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

Anonymous EMNLP submission

Abstract

Scientific surveys demand not just summarization but the organization of literature into coherent conceptual structures. Existing automatic survey generation methods, however, focus on linear text generation and largely ignore hierarchical relations and structured methodological comparisons, leaving substantial gaps to expert-written surveys. Evaluation has also been confined to ad-hoc CS topic lists, leaving cross-domain generalization untested. We address these limitations along two complementary lines. First, we propose **MVSS**, a constraint-driven multi-view survey generation framework: it constructs a citation-grounded Hierarchical Knowledge Tree (HKT) that serves as a strict anchor for tree-induced comparison tables, outline planning, and narrative writing, followed by a cross-view alignment step that resolves coverage, table-narrative binding, citation, and traversal inconsistencies. Second, we construct **MVS-Bench**, the first cross-domain benchmark for automated survey generation, comprising 100 expert-curated topics across computer science, economics, electrical engineering, and systems science, paired with high-impact reference surveys and a multi-signal evaluation suite. On MVS-Bench, MVSS significantly outperforms existing methods in survey organization and evidence grounding across all four disciplines, and substantially narrows the gap to expert-written surveys. We open our resources at `https://anonymous.4open.science/r/MVSS-824F`.

1 Introduction

In rapidly evolving areas of natural language processing—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 emerge within months, while comprehensive surveys lag behind by one or more publication cycles. This gap has motivated a growing body of LLM-based automated survey generation systems, including academic agents such as AutoSurvey (Wang et al., 2024), SurveyGen (Bao et al., 2025), SurveyX (Liang et al., 2025), SurveyForge (Yan et al., 2025), SurveyG (Nguyen et al., 2025), InteractiveSurvey (Wen et al., 2025), and LLMxMapReduce (Chao et al., 2025; Qi et al., 2025; Ali et al., 2024), as well as industrial research assistants (OpenAI, 2024; DeepMind, 2023). However, most of these systems produce surveys as linear narratives, overlooking the citation trees and citation tables that are central to human-written surveys.

Human-written surveys organize literature through two tightly coupled structures. A *citation tree* is a hierarchy whose nodes correspond to research concepts or sub-areas—each linked to representative supporting papers—and progressively decompose a topic from high-level themes into finer-grained directions. A *citation table*, in turn, is a tabular view whose rows are methods or research lines, columns are key comparison dimensions, and entries are explicitly linked to cited evidence. Their prevalence is striking: over 80% of LLM surveys published in 2021–2025 contain explicit hierarchical trees, and a 30-participant pilot of CS graduate students shows that 29/30 (96.7%) consult a hierarchy/figure or a comparison table before reading raw prose (Appendix D.6). Yet existing methods exploit them poorly. HiReview (Hu et al., 2024), the most relevant prior work, builds citation trees from title-based clustering, which misses papers whose relevance is not evident in their titles and yields unstable, noisy hierarchies; it further assesses tree quality only indirectly via downstream text quality, and does not model tables at all.

To address these limitations, we propose **Multi-View Structured Survey (MVSS)**, a framework

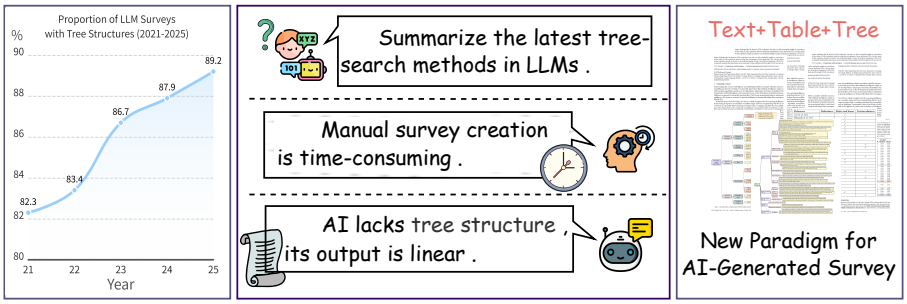

Figure 1: **Motivation for structure-first survey generation.** Linear generation obscures conceptual relations, whereas our structure-first paradigm aligns multi-view representations. Empirically, hierarchical trees are an established practice in expert-written LLM surveys: across the top-100-cited LLM-related surveys per year (2021–2025), the proportion containing an explicit hierarchical taxonomy is consistently above 80%.

that formulates survey generation as a *constraint-driven synthesis* problem rather than a linear text pipeline. At its core is the **Hierarchical Knowledge Tree (HKT)**, a citation-grounded representation that externalizes the latent conceptual hierarchy of a research domain. Conditioned on the HKT, a structure-aware mechanism generates comparison tables aligned with the tree across concepts, dimensions, and supporting evidence; both views then jointly constrain outline construction and survey text generation. MVSS is therefore not a module assembly but a reformulation of survey generation as *cross-view structural alignment*.

