MVSS: A Unified Framework for Multi-View Structured Survey Generation

Anonymous ACL submission

Abstract

Scientific surveys require not only summarizing large bodies of literature, but also organizing them into clear and coherent conceptual structures. However, existing automatic survey generation methods typically focus on linear text generation and struggle to explicitly model hierarchical relations among research topics and structured methodological comparisons, resulting in substantial gaps in structural organization and evidence presentation compared to expert-written surveys. To address this limitation, we propose **MVSS**, a *multi-view structured survey generation* framework that jointly generates and aligns citation-grounded hierarchical trees, structured comparison tables, and survey text. MVSS follows a structure-first paradigm: it first constructs a tree that captures the conceptual organization of a research domain, then generates comparison tables constrained by the tree structure, and finally uses both the tree and tables as joint structural constraints to guide outline construction and survey text generation. This design enables complementary and aligned multi-view representations across structure, comparison, and narrative. In addition, we introduce a dedicated evaluation framework that systematically assesses generated surveys from multiple dimensions, including structural quality, comparative completeness, and citation fidelity. Through large-scale experiments on **76** computer science topics, we demonstrate that MVSS significantly outperforms existing methods in survey organization and evidence grounding, and achieves performance comparable to expert-written surveys across multiple evaluation metrics.

1 Introduction

In several frontier areas of natural language processing, methodological innovations are advancing at a pace that far outstrips the ability of manually written surveys to keep up. In rapidly evolving domains such as large language models, retrieval-augmented generation, and multimodal reasoning (Brown et al., 2020; OpenAI, 2023; Zhao et al., 2023; Lewis et al., 2020; Gao et al., 2023; Yin et al., 2024), major breakthroughs often emerge within months, while comprehensive surveys typically lag behind by one or more publication cycles.

Recently, LLM-based automated survey generation agents have attracted growing attention in both academia and industry. Representative academic systems include AutoSurvey (Huang et al., 2020; Wang et al., 2024), SurveyGen (Bao et al., 2025), and related agent-based approaches (Qi et al., 2025; Ali et al., 2024). In parallel, industrial research assistants such as OpenAI's Deep Research (OpenAI, 2024) and Google Gemini (DeepMind, 2023) have demonstrated strong capabilities in large-scale literature understanding and synthesis. However, the majority of these systems generate surveys primarily as linear narratives, overlooking other critical dimensions commonly present in human-written surveysmost notably citation trees and citation tables.

A *citation tree* is a hierarchical structure used to organize survey content, where each node corresponds to a research concept or sub-area and is explicitly associated with a set of representative supporting papers. The hierarchical relations between nodes reflect the progressive decomposition of a research topic from high-level themes to finer-grained directions. A *citation table*, in contrast, provides a structured tabular representation of survey content, where each row corresponds to a method or research line, each column captures a key attribute or comparison dimension, and explicit citations link table entries to concrete evidence in the literature. By analyzing surveys published between 2021 and 2025 in the large language model domain, we find that over 80% of surveys include explicit hierarchical trees. Moreover, user studies indicate that more than 99% of

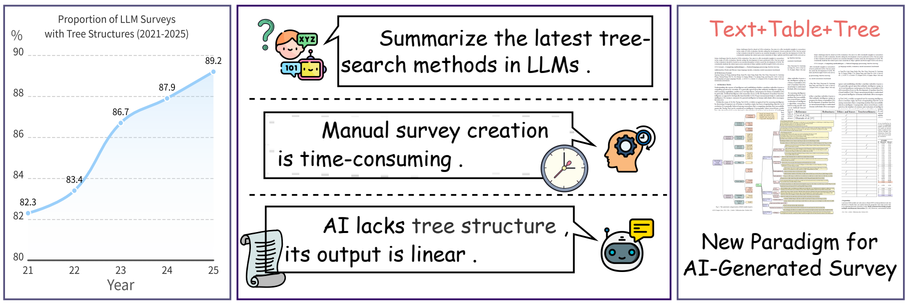

Figure 1: **Motivation for structure-first survey generation.** Linear-text survey generation obscures evolving comparison dimensions, motivating aligned multi-view structures. Expert surveys predominantly adopt explicit hierarchies (2021–2025), supporting a structure-first perspective.

readers tend to first consult trees or tables, rather than raw text, when reading survey papers.

Among existing methods, HiReview (Hu et al., 2025) constructs citation trees using clustering-based taxonomies derived primarily from paper titles. Such title-driven retrieval often misses papers whose relevance is not explicitly reflected in their titles and fails to leverage information beyond surface-level metadata, leading to unstable and noisy hierarchies. Furthermore, HiReview evaluates tree quality only indirectly through downstream survey text quality, rather than performing principled assessments of the taxonomy itself. In human-written surveys, citation trees and comparison tables are highly correlated and jointly used to organize content, yet HiReview does not generate or reason over tables at all.

To address these limitations, we propose **Multi-View Structured Survey (MVSS)**, a framework that formulates survey generation as a *joint structural generation* problem rather than a purely linear text generation pipeline. To construct rich and well-founded hierarchies, we introduce **Hierarchical Knowledge Trees (HKT)**, a principled representation for modeling the conceptual organization of a research domain. HKT explicitly externalizes the latent conceptual hierarchy in surveys as a citation-grounded tree, enabling systematic modeling of the branching evolution of ideas, the formation of subfields, and relationships among methodological paradigms. Building on HKT, we further propose a structure-aware generation mechanism that logically anchors and constrains table construction, producing structured comparison tables that remain consistent with the overall survey in terms of conceptual level, comparison dimensions, and evidence support. Finally, MVSS uses both trees and tables as joint structural constraints to guide outline construction and generate high-quality survey text. From a technical perspective, MVSS is not a simple assembly of modules, but a reformulation of survey generation as a cross-view structural alignment problem.

