextualized* summaries focusing on core contributions and comparative analysis. The **Organization Agent** builds a multidimensional DAG taxonomy (*Task*, *Method*, *Dataset*, *Evaluation*) and performs **DAG-Guided Outline Generation** to create a structured outline. The **Integration Agent** composes and refines the draft for coherent flow and critical synthesis. The **Fact-Checking Agent** verifies atomic claims and citation accuracy via dual-model verification and automated correction.

$$s_i' = \text{LLM}_{\text{sum}}\left(\begin{array}{c}\text{Summarize [Paper Title]:}\\ a_i,\ b_i\\ \text{within the context of } [S].\\ \text{Key pages: } \{p_{i1}^*, \ldots, p_{iD}^*\}\end{array}\right). \tag{2}$$

This design encourages s_i' to faithfully capture the paper's core contributions while remaining tightly aligned with the research context represented by the set $\mathcal{S}_t$.

3.3 Organization Agent

The **organization agent** constructs a multidimensional taxonomy by structuring input papers along four orthogonal dimensions $D = \{\text{Task}, \text{Methodology}, \text{Dataset}, \text{Evaluation}\}$, inspired by [Kargupta *et al.*, 2025]. To begin this process, we utilize a LLM-based multi-label classification approach. Given each paper's title and abstract, the LLM first categorizes the paper into one or more of the four primary dimensions in D, forming possibly overlapping subsets. Subsequently, for each $d \in D$, it generates a directed acyclic graph T_d supporting multi-parent relationships.

Initial DAG Construction. The initial DAG construction proceeds through two integrated stages–hierarchical expansion and semantic-aware clustering. This process terminates when either the density signal (node paper count or unclassified paper count) falls below a threshold δ, or the depth reaches $l_{\max}$. We frame DAG construction as an LLM-based multi-label classification task: Given each paper's title and abstract, the model partitions the collection into four (potentially overlapping) subsets corresponding to D. Formally, for each topic node $n_{i,d}$ (where i indexes nodes under dimension $d \in D$), we compute:

$$\rho(n_{i,d}) := |P_{i,d}|, \quad \tilde{\rho}(n_{i,d}) := |U_{i,d}|, \tag{3}$$

where:

- $P_{i,d} \subseteq \mathcal{S}_{\text{short}}$ denotes the set of papers explicitly associated with node $n_{i,d}$;
- $\rho(n_{i,d})$ (node paper count) is the size of $P_{i,d}$;

- $U_{i,d} \subseteq P_{i,d}$ denotes the subset of papers in $P_{i,d}$ that are *not classified by any child node of* $n_{i,d}$;
- $\tilde{\rho}(n_{i,d})$ (unclassified paper count) is the size of $U_{i,d}$, excluding papers partially classified by child nodes.

Expansion is triggered by the following conditions:

$$\begin{cases} \rho(n_{i,d}) \geq \delta \ \wedge \ \text{level}(n_{i,d}) < l_{\max}, \\ \tilde{\rho}(n_{i,d}) \geq \delta \ \wedge \ \mathcal{C}(n_{i,d}) \neq \emptyset, \end{cases} \tag{4}$$

where:

- $\text{level}(n_{i,d}) \in \mathbb{N}^+$ is the depth of node $n_{i,d}$ in T_d (root nodes have level $= 1$);
- $\mathcal{C}(n_{i,d})$ is the set of child nodes of $n_{i,d}$ (*i.e.*, $\mathcal{C}(n_{i,d}) = \{n_{j,d} \mid n_{i,d} \text{ is a parent of } n_{j,d}\}$).

Hierarchical expansion is achieved through two modes:

- *Depth expansion*: Adding fine-grained child nodes to leaf nodes (where $\mathcal{C}(n_{i,d}) = \emptyset$) with high paper density ($\rho(n_{i,d}) \geq \delta$);
- *Width expansion*: Adding new sibling nodes to non-leaf nodes (where $\mathcal{C}(n_{i,d}) \neq \emptyset$) whose existing children fail to cover unclassified papers ($\tilde{\rho}(n_{i,d}) \geq \delta$).

This process is iterative, with ρ and $\tilde{\rho}$ guiding both the direction of growth (depth vs. width) and termination (when neither condition is met).