To rigorously evaluate this paradigm beyond the ad-hoc CS topic lists used in prior work, we further construct **MVS-Bench**, a cross-domain benchmark of 100 expert-curated survey topics spanning four disciplines, paired with high-impact reference surveys, unified retrieval corpora, and a four-annotator consensus protocol. On MVS-Bench, MVSS consistently achieves the strongest overall performance across three LLM judges (GPT-4o, Gemini-2.5-Pro, DeepSeek-chat), with a substantial lead in structural organization and relevance, citation precision and recall both exceeding 75%, and gains that transfer to all three non-CS disciplines. While concurrent work such as DeepSurvey-Bench (Zhang et al., 2026) evaluates automated surveys using academic-value-oriented metrics on linear text, MVS-Bench is complementary in its explicit focus on multi-view structural fidelity and cross-domain generalization.

Our contributions are:

- **Method.** We introduce **MVSS**, a constraint-driven survey generation framework whose core is a citation-grounded **Hierarchical Knowledge Tree (HKT)** used as a strict anchor to constrain comparison-table generation, outline planning, and narrative writing.
- **Alignment.** We design $\mathrm{Align}(K, B, O, S)$, a four-way cross-view consistency operator that resolves coverage gaps, table–narrative binding violations, citation hallucinations, and tree–outline traversal inconsistencies after initial generation.
- **Benchmark.** We release **MVS-Bench**, the first automated-survey benchmark featuring cross-domain coverage, paired expert references, and cross-system evidence metrics .

2 Related Work

2.1 Automatic Survey Generation

Automated survey generation (Portenoy and West, 2020; Kasanishi et al., 2023; Darrin et al., 2024; Gonzalez Bonorino, 2023) is mostly framed as multi-document summarization producing linear narratives (Christensen et al., 2014; Celikyilmaz et al., 2010; Liu and Lapata, 2019; Yasunaga et al., 2019; Li et al., 2023; Zhang et al., 2024). Recent LLM systems add structural biases (Liu et al., 2021), retrieval augmentation (Izacard and Grave, 2021; Nogueira and Cho, 2019), and citation-aware critique (Kryściński et al., 2020; Dixit et al., 2023; Madaan et al., 2023; Shinn et al., 2023), yet remain fundamentally *text-centric*, treating hierarchies as byproducts. Existing benchmarks—Multi-XScience (Lu et al., 2020), BigSurvey (Liu et al., 2022), SciReviewGen (Kasanishi et al., 2023), SumSurvey (Liu et al., 2024) and SurveySum (Fernandes et al., 2024) reinforce this limitation: they are restricted to computer science, rely on reference abstracts, and lack paired expert surveys or cross-domain testing. MVSS adopts a *structure-*