We systematically evaluate MVSS on 76 computer science topics under different target lengths (8k, 16k, 32k, and 64k tokens), comparing against AutoSurvey, HiReview, naive RAG-based generation, and expert-written surveys. Across all settings, MVSS consistently achieves the strongest overall performance. In particular, at shorter lengths (8k and 16k tokens), MVSS improves coverage, structural quality, and relevance by a clear margin over automated baselines, while at longer lengths (32k and 64k tokens) it reaches near-expert performance, achieving average scores close to the maximum. Additional citation-level and structural evaluations further show that MVSS maintains reliable evidence alignment, with citation precision and recall both exceeding 75% across different LLM judges, and produces high-quality trees and tables with scores consistently above 4.8. Overall, these results indicate that MVSS effectively bridges the gap between automated survey generation and expert-written surveys in both structural organization and evidence grounding.

2 Related Work

2.1 Automatic Survey Generation

Automated survey generation has attracted increasing attention, with recent work exploring neural systems for assisting or automating literature reviews and survey-style overviews (Portenoy and West, 2020; Kasanishi et al., 2023; Darrin et al., 2024; Gonzalez Bonorino, 2023). Most approaches formulate the task as multi-document summarization, producing a single linear narrative

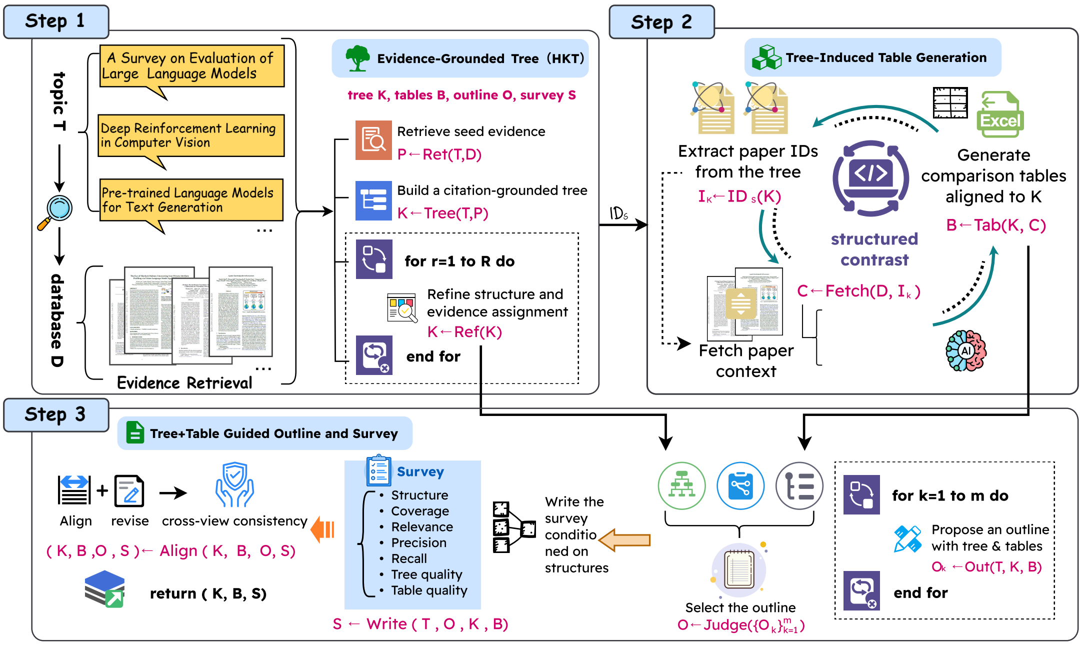

Figure 2: **Overview of MVSS.** Given a topic T and a paper database D, MVSS constructs an evidence-grounded hierarchical tree, generates aligned comparison tables, and produces a structured survey via cross-view alignment.

synthesized from retrieved papers (Christensen et al., 2014; Celikyilmaz et al., 2010; Liu and Lapata, 2019; Yasunaga et al., 2019; Li et al., 2023; Zhang et al., 2024). Some methods introduce limited structural bias, such as keyphrase-aware modeling (Liu et al., 2021), while retrieval-augmented pipelines improve factual grounding via evidence retrieval and neural ranking (Izacard and Grave, 2021; Nogueira and Cho, 2019). Recent LLM-based systems further emphasize citation-aware generation and critique-based refinement to improve reliability (Kryściński et al., 2020; Dixit et al., 2023; Kasanishi et al., 2023; Madaan et al., 2023; Shinn et al., 2023).

Despite these advances, existing systems remain fundamentally *text-centric*, relying on intermediate artifacts such as paper lists or outlines that are not optimized as explicit knowledge structures, limiting their ability to capture conceptual hierarchies or systematic comparisons. MVSS instead adopts a *structure-first* perspective, treating hierarchical knowledge trees, structured comparison tables, and cross-view alignment as primary optimization objectives rather than byproducts of text generation.

2.2 Knowledge Structuring and Multi-Signal Evaluation

Prior work on scientific knowledge organization includes document embeddings, graph-based retrieval, and bibliometric clustering. Domain-specific encoders capture semantic similarity at the paper level (Beltagy et al., 2019; Cohan et al., 2020), while graph-based and topic-structure models often lack clear hierarchical abstraction (Kasela et al., 2025). Bibliometric methods provide coarse global structure but are insufficient for fine-grained, citation-grounded survey synthesis (Waltman and van Eck, 2012).