Semantic-aware clustering is triggered when a node requires expansion, ensuring semantic validity of new entities through two sequential sub-steps:

1. *Subtopic pseudo-labeling*: The LLM generates pseudo-labels for papers in $U_{i,d}$, tailored to the dimension d and granularity matching $\text{level}(n_{i,d})$;
2. *Subtopic clustering*: These pseudo-labels are grouped into coherent subtopic clusters (via cosine similarity on label embeddings), forming new child nodes.

This dual-phase process optimizes corpus coverage, minimizes redundancy, and preserves semantic integrity during expansion.

DAG-Guided Outline Generation. At this stage, the system constructs a structured prompt:

$$\mathcal{T}_{\text{outline}} = f_{\text{template}}(A, \mathcal{G}). \tag{5}$$

Here, A denotes the input abstract, and $\mathcal{G}$ represents the complete set of dimension-specific DAGs:

$$\mathcal{G} = \{T_d \mid d \in D\}. \tag{6}$$

This prompt directs the LLM to analyze the hierarchical organization of $\mathcal{G}$ and to select an appropriate granularity for the resulting outline. Conditioned on this input, the LLM $\mathcal{M}$ generates an outline that is thematically coherent and hierarchically aligned:

$$\mathcal{O} = \mathcal{M}(\mathcal{T}_{\text{outline}}), \tag{7}$$

where $\mathcal{O} = \{(t_i, l_i) \mid 2 \leq |\mathcal{O}| \leq 4,\ l_i \leq 10\}$. In this expression, $|\mathcal{O}|$ gives the number of topics in the outline, which ranges from 2 to 4. Meanwhile, l_i specifies the maximum allowable word length for each topic t_i, which is capped at 10 words. The density-based distribution analysis performed by the LLM on $\mathcal{G}$ ensures that the resulting outline aligns with the semantic structure found in the input papers.

3.4 Integration Agent

The process for generating and refining the related work text within our system comprises two main stages: feedback generation and content integration.

Initially, an initial draft of the related work is produced by a large language model, leveraging reclassified papers, their original abstracts, and contextualized summaries.

Subsequently, feedback is provided on this initial draft for revision, evaluating four dimensions: structural logic, critical analysis, prominent contribution, and concise language.

Finally, an integration agent refines the related work section based on this feedback. The revision process, applied within each subsection, focuses on four key aspects:

The revision focuses on four dimensions. *Reorganized Structure* ensures that the content is presented following a clear logical progression from problem formulation to method classification, then to limitations, and finally to our innovation. *Strengthen criticism* emphasizes that, after introducing each category of methods, their limitations are explicitly analyzed and directly connected to the motivation of our work. *Simplified language* aims to merge citations, remove redundant descriptions, and highlight the core viewpoints to improve clarity and conciseness. *Comparison clarity* explicitly states in the conclusion why our method outperforms prior approaches.

Throughout the revision process, the system strictly adheres to the established outline of the related work section.

3.5 Fact-Checking Agent

To ensure rigorous academic integrity, the agent implements a dual-model verification system for every pair of claim-source(s) in the generated text. For each claim c with cited passages $\mathcal{R}_c = \{r_1, \ldots, r_k\}$:

$$\text{Valid}(c) = \begin{cases} 1 & \text{if Model}_1\,(c, \mathcal{R}_c) = \text{supported} \\ & \quad \wedge\ \text{Model}_2\,(c, \mathcal{R}_c) = \text{supported} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

Claims are classified into error categories via ensemble labeling:

$$\text{ErrorType}(c) = \text{mode}\left(\left[f_{\text{cls}}^{\text{Model}_1}(c, \mathcal{R}_c),\ f_{\text{cls}}^{\text{Model}_2}(c, \mathcal{R}_c)\right]\right) \tag{9}$$

where error types:

$$\mathcal{E} = \{\text{Direct Contradiction}, \text{Unsubstantiated}, \\ \text{Misrepresentation}, \text{Incorrect Attribution}, \text{Other}\}.$$

Detected erroneous claims are automatically rectified by Model_1:

$$c' = \text{Model}_1\begin{pmatrix} \texttt{Correct claim: [c]} \\ \texttt{for error: [ErrorType]} \\ \texttt{using sources: } \mathcal{R}_c \end{pmatrix} \tag{10}$$