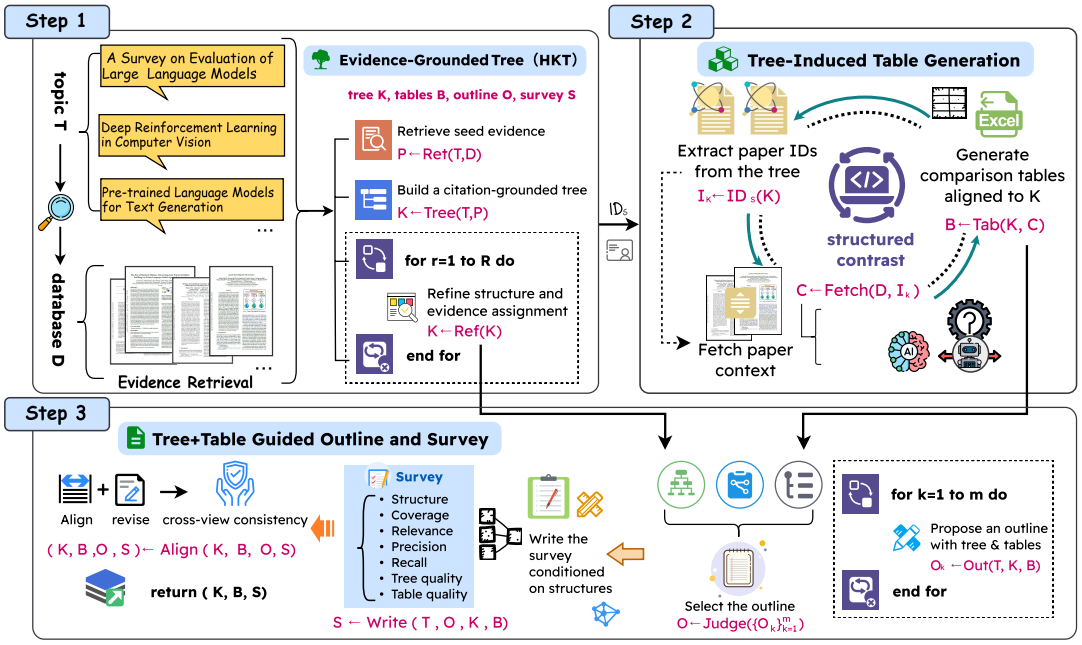

Figure 2: **Overview of MVSS.** Given a topic T and a paper database D, MVSS constructs an evidence-grounded hierarchical tree, generates aligned comparison tables, and produces a structured survey via cross-view alignment.

first perspective, treating hierarchical trees, tables, and cross-view alignment as explicit objectives, evaluated on **MVS-Bench**, the first four-discipline benchmark with paired expert references.

2.2 Knowledge Structuring and Evaluation

Prior knowledge organization via document embeddings (Beltagy et al., 2019; Cohan et al., 2020), graph retrieval (Kasela et al., 2025), and bibliometric clustering (Waltman and van Eck, 2012) yields structures too coarse for fine-grained surveys. Recent LLM taxonomies (Hsu et al., 2024; Kargupta et al., 2025; Zhu et al., 2025) are primarily standalone exploration tools, whereas our Hierarchical Knowledge Tree (HKT) acts as a strict anchor dynamically constraining table and narrative generation for cross-view consistency. Existing multi-signal evaluations target text quality in CS by leveraging techniques such as reflection (Madaan et al., 2023), LLM-as-a-judge (Fu et al., 2024; Bhat and Varma, 2023; Bansal and Sharma, 2023), fact verification (Kryściński et al., 2020; Dixit et al., 2023), and reranking (Nogueira and Cho, 2019). However, these approaches leave structural fidelity and cross-domain metrics underexplored.**MVSS addresses this:** methodologically, by optimizing structure to form coherent, grounded representations across trees, tables, and text; and via **MVS-Bench**, providing a unified multi-signal protocol across four disciplines with paired expert surveys. Concurrently, DeepSurvey-Bench (Zhang et al., 2026) proposes academic-value-oriented metrics across information, scholarly communication, and research guidance dimensions to evaluate generated surveys; while complementary in spirit to our work, their evaluation operates on linear surveys and does not assess cross-view structural alignment between trees, tables, and narrative. MVS-Bench instead targets multi-view structural fidelity and provides paired expert references with cross-domain coverage.

3 Method

In this section, we describe MVSS, a multi-view structured framework for automated survey generation. Unlike pipeline approaches that treat structure as a by-product of writing, MVSS formulates survey synthesis as a *joint structural generation* problem: a hierarchical tree, comparison tables, an outline, and the final text are constructed as mutually constrained views and aligned in a coordinated manner. Concretely, 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 addressing a key challenge in automated survey creation—conceptual

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)

organization, comparative analysis, and evidence-consistent writing, respectively. Figure 2 provides an overview of the 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 retrieve an initial evidence pool $P \leftarrow \mathrm{Ret}(T, D)$ and induce an initial tree

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

where the Tree operator elicits concept nodes top-down via an LLM conditioned on a structural-decomposition prompt, and binds each node to up to k supporting papers retrieved from D (we set k=60). This yields *citation-anchored* nodes rather than the free-floating, title-based labels common in prior clustering-based approaches.