Separately, multi-signal evaluation has improved the reliability of long-form LLM outputs. Reflection-based self-revision (Madaan et al., 2023; Shinn et al., 2023), LLM-as-a-judge scoring (Fu et al., 2024; Bhat and Varma, 2023; Bansal and Sharma, 2023), fact verification (Kryściński et al., 2020; Dixit et al., 2023), and reranking techniques (Nogueira and Cho, 2019) provide complementary correction signals. However, these methods still treat structure as secondary to text generation and do not enforce coherent organization across survey views.

MVSS bridges this gap by making structure itself the optimization target. By jointly modeling hierarchical abstraction and multi-signal verification, MVSS ensures that trees, tables, and text form a coherent, navigable, and faithfully grounded representation of the research landscape, enabling more interpretable and reliable survey generation.

Algorithm 1 MVSS: Multi-View Structured Survey Generation

Require: topic T, paper database D
Ensure: tree K, tables B, outline O, survey S
  Phase 1: Evidence-Grounded Tree (HKT)
  Retrieve seed evidence: $P \leftarrow \mathrm{Ret}(T, D)$
  Build a citation-grounded tree: $K \leftarrow \mathrm{Tree}(T, P)$
  for $r = 1$ to R **do**
    Refine structure and evidence assignment: $K \leftarrow \mathrm{Ref}(K)$
  end for
  Phase 2: Tree-Induced Table Generation
  Extract paper IDs from the tree: $I_K \leftarrow \mathrm{IDs}(K)$
  Fetch paper context: $C \leftarrow \mathrm{Fetch}(D, I_K)$
  Generate comparison tables aligned to K: $B \leftarrow \mathrm{Tab}(K, C)$
  Phase 3: Tree+Table Guided Outline and Survey
  for $k = 1$ to m **do**
    Propose an outline with tree & tables: $O_k \leftarrow \mathrm{Out}(T, K, B)$
  end for
  Select the outline: $O \leftarrow \mathrm{Judge}(\{O_k\}_{k=1}^{m})$
  Write the survey conditioned on structures: $S \leftarrow \mathrm{Write}(T, O, K, B)$
  Align and revise for cross-view consistency: $(K, B, O, S) \leftarrow \mathrm{Align}(K, B, O, S)$
  return (K, B, S)

3 Method

In this section, we describe the methodology of MVSS, a multi-view structured framework for automated survey generation. MVSS formulates survey synthesis as a joint structural generation problem, where multiple representations are constructed and aligned in a coordinated manner. Specifically, MVSS proceeds through three structured stages: (1) evidence-grounded hierarchical knowledge tree construction, (2) tree-induced structured table generation, and (3) tree- and table-guided outline and survey text generation. Each stage is designed to address a key challenge in automated survey creation, including conceptual organization, comparative analysis, and evidence-consistent writing. Figure 2 provides an overview of the complete workflow, and the overall procedure is summarized in Algorithm 1.

3.1 Evidence-Grounded Tree (HKT)

Given a survey topic T and a paper database D, MVSS first constructs a Hierarchical Knowledge Tree (HKT) to explicitly model the conceptual organization of the target domain. We begin by retrieving an initial set of candidate papers:

$$P \leftarrow \mathrm{Ret}(T, D),$$

which serves as the evidence pool for subsequent structure induction.

Using the topic T and retrieved papers P, we construct an initial tree:

$$K \leftarrow \mathrm{Tree}(T, P),$$

where each node corresponds to a research concept or subtopic and is explicitly associated with a set of supporting papers. The resulting tree provides a structured abstraction of the domain, capturing major branches and their conceptual relations.

To improve structural stability and evidence consistency, we further apply an iterative refinement procedure:

$$K \leftarrow \mathrm{Ref}(K),$$

which adjusts the hierarchy by resolving redundant nodes, correcting parent–child relations, and reassigning supporting papers when necessary. After R refinement rounds, the tree K serves as a citation-grounded conceptual backbone for subsequent stages.

3.2 Tree-Induced Table Generation

Based on the refined hierarchical knowledge tree K, MVSS generates structured comparison tables to explicitly expose discriminative dimensions among methods and subfields. We first extract all paper identifiers associated with tree nodes:

$$I_K \leftarrow \mathrm{IDs}(K),$$

and retrieve their corresponding contextual information from the database:

$$C \leftarrow \mathrm{Fetch}(D, I_K),$$

including titles, abstracts, and other descriptive metadata.

Table generation is then formulated as a conditional mapping:

$$B = \mathrm{Tab}(K, C),$$

where the tree structure K determines the semantic scope and placement of each table, and the paper context C provides evidence for populating table entries. Each table is aligned with a specific node or subtree, ensuring that comparison rows and columns operate at a consistent conceptual level. As a result, the generated tables present structured, evidence-grounded comparisons that are coherent with the global survey organization.

3.3 Tree+Table Guided Outline and Survey

In the final stage, MVSS jointly leverages the hierarchical tree K and comparison tables B to guide outline construction and survey writing. Given the topic and structural representations, we generate multiple outline candidates:

$$O_k \leftarrow \mathrm{Out}(T, K, B), \quad k = 1, \ldots, m,$$

and select them using a judge model:

$$O \leftarrow \mathrm{Judge}(\{O_k\}_{k=1}^{m}).$$

Conditioned on the final outline and structural constraints, the survey text is generated as:

$$S = \mathrm{Write}(T, O, K, B),$$

where the outline controls section ordering, the tree enforces conceptual hierarchy, and the tables guide comparative analysis and evidence usage.