Corrected claims are re-verified:

$$c_{\text{final}} = \begin{cases} c' & \text{if Valid}(c') = 1 \\ \text{REMOVE} & \text{otherwise} \end{cases} \tag{11}$$

3.6 Integrated Workflow

The pipeline is formalized as functional composition:

$$R_{\text{final}} = \mathcal{F}_{\text{fact-check}} \circ \mathcal{F}_{\text{integrate}} \circ \mathcal{F}_{\text{organize}} \circ \mathcal{F}_{\text{summarize}} \circ \mathcal{F}_{\text{preprocess}}(S; \Theta) \tag{12}$$

where the final result R_{final} is generated by sequentially applying the preprocessing function $\mathcal{F}_{\text{preprocess}}$, the summarization function $\mathcal{F}_{\text{summarize}}$, the organization function $\mathcal{F}_{\text{organize}}$, the integration function $\mathcal{F}_{\text{integrate}}$, and finally the fact-checking function $\mathcal{F}_{\text{fact-check}}$ to the input source S.

This functional sequence is controlled by a global parameter set, $\Theta = \{B, D, \lambda\}$, which jointly governs the workflow's behavior and output characteristics. The structural granularity is specified by two parameters: the paper breadth B defines the cross-document scope, while the page depth D determines the level of detail within each document. λ controls diversity among intermediate candidate sets.

4 Experiments

4.1 Experimental Settings

Baselines. We compare against four representative baselines, covering both retrieval-augmented generation (RAG) and agent-based survey generation: (i) two RAG variants, Naive RAG and CiteGeist; and (ii) two agent-based methods, LitLLM and SurveyForge. CiteGeist [Beger and Henneking, 2025] is a dynamic arXiv-based retrieval-augmented system for related-work generation. Naive RAG follows a standard retrieve-then-generate pipeline, using Gemini 2.5 Flash as the generator [Google, 2025]. LitLLM [Agarwal *et al.*, 2024a] is a multi-agent framework that integrates academic paper retrieval into the generation loop and employs multiple prompts to iteratively produce citation-aware academic surveys. SurveyForge [Yan *et al.*, 2025] likewise uses a structured outline to steer generation; additionally, it introduces memory-driven generation and multi-dimensional evaluation to automatically produce structured academic surveys. For fairness, all baselines share the same retrieval source, embedding model, vector database, and retrieval hyperparameters where applicable; we only replace the retrieval module and preserve each baseline's original generation logic.

Datasets. We construct a benchmark consisting of 200 research topics (including title and abstract) and a literature retrieval corpus of 8,000 papers. Each research topic is paired with a human-written related work section serving as the gold label. We sample 200 AI/ML papers published between 2020 and 2024 from ScholarCopilot [Wang *et al.*, 2025], covering ten themes such as large language models, multimodal foundation models, diffusion models, graph neural networks, reinforcement learning, federated learning, interpretability, compression, adversarial robustness, and continual learning. For each paper, we collect references cited in its related-work section, yielding 6,519 relevant papers, and add 1,481 same-area distractor papers. For the reported full-system comparisons, we evaluate all methods on a 50-paper subset sampled from this benchmark, using the same fixed corpus and metric protocol. This construction gives controlled gold references for citation diagnostics, while still leaving open-world retrieval and hard-negative mining as limitations.

Implementation Details. We implement the dual-model verification module by coupling a lightweight generator with an independent verifier. Concretely, the generator is instantiated with Gemini 2.5 Flash, while the verifier is DeepSeek -V3-0324 [Liu *et al.*, 2024], which re-checks the drafted response for factual consistency. We set the maximum number of verification attempts to 1, *i.e.*, the system performs exactly one fact-checking pass before finalizing the output.

Hyperparameters Setup. The basic hyperparameter settings used in the experiments are as follows. Breadth and depth parameters are set to B (corresponding to the preset number of citations for ground truth related work per round) and $D = 3$ respectively. The diversity weight was set to $\lambda = 0$ in the main experiments, reducing the retrieval score to relevance-only ranking for controlled comparison. The diversity term is retained as an optional component of the framework. As part of the DAG construction process, the density signal threshold (δ) is set to 5, and the maximum depth ($l_{\max}$) is set to 3.