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

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

where, at each round $r \in \{1, \ldots, R\}$ (we set R=3), Ref inspects the current tree together with a judge-returned tree-quality and citation-precision score, and performs three operations: pruning redundant nodes, repairing parent–child mis-attachments, and reassigning misplaced supporting papers. **The judge component of Ref is realized by `CRITERIA_BASED_JUDGING_PROMPT`, and node-level edits are formalized via `EDIT_FINAL_OUTLINE_PROMPT` (Appendix E).** The resulting tree K serves as a citation-grounded conceptual backbone for subsequent stages.

3.2 Tree-Induced Table Generation

Based on the refined tree K, MVSS generates structured comparison tables that explicitly expose discriminative dimensions among methods and subfields. We extract paper IDs $I_K \leftarrow \mathrm{IDs}(K)$ and fetch contextual metadata $C \leftarrow \mathrm{Fetch}(D, I_K)$, including titles, abstracts, and other descriptive fields. Table generation is formulated as a conditional mapping

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

where Tab is invoked at every subtree spanning ≥ 3 distinct methods: its columns are discriminative dimensions induced from the subtree's children, and its rows are anchored to exact paper titles in C. This guarantees each table operates at a consistent conceptual level and aligns one-to-one with

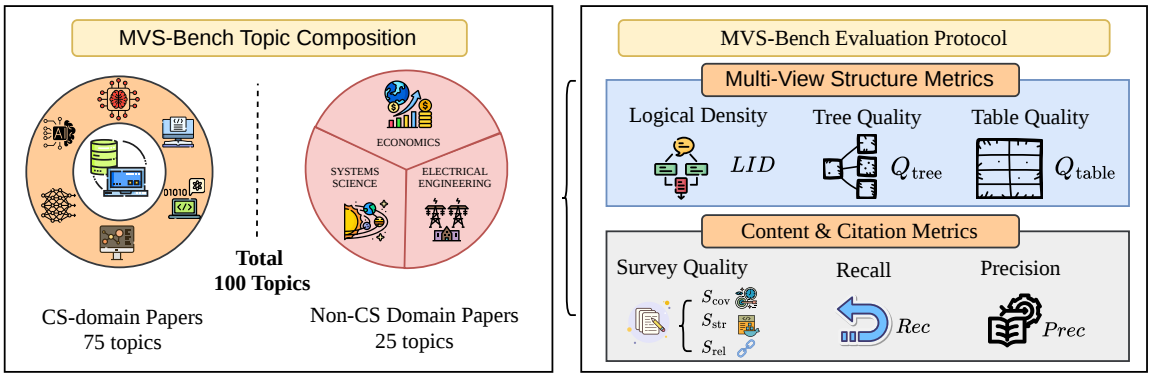

Figure 3: **MVS-Bench**: Dataset composition and evaluation protocol. **Left**: The dataset comprises 100 expert-curated topics paired with reference papers, explicitly stratified into 75 Computer Science (CS) domains and 25 non-CS domains (Economics, Systems Science, and Electrical Engineering) to evaluate cross-domain generalization. **Right**: The multi-dimensional evaluation protocol. Metrics are decoupled into Multi-View Structure Metrics (e.g., Logical Contribution Density (LID), Tree Quality Q_{tree}, and Table Quality Q_{table}) and Content & Citation Metrics (e.g., Survey Quality Q_{survey}, Logical Contribution Density, alongside Citation Recall and Precision).

its node, yielding evidence-grounded comparisons coherent with the global survey. **Operationally, `LOCAL_TABLE_REFLECT_PROMPT` executes** Tab **to emit a Markdown table (Appendix E).**

3.3 Tree+Table Guided Outline and Survey

In the final stage, MVSS jointly leverages the tree K and tables B to guide outline construction and survey writing. We generate m outline candidates and select one with a judge model:

$$O_k \leftarrow \text{Out}(T, K, B), \quad k = 1, \ldots, m$$
$$O \leftarrow \text{Judge}(\{O_k\}_{k=1}^{m})$$

We initialize outline candidates utilizing `ROUGH_OUTLINE_PROMPT`, expand their structures via the `SUBSECTION_OUTLINE_PROMPT`, and merge them into the final O using `MERGING_OUTLINE_PROMPT` (Appendix E). Conditioned on the final outline and structural constraints, the survey text is generated as

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

where Write is realized by `SUBSECTION_WRITING_PROMPT`, which enforces explicit constraints (minimum subsection length, non-repetition of subsection titles, and `[Title]`-only citations), with O controlling section ordering, K enforcing conceptual hierarchy, and B guiding comparative analysis and evidence usage.