To ensure consistency across multiple representations, we apply a cross-view alignment operation:

$$(K, B, O, S) \leftarrow \mathrm{Align}(K, B, O, S),$$

which detects and revises structural conflicts, missing coverage, or inconsistent citations among the tree, tables, outline, and text. This alignment step yields a coherent, multi-view survey with consistent structure and evidence grounding.

4 Experiments

4.1 Experimental Setup

We conduct comprehensive experiments to evaluate MVSS and its hierarchical knowledge tree (HKT) module from three perspectives: (1) quality of the generated knowledge trees, (2) quality of the structured comparison tables, and (3) overall survey quality when trees and tables are used jointly to guide text generation.

Dataset and Corpus. We evaluate on 76 computer science topics spanning machine learning, natural language processing, computer vision, and systems research. The retrieval corpus contains 530,000 arXiv papers (2018–2024), preprocessed following standard practices for scientific document retrieval. We use **deepseek-chat** as the primary generator to produce MVSS outputs.

LLM-as-judge evaluation. Following recent work on automatic evaluation of long-form generation, we rely on calibrated LLM judges to score trees, tables, and surveys. All prompts are aligned with human-written guidelines, and a small expert-annotated set is used for scale calibration, echoing findings that large language models can serve as reliable automatic evaluators when appropriately designed and validated against human judgments (Kocmi and Federmann, 2023; Fabbri et al., 2021).

4.2 Metrics

We develop a new evaluation scheme tailored to multi-view structured survey generation. It includes three dimensions for surveys (**Coverage**, **Structure**, **Relevance**), together with a single holistic criterion for trees (**TreeQuality**) and for tables (**TableQuality**). All criteria are rated on a 1–5 Likert scale by calibrated LLM judges.

Survey quality. We evaluate surveys along three 5-point dimensions summarized in Table 1. Coverage assesses whether the survey covers key and peripheral aspects comprehensively. Structure evaluates logical organization, coherence, and non-redundant flow. Relevance measures alignment with the target topic and focus with minimal digressions. We define the overall survey score as $Q_{\text{survey}} = \frac{1}{3}(S_{\text{cov}} + S_{\text{str}} + S_{\text{rel}})$.

Tree quality. TreeQuality evaluates the taxonomy/topic tree quality, including hierarchy correct-

Table 1: New evaluation criteria for MVSS. All dimensions use a 1–5 Likert scale. For brevity, we show representative anchors (1/5).

Criterion	Anchors of 1–5 scale
Coverage	1: Very limited coverage; misses most key areas. 5: Fully comprehensive; covers key and peripheral topics in depth.
Structure	1: No clear logic or connections between sections. 5: Tightly structured, clear logic, smooth transitions, no redundancy.
Relevance	1: Outdated/unrelated; not aligned with the topic. 5: Exceptionally focused; every detail supports understanding of the topic.
TreeQuality	1: No meaningful tree or totally wrong hierarchy. 5: Excellent tree: comprehensive, correct, clear grouping, useful abstraction.
TableQuality	1: No usable table or incorrect/misleading. 5: Excellent tables: comprehensive comparisons with consistent formatting.

ness, coverage of major branches, and clarity of grouping. We set $Q_{\text{tree}} = S_{\text{tq}}$.

Table quality. TableQuality evaluates the usefulness of comparison tables, focusing on correctness, completeness, consistency, and practical utility for comparison. We set $Q_{\text{table}} = S_{\text{tab}}$.

Citation quality for trees and surveys. Following scientific fact verification, we extract a set of claims $C = \{c_i\}_{i=1}^{m}$ and model-proposed (claim, reference) pairs $P = \{(c_i, r_j)\}$. An NLI model $V(c, r) \in \{0, 1\}$ returns 1 if r supports c. Citation recall and precision are

$$\text{Rec}_{\text{cite}} = \frac{|\{c \in C : \exists r, (c, r) \in P, V(c, r) = 1\}|}{|C|}, \tag{1}$$

$$\text{Prec}_{\text{cite}} = \frac{|\{(c, r) \in P : V(c, r) = 1\}|}{|P|}. \tag{2}$$

4.3 Baselines

We compare MVSS against representative human and automatic survey-generation baselines:

- **Human Experts.** Expert-written surveys with manually curated hierarchies, used as an upper bound.
- **Naive RAG-based LLM Generation.** This baseline follows a simple retrieve-and-generate paradigm. Given a topic T and a target survey length, we first retrieve a set of relevant papers and then directly prompt an LLM to generate a linear survey text until completion, without any explicit tree, table, or outline planning.
- **AutoSurvey.** An automatic survey generator that produces an outline and expands it into text with sentence-level citations.
- **HiReview.** A hierarchical review system that retrieves a fixed set of papers and generates taxonomy-guided survey text.

4.4 Main Results

We now evaluate MVSS as a full survey generator and analyze how trees and tables affect the resulting text.

The key findings from our end-to-end survey evaluation are summarized as follows.