4.2 Main Evaluation

Evaluation Metrics. We report three complementary metric families (Table 1). **(1) Citation quality** measures claim–source correctness and evidence coverage: *Claim Precision (ClaP)* is the fraction of correctly supported cited claims; *Citation Precision (CitP)* is the fraction of correct citations among all citations; *Reference Coverage (RC)* measures how well gold references are covered by generated citations; and *Average Citations per Claim (CC)* is the mean number of citations per cited sentence. **(2) Text quality** evaluates overall writing quality via an LLM-judged *Text Quality (TQ)* score averaged over five dimensions (*Clarity*, *Structure*, *Relevance*, *Motivation*, and *Critics*). **(3) Semantic alignment** quantifies semantic overlap with human-written related work using embedding-based *SN-Precision (SN-P)*, *SN-Recall (SN-R)*, and *SN-F1*. These definitions make citation correctness, coverage, and semantic overlap explicit at the sentence/reference level.

Results. Table 1 presents the main experimental results. MA-RWG delivers the best overall performance, showing consistent advantages in citation quality, text quality, and semantic alignment. It achieves the highest scores in **CitP** (0.910), **RC** (0.812), **TQ** (4.046), and **SN-F1** (0.624), and remains highly competitive in **ClaP** (0.900). This indicates that the proposed framework enhances evidence coverage and writing quality in a balanced manner, instead of relying on a trade-off between the two.

The comparison also clarifies the difference between RWG and survey generation. SurveyForge has competitive text quality but much lower reference coverage, whereas RAG baselines improve claim-level fidelity but lack reflective polishing and citation validation. MA-RWG combines retrieval completeness with post-hoc verification, and averages about 1100 ± 30 seconds per article in our implementation; the DAG stage dominates runtime, so simpler RAG pipelines may be preferable for latency-critical use. The ICC across three LLM-scoring runs is 0.9713, but we still treat LLM-judged text quality as a limitation and complement it with

Method	ClaP	CitP	RC	CC	TQ	SN-P	SN-R	SN-F1
MA-RWG	0.900	**0.910**	**0.812**	**1.746**	**4.046**	0.607	**0.667**	**0.624**
LitLLM	0.385	0.200	0.751	1.323	3.290	0.320	0.157	0.204
NaiveRAG	**0.917**	0.827	0.764	1.137	2.950	0.455	0.301	0.359
CiteGeist	0.878	0.757	0.789	1.400	3.178	0.554	0.368	0.433
SurveyForge	0.470	0.510	0.350	1.200	3.992	**0.609**	0.374	0.457

Table 1: Results of the five methods on the main experiments

Method	ClaP	CitP	RC	CC	TQ	SN-P	SN-R	SN-F1
w/o summarization	0.803	0.745	0.786	1.194	3.37	0.568	0.380	0.445
w/o DAG	0.832	0.860	0.782	1.616	3.23	0.537	0.435	0.475
w/o fact check	0.452	0.489	0.804	1.472	3.49	0.409	0.237	0.289
w/o feedback	0.890	0.900	0.793	1.189	3.35	0.567	0.349	0.423
MA-RWG (Full)	0.900	0.910	0.812	1.746	3.96	0.607	0.667	0.624

Table 2: Ablation results on citation quality with extended evaluation dimensions.

citation-level and semantic-overlap metrics.

4.3 Ablation Study

Table 2 shows that all four components are necessary. Removing the Summarization Agent weakens context-sensitive evidence alignment; removing the DAG disrupts narrative structure and reduces TQ by 18.4%; removing the Fact-Checking Agent causes the steepest drop in citation precision (46.3%) and claim precision (49.8%); and removing feedback-based integration reduces evaluative depth and clarity. These distinct failure modes support the need for coordinated retrieval, organization, refinement, and verification.

4.4 Domain Generalization Analysis

We further evaluate MA-RWG on 50 cross-domain papers from ICML/NeurIPS AI4Science workshops, including biology, materials science, and fluid physics. MA-RWG again outperforms the baselines on the key indicators (CitP 0.899, TQ 3.922, SN-F1 0.576). By contrast, LitLLM drops to CitP 0.256 and SN-F1 0.254, while SurveyForge remains comparatively diffuse despite reasonable writing quality. This suggests that the gains of MA-RWG come from RWG-specific grounding and organization rather than domain-specific overfitting.