Finally, a cross-view alignment operation $(K, B, O, S) \leftarrow \text{Align}(K, B, O, S)$ systematically resolves four types of inconsistency by inspecting the generated artifacts: (i) **Coverage check**: Identifies tree nodes in K that lack narrative coverage in S; affected segments trigger a targeted rewrite to incorporate the missing concepts. (ii) **Table-narrative binding**: Detects table rows in B that remain uncited in S; uncited rows trigger a targeted revision to ensure the comparative discussion reflects the table entries via `[Title]` citations. (iii) **Citation faithfulness**: Flags hallucinated citations in S that are absent from the corresponding node's bound papers. These trigger a rewrite where citation faithfulness is strictly re-verified by `CHECK_CITATION_PROMPT`, and claim-source entailment is explicitly scored by `NLI_PROMPT`. (iv) **Traversal consistency**: Evaluates whether the section ordering in O violates the structural traversal of K; violations trigger targeted reordering to guarantee that the narrative flow strictly follows the hierarchical tree. This step yields a coherent, multi-view survey with consistent structure and evidence grounding. The full prompt templates for each operator (Tree, Ref, Tab, Out, Write, Align) are listed in Appendix E.

4 Experiments

Existing automated survey systems are evaluated almost exclusively on ad-hoc CS topic lists with heterogeneous corpora, no paired expert references, and no test of cross-domain generalization; their metrics are borrowed piecemeal from summarization and miss the two properties that distinguish a survey from a literature dump—*structural organization* and *evidence quality*. To enable rigorous and discipline-general evaluation, we first con-

struct **MVS-Bench**, a cross-domain benchmark for automated survey generation (§4.1, §4.2), and then use it to evaluate MVSS along three axes: knowledge-tree quality, comparison-table quality, and overall survey quality.

4.1 MVS-Bench: Dataset

MVS-Bench contains **100 topics** curated by citation-impact ranking on Google Scholar with explicit disciplinary stratification: **75 CS topics** (ML, NLP, CV, systems) and **25 non-CS topics** (economics, electrical engineering, systems science). Non-CS disciplines are selected for their established survey traditions, distinct methodological vocabularies (a stringent generalization test), and the availability of domain-matched annotators. Each topic is paired with a high-impact *expert-written reference survey* (CS subset: 50–674 citations; top-20 in Appendix F), which serves both as an upper bound and as an anchor for structural evaluation.

All systems retrieve from a *unified corpus*: ∼530,000 arXiv papers (2018–2024) for CS/EE/systems science, and ∼80,000 RePEc/SSRN papers for economics, with an identical budget of at most 60 papers per local query. All survey- and tree-level judgments follow a *four-annotator consensus protocol* ($\geq$ 75% agreement for inclusion), with domain-matched annotators (M.S./Ph.D. students from four top universities for CS; discipline-matched Ph.D. students for the rest). The protocol achieves Fleiss' $\kappa = 0.81$ overall, with Relevance and screening reaching $\kappa \geq 0.83$. Full qualifications, screening, per-dimension agreement, and a side-by-side comparison against prior evaluation setups are in Appendix A.

4.2 MVS-Bench: Evaluation Protocol

Built on the dataset, MVS-Bench specifies a unified protocol with three 5-point survey dimensions (**Coverage**, **Structure**, **Relevance**), two holistic structure-level criteria (**TreeQuality**, **TableQuality**), citation-level recall/precision, and a corpus-agnostic evidence-quality metric (**LID**). All Likert criteria use the anchors in Appendix B and are scored by three calibrated LLM judges (**GPT-4o**, **Gemini-2.5-Pro**, **DeepSeek-chat**), validated against expert annotations (LLM–human correlations in §4.7, Appendix D.1).