- **MVSS consistently achieves the highest overall survey quality across all target lengths.** As shown in Table 2, MVSS outperforms all automatic baselines under every length setting from 8k to 64k tokens. At 16k tokens, MVSS attains an average score of 4.90, exceeding AutoSurvey (4.60) and HiReview (3.65). This performance gap further widens at longer lengths, indicating that MVSS frequently approaches, and in some cases matches, expert-level organization under direct human comparison.
- **At long-context settings, MVSS matches expert-written surveys in structure and relevance.** Under 32k and 64k token budgets, MVSS achieves near-perfect or perfect scores on all three dimensions. Notably, at 64k tokens, MVSS reaches 5.00 on Coverage, Structure, and Relevance, matching human-written surveys under the same evaluation protocol. This result highlights the effectiveness of structure-first planning and cross-view alignment for long-form survey synthesis.
- **Structural organization is the primary source of MVSSs advantage over baselines.** Across all lengths, MVSS consistently yields the highest Structure scores, with especially large margins over naive RAG-based generation and HiReview. For example, at 16k tokens, MVSS improves Structure by +0.52 over AutoSurvey and +1.85 over HiReview. These gains demonstrate that explicitly modeling and enforcing hierarchical and comparative structures is critical for producing high-quality surveys.

Table 2: Survey quality under different target lengths (#tokens). We report content quality on a 1–5 Likert scale: Coverage, Structure, Relevance, and their average.

Survey Length	Methods	Coverage	Structure	Relevance	Avg.
8k	Human writing	4.50	4.16	5.00	4.52
	Naive RAG-based LLM generation	4.40±0.48	3.86±0.71	4.86±0.33	4.33
	AutoSurvey	4.60±0.48	4.46±0.49	4.80±0.39	4.61
	HiReview	3.67±0.47	3.00	4.00	3.56
	MVSS (ours)	**4.83±0.39**	**4.79±0.35**	**4.88±0.19**	**4.83**
16k	Human writing	4.66	4.38	5.00	4.66
	Naive RAG-based LLM generation	4.46±0.49	3.66±0.69	4.73±0.44	4.23
	AutoSurvey	4.66±0.47	4.33±0.59	4.86±0.33	4.60
	HiReview	3.94±0.24	3.00	4.00	3.65
	MVSS (ours)	**4.94±0.35**	**4.85±0.48**	**4.92±0.29**	**4.90**
32k	Human writing	4.66	4.50	5.00	4.71
	Naive RAG-based LLM generation	4.41±0.64	3.75±0.72	4.66±0.47	4.23
	AutoSurvey	4.73±0.44	4.26±0.69	4.80±0.54	4.58
	HiReview	4.88±0.33	4.56±0.50	5.00	4.81
	MVSS (ours)	**4.91±0.24**	**4.93±0.33**	**5.00**	**4.95**
64k	Human writing	5.00	4.66	5.00	4.88
	Naive RAG-based LLM generation	4.46±0.61	3.66±0.47	4.66±0.47	4.19
	AutoSurvey	4.73±0.44	4.33±0.47	4.86±0.33	4.62
	HiReview	4.70±0.46	4.40±0.49	5.00	4.70
	MVSS (ours)	**5.00**	**5.00**	**5.00**	**5.00**

4.5 Human Evaluation

To complement LLM-based evaluation with human-grounded evidence, we conduct a human study on 30 randomly sampled topics in the form of a *double-blind pairwise preference evaluation*.

Double-blind pairwise evaluation. Annotators are shown two anonymized system outputs (A/B) in random order and asked to choose which one is better, or mark a tie, along three dimensions: (1) *Coverage*, (2) *Structure*, and (3) *Relevance*. This protocol directly measures relative superiority while mitigating scale bias.

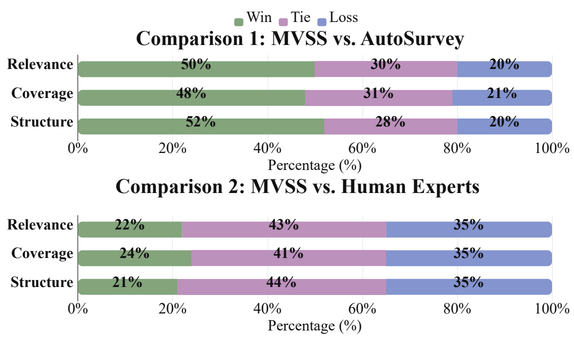

Figure 3: Double-blind pairwise human evaluation results. The figure shows win/tie/loss counts for MVSS compared with AutoSurvey and human-written surveys.

Figure 3 shows that MVSS is preferred over AutoSurvey across dimensions. When compared to human-written surveys, MVSS is selected in a non-trivial fraction of cases with a substantial tie rate, indicating that MVSS often approaches expert-level organization in pairwise judgments. This result suggests that MVSS not only improves over automated baselines, but can also produce structures that human evaluators frequently find comparable to expert-written surveys, particularly in terms of structural clarity and organization.

4.6 Ablation Studies

We analyze the role of major design choices in MVSS through targeted ablations to understand how each component contributes to the overall performance. Specifically, we examine the effects of removing hierarchical tree guidance, disabling iterative tree refinement, and replacing multi-model outline consensus with a single-model outline. These ablations isolate the impact of structural planning, refinement dynamics, and cross-model agreement on both survey quality and citation reliability. The complete MVSS system serves as the reference point, enabling a controlled comparison that highlights which design choices are critical for achieving robust structure and faithful evidence grounding.

Results and discussion. Table 3 shows that removing tree guidance causes the largest degradation in structural quality and citation precision, highlighting the central role of hierarchical planning in organizing survey content and anchoring

Table 3: Ablation study results for MVSS with different components removed.