5 Limitations and Future Work

MA-RWG is an author-assistive system, not a replacement for scholarly judgment. It currently uses only title and abstract inputs, whereas full text and multimodal materials could support deeper comparison. The benchmark uses a fixed 8,000-paper corpus for reproducibility; real deployments should refresh embeddings and check coverage of canonical works as new papers appear. The pipeline also relies on proprietary LLMs, incurs about 1100 ± 30 seconds per article, and uses partly LLM-judged text-quality metrics, motivating future work on open models, hard-negative retrieval diagnostics, human evaluation, and cost-aware agent selection.

6 Conclusion

We presented **MA-RWG**, a multi-agent framework that generates thematically structured and fact-grounded related work sections from only a title and abstract. By integrating contextual summarization, DAG-based organization, feedback-guided synthesis, and dual-model fact checking, MA-RWG provides a principled pipeline for improving both content organization and factual reliability. Experiments on the dedicated RWG benchmark and cross-domain evaluation demonstrate consistent improvements in structure, coverage, and attribution quality.

Ethical Statement

Automated related-work generation can reduce literature-review burden, but it also risks encouraging over-reliance on generated scholarly text. MA-RWG therefore includes claim-level citation verification and correction, and its outputs should be treated as author-assistive drafts that require human review before submission. Authors remain responsible for checking whether cited papers genuinely support each comparative claim, whether important literature is missing, and whether the final text accurately represents prior work.

Acknowledgments

This work was supported by the National Science Fund for Excellent Young Scholars (Overseas) under Grant No. KZ37117501, the National Natural Science Foundation of China under Grant No. 62306024, the Fundamental Research Funds for the Central Universities, and the Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing.

Contribution Statement

Zhuang Liu and Jian Liu contributed equally to this work and are co-first authors. Lei Sha and Chun Kang are co-corresponding authors. All authors contributed to the development of this work, discussed the results, reviewed the manuscript, and approved the final version.

References

[Achiam *et al.*, 2023] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[Agarwal *et al.*, 2024a] Shubham Agarwal, Gaurav Sahu, Abhay Puri, Issam H Laradji, Krishnamurthy DJ Dvijotham, Jason Stanley, Laurent Charlin, and Christopher Pal. Litllm: A toolkit for scientific literature review. *arXiv preprint arXiv:2402.01788*, 2024.

[Agarwal *et al.*, 2024b] Shubham Agarwal, Gaurav Sahu, Abhay Puri, Issam H Laradji, Krishnamurthy Dj Dvijotham, Jason Stanley, Laurent Charlin, and Christopher Pal. Litllms, llms for literature review: Are we there yet? *arXiv preprint arXiv:2412.15249*, 2024.

[Anand *et al.*, 2023] Avinash Anand, Mohit Gupta, Kritarth Prasad, Ujjwal Goel, Naman Lal, Astha Verma, and Rajiv Ratn Shah. Kg-ctg: Citation generation through knowledge graph-guided large language models. In *International Conference on Big Data Analytics*, pages 37–49. Springer, 2023.

[Beger and Henneking, 2025] Claas Beger and Carl-Leander Henneking. Citegeist: Automated generation of related work analysis on the arxiv corpus. *arXiv preprint arXiv:2503.23229*, 2025.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[Docekal *et al.*, 2024] Martin Docekal, Martin Fajcik, and Pavel Smrz. Oarelatedwork: A large-scale dataset of related work sections with full-texts from open access sources. *arXiv preprint arXiv:2405.01930*, 2024.

[Google, 2025] Google. Gemini 2.5 flash. https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash, 2025. Accessed: 2026-05-11.

[Grootendorst, 2022] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[Guo *et al.*, 2025] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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

[Kang *et al.*, 2023] Hyeonsu B Kang, Tongshuang Wu, Joseph Chee Chang, and Aniket Kittur. Synergi: A mixed-initiative system for scholarly synthesis and sensemaking. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–19, 2023.

[Kargupta *et al.*, 2025] Priyanka Kargupta, Nan Zhang, Yunyi Zhang, Rui Zhang, Prasenjit Mitra, and Jiawei Han. Taxoadapt: Aligning llm-based multidimensional taxonomy construction to evolving research corpora. *arXiv preprint arXiv:2506.10737*, 2025.