Survey, tree, and table quality. Coverage assesses comprehensiveness, Structure evaluates logical organization and non-redundant flow, Relevance measures focus. We define $Q_{\text{survey}} = \frac{1}{3}(S_{\text{cov}} + S_{\text{str}} + S_{\text{rel}})$. TreeQuality scores taxonomy correctness, branch coverage, and grouping clarity ($Q_{\text{tree}} = S_{\text{tq}}$); TableQuality scores comparison-table correctness, completeness, and utility ($Q_{\text{table}} = S_{\text{tab}}$).

Citation quality. Following scientific fact verification, we extract claims $C = \{c_i\}$ and model-proposed (claim, reference) pairs $P = \{(c_i, r_j)\}$, with an NLI verifier $V(c, r) \in \{0, 1\}$. Citation recall and precision are

$$\begin{aligned} \text{Rec} &= \frac{|\{c \in C : \exists r, (c, r) \in P, V{=}1\}|}{|C|}, \\ \text{Prec} &= \frac{|\{(c, r) \in P : V{=}1\}|}{|P|} \end{aligned} \tag{1}$$

Logical Contribution Density (LID). Because baselines retrieve over *different* paper sets, Rec/Prec are **not directly comparable across systems**—a system retrieving a small set of obvious "safe" papers can trivially inflate both. We therefore introduce **LID**, a corpus-agnostic measure on each system's top-N retrieved papers ($N{=}60$): we extract a one-sentence core idea per paper, embed with Sentence-BERT, cluster with DBSCAN, and set $\text{LID} = N_{\text{cluster}}/N$. Higher LID indicates more distinct logical contributions and less redundant padding, *regardless of which papers were retrieved*. LID is validated against human judgments of evidence quality in Appendix B (Pearson $r{=}0.84$).

4.3 Experimental Setup

The main analysis focuses on the 75 CS topics for direct comparability with prior work; cross-domain results on the 25 non-CS topics are in §4.6. We use **deepseek-chat** as the primary generator for MVSS. To ensure fair comparison, we enforce four controls: (i) unified retrieval budget ($\leq$60 papers per local query); (ii) unified backbone (deepseek-chat for all systems); (iii) unified iteration budget ($R{=}3$ for iterative baselines); (iv) identical topic set. Full hyperparameters, corpus preprocessing, and runtime/cost details are in Appendix C.

Table 1: Performance comparison combining **multi-model LLM judges** and **double-blind human evaluation** on the 75-CS-topic main benchmark. **Avg**: (C+S+R)/3. **H-Str/H-Ovr**: expert human ratings on Structure/Overall. **LID**: Logical Contribution Density. Human writing is a reference upper bound. Cross-domain results on 25 non-CS topics are in §4.6.

Judge	**GPT-4o**				**Gemini-2.5-pro**				**DeepSeek-chat**				**Human**		**Metric**
Method	C	S	R	Avg	C	S	R	Avg	C	S	R	Avg	H-Str	H-Ovr	LID(%)
LLMxMapReduce	4.27	4.43	4.00	4.23	4.35	4.29	4.73	4.46	4.08	4.10	4.00	4.06	3.95	4.13	84
SurveyForge	4.13	4.05	4.58	4.25	4.03	3.38	4.66	4.02	4.05	3.95	4.05	4.02	3.90	4.08	90
SurveyG	4.45	4.76	3.98	4.40	4.64	4.80	4.97	4.80	4.10	4.00	4.50	4.20	4.12	4.29	64
SurveyX	4.07	3.95	4.10	4.04	3.80	3.55	4.52	3.96	4.02	3.98	4.00	4.00	3.75	3.92	84
InteractiveSurvey	4.07	4.36	4.02	4.15	2.93	2.91	3.56	3.13	3.95	3.90	4.03	3.96	3.80	3.92	68
AutoSurvey	4.60	4.60	4.46	4.55	4.66	4.33	4.86	4.62	4.13	4.06	4.48	4.22	4.15	4.35	76
HiReview	3.67	3.00	4.00	3.56	3.94	3.00	4.00	3.65	3.79	3.77	4.00	3.85	3.50	3.65	72
Human writing (Ref.)	*4.50*	*5.00*	*4.76*	*4.75*	*4.90*	*4.62*	*5.00*	*4.84*	*4.66*	*4.50*	*5.00*	*4.72*	*4.89*	*4.88*	*98*
MVSS (Ours)	**4.86**	**4.80**	4.47	**4.71**	**4.98**	4.45	**4.99**	**4.81**	4.11	**4.39**	**4.99**	**4.50**	**4.48**	**4.58**	**92.0**