Variant	Cov	Str	Rel	Q_{survey}	Rec(%)	Prec(%)	TreeQ	TableQ
MVSS	**4.12±0.18**	**3.35±0.21**	3.92±0.19	**3.88±0.16**	82.31±5	**76.94±4**	3.85±0.31	**3.77±0.43**
MVSS w/o tree generation	4.00±0.38	3.20±0.41	3.87±0.35	3.69±0.38	**82.68±5**	76.70±6	1	3.67±0.49
MVSS w/o tree refinement	4.10±0.29	3.10±0.29	3.76±0.43	3.65±0.22	80.26±6	74.58±6	3.76±0.53	3.57±0.51
MVSS w/o table generation	3.95±0.21	3.18±0.39	3.91±0.29	3.68±0.29	82.06±6	76.66±6	**3.91±0.29**	1
MVSS w/o multi-model outline	4.05±0.23	3.32±0.48	**3.95±0.23**	3.77±0.31	79.01±7	72.68±6	3.75±0.54	3.68±0.48

evidence. Disabling iterative tree refinement and multi-model outline consensus further degrades performance, suggesting that both iterative structure correction and cross-model agreement are important for producing stable hierarchies and consistent survey organization. These results indicate that MVSSs performance gains arise from the interaction of multiple structural components rather than any single design choice.

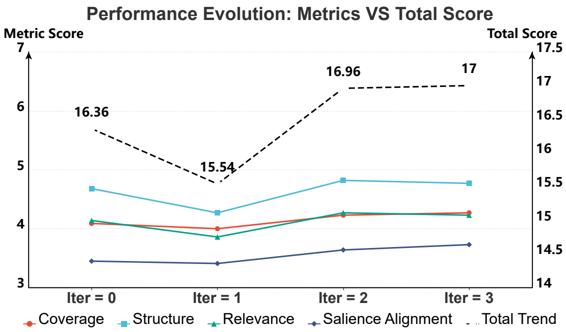

Figure 4: Performance evolution across iterative tree refinement rounds. Iter 0 corresponds to the initial tree without refinement.

Figure 4 illustrates how survey quality evolves across iterative tree refinement. The initial iteration (Iter 0) starts from a coarse hierarchy and yields moderate scores across all dimensions. After the first refinement step (Iter 1), performance temporarily dips, reflecting structural reorganization and pruning of noisy branches. Subsequent iterations consistently improve Coverage, Structure, and Relevance, leading to a monotonic increase in the overall score. This trend indicates that iterative refinement stabilizes the hierarchical organization and progressively aligns structural planning with downstream survey generation, validating the effectiveness of our refinement strategy *under realistic long-context generation settings*.

Table 4: MVSS evaluation under different LLM judges.

Metric	LLM Judge		
	deepseek-chat	gpt-4o	gemini-2.5-pro
Rec (%)	86.13±4.52	88.89±4.34	78.81±6.75
Prec (%)	80.72±4.98	83.82±4.89	74.35±6.39
Coverage	5.00	5.00	5.00
Structure	5.00	5.00	4.78±0.42
Relevance	5.00	5.00	5.00
TreeQ	4.78±0.42	4.96±0.21	4.35±0.93
TableQ	4.17±0.39	5.00	4.87±0.34
Avg	4.79±0.11	4.99±0.04	4.80±0.23

Across different LLM judges, MVSS consistently achieves strong performance, with high Coverage and Relevance scores and moderate variance across metrics, indicating robust behavior under diverse evaluation models.

5 Conclusion

We presented **MVSS**, a unified framework for *multi-view structured survey generation* that elevates conceptual structure from a secondary byproduct to a first-class optimization objective. By jointly constructing citation-grounded hierarchical knowledge trees, schema-driven comparison tables, and evidence-aware narrative text, MVSS enforces structural coherence and semantic alignment across survey views. Through multi-model outline consensus and dual-objective structural refinement, MVSS consistently improves structural clarity, comparative insight, and citation fidelity across 76 diverse topics, approaching expert-written quality while remaining orders of magnitude more efficient.

Beyond its empirical gains, MVSS reframes automated survey generation as a *structure-centric synthesis* problem, highlighting the role of explicit hierarchies and comparisons in literature understanding. We believe MVSS represents a step toward scalable systems that go beyond summarization to actively organize scientific knowledge.

6 Limitations

The system still depends on frontier LLMs for both generation and judgment, which raises cost, reproducibility, and bias concerns, especially when extending to domains underrepresented in pretraining data. Our structural and alignment objectives introduce additional hyperparameters whose robustness across domains and retrieval settings has not been fully characterized. Moreover, our evaluation focuses on 76 CS topics using an arXiv-based corpus, limiting generalizability to other disciplines, formats, or argumentative norms. Finally, MVSS models a static snapshot of a field and does not capture temporal evolution or uncertainty in conflicting evidence, suggesting opportunities for incorporating time-aware or uncertainty-aware structures in future work.

References

Nurshat Fateh Ali, Md Mahdi Mohtasim, Shakil Mosharrof, and T. Gopi Krishna. 2024. Automated literature review using nlp techniques and llm-based retrieval-augmented generation. In *2024 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, pages 1–6. IEEE.

Parikshit Bansal and Amit Sharma. 2023. Large language models as annotators: Enhancing generalization of nlp models at minimal cost. *arXiv preprint*, arXiv:2306.15766.

Tong Bao, Mir Tafseer Nayeem, Davood Rafiei, and Chengzhi Zhang. 2025. SurveyGen: Quality-aware scientific survey generation with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2712–2736, Suzhou, China. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the EMNLP-IJCNLP*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Savita Bhat and Vasudeva Varma. 2023. Large language models as annotators: A preliminary evaluation for annotating low-resource language content. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 100–107, Bali, Indonesia. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Asli Celikyilmaz, Dilek Hakkani-Tür, and Gokhan Tur. 2010. A hybrid hierarchical model for multi-document summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 815–824, Uppsala, Sweden. Association for Computational Linguistics.