[Li *et al.*, 2022] Xiangci Li, Biswadip Mandal, and Jessica Ouyang. Corwa: A citation-oriented related work annotation dataset. *arXiv preprint arXiv:2205.03512*, 2022.

[Li *et al.*, 2023] Xiangci Li, Yi-Hui Lee, and Jessica Ouyang. Cited text spans for citation text generation. *arXiv preprint arXiv:2309.06365*, 2023.

[Li *et al.*, 2024] Xiangci Li, Yi-Hui Lee, and Jessica Ouyang. Cited text spans for scientific citation text generation. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 90–104, 2024.

[Li *et al.*, 2025] Yutong Li, Lu Chen, Aiwei Liu, Kai Yu, and Lijie Wen. Chatcite: Llm agent with human workflow guidance for comparative literature summary. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3613–3630, 2025.

[Liu *et al.*, 2023] Jiachang Liu, Qi Zhang, Chongyang Shi, Usman Naseem, Shoujin Wang, and Ivor Tsang. Causal intervention for abstractive related work generation. *arXiv preprint arXiv:2305.13685*, 2023.

[Liu *et al.*, 2024] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[Liu *et al.*, 2025] Xiaochuan Liu, Ruihua Song, Xiting Wang, and Xu Chen. Select, read, and write: A multi-agent framework of full-text-based related work generation. *arXiv preprint arXiv:2505.19647*, 2025.

[Liu *et al.*, 2026] Zhuang Liu, Jian Liu, Chun Kang, Chenbin Zhang, Rui Li, Fanhu Zeng, Yong Dai, and Lei Sha. MA-RWG: Multi-agent framework source code. https://github.com/Felix-liu0989/Multi_Agents_for_Citation_Verification, 2026. Accessed: 2026-05-11.

[Mandal *et al.*, 2024] Biswadip Mandal, Xiangci Li, and Jessica Ouyang. Contextualizing generated citation texts. *arXiv preprint arXiv:2402.18054*, 2024.

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

[Reimers and Gurevych, 2019] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 3982–3992, 2019.

[Shah and Barzilay, 2021] Darsh J Shah and Regina Barzilay. Generating related work. *arXiv preprint arXiv:2104.08668*, 2021.

[Torres *et al.*, 2024] João Pedro Fernandes Torres, Catherine Mulligan, Joaquim Jorge, and Catarina Moreira. Promptheus: A human-centered pipeline to streamline slrs with llms. *arXiv preprint arXiv:2410.15978*, 2024.

[Wang *et al.*, 2019] Yongzhen Wang, Xiaozhong Liu, and Zheng Gao. Neural related work summarization with a joint context-driven attention mechanism. *arXiv preprint arXiv:1901.09492*, 2019.

[Wang *et al.*, 2021] Jianguo Wang, Xiaomeng Yi, Rentong Guo, Hai Jin, Peng Xu, Shengjun Li, Xiangyu Wang, Xiangzhou Guo, Chengming Li, Xiaohai Xu, et al. Milvus: A purpose-built vector data management system. In *Proceedings of the 2021 international conference on management of data*, pages 2614–2627, 2021.

[Wang *et al.*, 2024] Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Qingsong Wen, Wei Ye, et al. Autosurvey: Large language models can automatically write surveys. *Advances in neural information processing systems*, 37:115119–115145, 2024.

[Wang *et al.*, 2025] Yubo Wang, Xueguang Ma, Ping Nie, Huaye Zeng, Zhiheng Lyu, Yuxuan Zhang, Benjamin Schneider, Yi Lu, Xiang Yue, and Wenhu Chen. Scholarcopilot: Training large language models for academic writing with accurate citations. *arXiv preprint arXiv:2504.00824*, 2025.

[Yan *et al.*, 2025] Xiangchao Yan, Shiyang Feng, Jiakang Yuan, Renqiu Xia, Bin Wang, Bo Zhang, and Lei Bai. Surveyforge: On the outline heuristics, memory-driven generation, and multi-dimensional evaluation for automated survey writing. *arXiv preprint arXiv:2503.04629*, 2025.