4.4 Baselines

We compare MVSS against representative human and automatic baselines:

- **Human Experts.** Expert-written surveys with manually curated hierarchies (upper bound).
- **AutoSurvey.** Drafts an outline and expands it into text with sentence-level citations.
- **HiReview.** Retrieves a fixed paper set and generates taxonomy-guided text.
- **LLMxMapReduce.** A MapReduce-style baseline summarizing and synthesizing retrieved papers.
- **SurveyForge.** Extracts and organizes key information into structured surveys.
- **SurveyG.** Structures multi-document summaries into surveys.
- **SurveyX.** A representative state-of-the-art automated survey system.
- **InteractiveSurvey.** Builds structure and content via iterative/human-in-the-loop feedback.

4.5 Main Results

- **MVSS consistently leads all automated baselines and closes the gap to expert writing.** Table 1 shows MVSS ranks first across all three judges. Under GPT-4o it achieves 4.71, surpassing AutoSurvey (4.55) and LLMxMapReduce (4.23); similar advantages hold under Gemini (4.81). MVSS tracks Human Experts closely under GPT-4o and Gemini (4.71 vs. 4.75; 4.81 vs. 4.84), with all improvements statistically significant by paired t-tests ($p < 0.05$; Appendix D.2).
- **Structural organization is the primary source of MVSS's advantage.** On *Structure* under GPT-4o, MVSS attains 4.80, substantially outperforming LLMxMapReduce (4.43) and SurveyForge (4.05). This confirms that explicit hierarchical modeling via HKT, with cross-view alignment, eliminates redundancy and yields coherent narrative flow.
- **Tree-guided retrieval maximizes logical contribution density.** MVSS achieves $\mathbf{92.0}\%$ LID, exceeding all baselines. Structure-conditioned retrieval not only ensures factual accuracy but actively filters redundant padding.

4.6 Cross-Domain Generalization

To evaluate survey generation *across disciplines*, we test all systems on 25 non-CS topics (economics, EE, systems science) while preserving the fairness controls of §4.3.

Averaged across three LLM judges (Table 3), MVSS achieves **4.66**, leading the best baseline (AutoSurvey, 4.32) and narrowing the gap to human experts (4.79) to 0.13. This advantage peaks on Structure (4.65 vs. 4.20) and within economics (Δ=+0.40)—the discipline whose corpus differs most from CS. Detailed breakdowns and domain-matched human evaluations (Appendix D.3) confirm that *structure-first generation is a discipline-general principle for automated survey synthesis*.

4.7 Human Evaluation

Complementing automated metrics, we conducted a *double-blind* human study on 75 CS topics; expert ratings parallel LLM scores in Table 1 (H-Str,

Table 2: Ablation study for MVSS with components removed. Stricter single-judge (GPT-5.1) protocol; relative rankings are the meaningful quantity.

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 alignment	4.05±0.21	3.22±0.23	3.84±0.18	3.75±0.22	81.15±7	76.10±5	3.88±0.28	3.65±0.45
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
MVSS w/o explicit formatting	4.02±0.19	3.17±0.32	3.88±0.21	3.79±0.18	79.85±6	77.25±4	1	1

Table 3: Cross-domain generalization on the 25 non-CS topics. Scores averaged across GPT-4o, Gemini-2.5-Pro, and DeepSeek-chat.

Method	Cov	Str	Rel	Avg
LLMxMapReduce	4.05	4.12	4.21	4.13
SurveyForge	3.92	3.78	4.30	4.00
SurveyG	4.05	4.10	4.10	4.08
SurveyX	3.85	3.65	4.18	3.89
InteractiveSurvey	3.70	3.55	3.92	3.72
AutoSurvey	4.28	4.20	4.48	4.32
HiReview	3.58	3.20	3.85	3.54
Human writing (Ref.)	*4.72*	*4.80*	*4.85*	*4.79*
MVSS (Ours)	**4.60**	**4.65**	**4.72**	**4.66**

H-Ovr). MVSS achieves 4.58 ± 0.38 overall, significantly outperforming all automated baselines ($p < 0.05$) and approaching human-written reference quality. LLM-human correlations (Pearson/Spearman) exceed 0.79 across all dimensions (Appendix D.1). Protocols, qualifications, and agreement (Fleiss' $\kappa = 0.81$) are in Appendix A; non-CS evaluations are in Appendix D.3.