Janara Christensen, Oren Etzioni, and 1 others. 2014. Towards coherent multi-document summarization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, USA. Association for Computational Linguistics.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Maxime Darrin, Ines Arous, Pablo Piantanida, and Jackie Chi Kit Cheung. 2024. GLIMPSE: Pragmatically informative multi-document summarization for scholarly reviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Bangkok, Thailand. Association for Computational Linguistics.

Google DeepMind. 2023. Gemini: A family of highly capable multimodal models. Technical report, Google. Technical report.

Tanay Dixit, Fei Wang, and Muhao Chen. 2023. Improving factuality of abstractive summarization without sacrificing summary quality. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, Toronto, Canada. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. In *Transactions of the Association for Computational Linguistics*, volume 9, pages 391–409. MIT Press.

Jinlan Fu and 1 others. 2024. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Mexico City, Mexico. Association for Computational Linguistics.

Luyu Gao, Zhuyun Dai, Zhiyu Chen, Angela Fan, Xilun Ma, and Danqi Chen. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Augusto Gonzalez Bonorino. 2023. Smart surveys: An automatic survey generation and analysis tool. In *Proceedings of the 15th International Conference on Computer Supported Education (CSEDU), Volume 2*, pages 113–119. SciTePress.

Xingyu Hu, Haoyu Zhang, Yuxuan Li, Hongshen Xu, and Dongyan Wang. 2025. HiReview: Hierarchical taxonomy-driven automatic literature review generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*. To appear.

Chung-Wei Huang, Hong-Jie Dai, Tomoya Imai, and Ying-Jui Ko. 2020. AutoSurvey: Automatic survey generation based on a research draft. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI) – Demonstrations Track*, pages 280–286.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 874–880, Online. Association for Computational Linguistics.

Fumiya Kasanishi, Shinsuke Mori, and Ichiro Sakata. 2023. Scireviewgen: A dataset for scientific literature review generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11602–11617, Toronto, Canada. Association for Computational Linguistics.

Pranav Kasela, Gabriella Pasi, and Raffaele Perego. 2025. PARK: Personalized academic retrieval with knowledge-graphs. *Information Systems*, 134:102574.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Miao Li, Eduard H. Hovy, and Jey Han Lau. 2023. Towards summarizing multiple documents with hierarchical relationships. *CoRR*, abs/2305.01498.

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5070–5081, Florence, Italy. Association for Computational Linguistics.

Yang Liu and 1 others. 2021. Highlight-transformer: Leveraging key phrase aware attention to improve abstractive multi-document summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

Aman Madaan, Amir Yazdanbakhsh, Gerard de Melo, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. In *Proceedings of the 2019 Conference on Neural Information Retrieval (NeurIPS IR Workshop)*.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

OpenAI. 2024. Deep research: An ai system for multi-step research and knowledge synthesis. Technical report, OpenAI. Technical report.

Jason Portenoy and Jevin D. West. 2020. Constructing and evaluating automated literature review systems. *Scientometrics*, 125(3):3233–3251.

Ruihua Qi, Weilong Li, and Haobo Lyu. 2025. Generation of scientific literature surveys based on large language models (llm) and multi-agent systems (mas). In *Natural Language Processing and Chinese Computing (NLPCC 2024)*, volume 15363 of *Lecture Notes in Computer Science*, pages 169–180. Springer.

Noah Shinn, Federico Cassano, Dorsa Gopinath, and 1 others. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ludo Waltman and Nees Jan van Eck. 2012. A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, 63(12):2378–2392.

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, Wei Ye, Shikun Zhang, and Yue Zhang. 2024. AutoSurvey: Large language models can automatically write surveys. In *Advances in Neural Information Processing Systems*, volume 37. NeurIPS 2024, Main Conference Track.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander Fabbri, Shulamit Ishiwatari, Zhehao Li, and Dragomir R. Radev. 2019. Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7386–7393.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Xin Zhang, Qiyi Wei, Qing Song, and Pengzhou Zhang. 2024. TOMDS (topic-oriented multi-document summarization): Enabling personalized customization of multi-document summaries. *Applied Sciences*, 14(5):1880.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Wenxuan Hou, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Appendix

Hyperparameters and Implementation Details

Implementation details. Unless otherwise stated, we retrieve 1,200 topic-relevant papers from the arXiv corpus and use abstracts for outline generation. For MVSS, 60 papers are retrieved per tree node or table row. To ensure a fair comparison, HiReview is configured to retrieve top-k=60 papers per topic, matching MVSSs per-subsection retrieval budget. All methods operate on the same retrieval corpus and use the same backbone LLM for generation. The reflection loop in the hierarchical knowledge tree (HKT) module runs for three iterations, selecting the best candidate based on TreeQuality and citation metrics. Temperature is fixed to 1.0 for all LLM calls. Remaining hyperparameters and prompt configurations are provided in Appendix 6.

Additional Experimental Results

Section Granularity Analysis

Table 5 shows that content quality improves from 3 to 4 sections and saturates at 45 sections.

Table 5: Impact of section numbers on HKT performance.

Sections	Cov	Str	Rel	Sal	Avg
3	4.27	4.82	4.55	4.00	**4.41**
4	4.33	5.00	4.72	4.06	**4.53**
5	4.50	4.70	4.70	4.10	**4.50**
6	4.25	5.00	4.75	4.00	**4.50**

Cost and Runtime Breakdown

We compare the computational and monetary costs of MVSS with those of baseline systems. Table 6 reports per-topic API cost (estimated using public pricing) and end-to-end wall-clock time.

Table 6: Cost analysis across systems.

Method	API Cost ($)	Time (min)
AutoSurvey	0.31	10.94
HKT (ours)	0.55	11.3
MVSS (ours)	0.94	30.48

MVSS incurs moderately higher computational cost than pipeline-style baselines due to cross-view refinement, but the gains in coverage, structure, and relevance suggest a favorable accuracy-cost trade-off for high-stakes survey generation.

LLM–Human Agreement and Significance Analysis

To evaluate the reliability of LLM-as-judge scores, we compute Pearson and Spearman correlations between LLM ratings and human Likert-scale ratings over 30 randomly sampled topics. Table 7 summarizes the alignment across dimensions.

Table 7: Correlation between LLM scores and human ratings.

Score Pair	Pearson r	Spearman ρ
Q_{survey} vs. Human Overall	0.61	0.57
LLM Cov vs. Human Cov	0.54	0.55
LLM Str vs. Human Str	0.62	0.58
LLM Rel vs. Human Rel	0.54	0.52

Table 8: Paired t-test significance of MVSS vs. baselines (two-tailed, paired).

Comparison	t-stat	p-value
MVSS vs AutoSurvey	18.97	$< 10^{-16}$

These results confirm that LLM judgments track human evaluations and that MVSSs improvements over baselines are statistically reliable.

Additional Tree-Level Results

Table 9: Tree-level comparison between Naive RAG-based LLM generation and HKT (MVSS).

Method	Cov	Str	Rel	Sal	Avg
Naive Generation	3.90	4.37	4.53	2.50	3.83
HKT (MVSS)	**4.50**	**4.80**	**4.30**	**3.60**	**4.30**

Prompt Templates

Table 10: Key prompts used in the MVSS system.

`CRITERIA_BASED_JUDGING_PROMPT`	Given an academic survey and specific criteria with Score 1–5 descriptions, evaluate the survey quality. Return the score only.
`NLI_PROMPT`	Given a Claim and a Source, determine if the Claim is faithful to the Source. Return only `Yes` or `No`.
`ROUGH_OUTLINE_PROMPT`	Given `[PAPER LIST]`, `[PRIOR KNOWLEDGE MD]`, and `[PRIOR KNOWLEDGE JSON]`, draft a comprehensive outline with `[SECTION NUM]` sections.
`MERGING_OUTLINE_PROMPT`	Given multiple outline candidates `[OUTLINE LIST]`, merge them into a single, logical, and comprehensive final outline.
`SUBSECTION_OUTLINE_PROMPT`	Given an overall outline, prior knowledge, and a specific section description, generate structural subsections using `[PAPER LIST]`.
`EDIT_FINAL_OUTLINE_PROMPT`	Refine a draft outline containing sections and subsections to remove duplicates and improve logical coherence. Return in LaTeX-style format.
`CHECK_CITATION_PROMPT`	Verify whether citations in a written subsection are supported by the corresponding papers in `[PAPER LIST]`. Fix incorrect citations or remove them.
`SUBSECTION_WRITING_PROMPT`	Write content (> `[WORD NUM]` words) for a specific subsection. Cite papers using only `[Title]` format. Strict constraints: do not repeat subsection titles and do not output prior knowledge trees.
`LOCAL_TABLE_REFLECT_PROMPT`	Analyze the written subsection and source papers. If ≥ 3 distinct methods are discussed, generate a comparison table in raw Markdown using exact paper titles.

Reference Survey Papers

From Google Scholar, we selected 76 influential surveys on diverse LLM topics, balancing citation counts and coverage.

Table 11: Survey papers used for evaluation (top 20 by number of references).

Topic	**Survey Title**	**Refs**
LLM Agents	The Rise and Potential of Large Language Model Based Agents: A Survey	674
Deep RL for Vision	Deep Reinforcement Learning in Computer Vision: A Comprehensive Survey	432
Vision Foundation Models	Foundational Models Defining a New Era in Vision: A Survey and Outlook	359
GNNs in IoT	Graph Neural Networks in IoT: A Survey	333
LLM Evaluation	A Survey on Evaluation of Large Language Models	269
RL/IL for Auto. Driving	A Survey of Deep RL and IL for Autonomous Driving Policy Learning	268
Blockchain & AI for 6G	A Survey of Blockchain and Artificial Intelligence for 6G Wireless Communications	264
Diffusion Models	A Survey on Generative Diffusion Models	258
PTMs in NLP	Pre-trained Models for Natural Language Processing: A Survey	249
PHY Security (Industry)	A Survey of Physical Layer Techniques for Secure Wireless Communications in Industry	248
KG Embeddings	Knowledge Graph Embedding: A Survey of Approaches and Applications	239
GNNs in RecSys	Graph Neural Networks in Recommender Systems: A Survey	231
PLMs for Text Gen.	Pre-trained Language Models for Text Generation: A Survey	226
Vehicular Network Sec.	Machine Learning for Security in Vehicular Networks: A Comprehensive Survey	224
Prompt Learning	Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in NLP	223
Text-to-SQL	A Survey of Text-to-SQL in the Era of LLMs: Where are We, and Where are We Going?	217
Hyperspectral Super-Res.	Hyperspectral Image Super-Resolution Meets Deep Learning: A Survey and Perspective	213
Federated Analytics	A Survey on Federated Analytics: Taxonomy, Enabling Techniques, Applications and Open Issues	202
Motion Planning	Motion Planning for Autonomous Driving: The State of the Art and Future Perspectives	182
RLHF / Human Feedback	Bridging the Gap: A Survey on Integrating (Human) Feedback for Natural Language Generation	153