4.8 Ablation Studies

We evaluate four design choices: hierarchical tree guidance, iterative refinement, multi-model outline consensus, and explicit formatting. Removing explicit formatting (retaining logical constraints in pure text) yields minimal drops in Q_{survey}, Rec, and Prec, proving MVSS's gains stem from logical constraints rather than formatting bias. Additionally, a format-agnostic analysis of introductions (Appendix D.4) shows MVSS leading on 4/5 academic-value criteria from DeepSurvey-Bench (Zhang et al., 2026), including critical analysis.

Removing tree guidance causes the largest drop in structural quality and citation precision, underscoring the necessity of hierarchical planning. Disabling iterative refinement or multi-model outline consensus degrades performance, confirming gains stem from synergistic interaction.

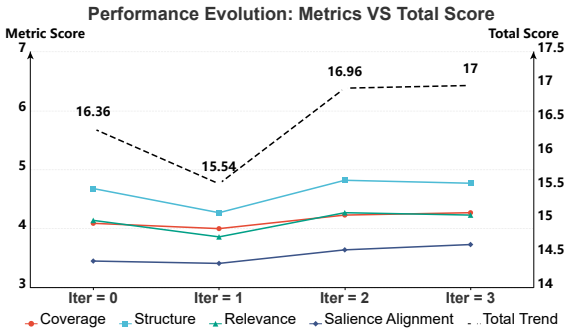

Figure 4: Performance across iterative tree-refinement rounds. Iter 0: initial tree without refinement.

Figure 4 shows quality evolution: Iter 1 dips as noisy branches are pruned; subsequent iterations monotonically improve Coverage, Structure, and Relevance, validating refinement.

5 Conclusion

We presented **MVSS**, a *multi-view structured survey generation* framework elevating conceptual structure to a primary optimization objective. By jointly constructing citation-grounded knowledge trees, comparison tables, and evidence-aware narratives, MVSS enforces robust structural and semantic alignment. Via outline consensus and structural refinement, MVSS improves structural clarity, comparative insight, and citation fidelity across **100 topics spanning computer science, economics, electrical engineering, and systems science**. Ultimately, MVSS reframes survey generation as a *structure-centric synthesis* problem, advancing scalable systems beyond mere summarization to actively organize scientific knowledge.

Limitations

While MVSS narrows the gap to expert-written surveys, several limitations remain. **First, failure-case analysis** on our 100 topics reveals two recurring failure modes: (i) emerging topics with fewer than $\sim$30 retrievable papers, where the induced tree collapses to shallow two-level hierarchies, and (ii) highly interdisciplinary topics whose evidence splits across disjoint sub-communities, occasionally producing over-merged tree nodes after cross-view alignment that obscure nuanced sub-field distinctions. **Second, hyperparameter sensitivity** is only partially characterized: while we report ablations on section granularity (Appendix D.5) and refinement rounds (Figure 4), the retrieval cutoff (k=60) is fixed throughout. A sensitivity sweep over $k \in \{30, 60, 90, 120\}$ would strengthen the empirical picture and help establish the optimal trade-off between structural breadth and computational overhead. **Third, handling conflicting evidence** remains a challenge; when scholarly contradictions arise within the retrieved pool, our current alignment mechanism prioritizes factual faithfulness to individual sources rather than explicitly synthesizing or highlighting the scientific debate. Incorporating a dialectical reasoning module to intelligently contrast opposing claims is a crucial next step. **Fourth, the NLI-based citation verifier** used in our Rec/Prec metrics inherits the calibration biases of its underlying model—multi-hop and aggregative claims are inherently harder to verify due to context fragmentation, so our 75%+ Rec/Prec figures should be read as *conservative lower bounds* rather than absolute fidelity guarantees. Finally, MVSS currently models a *static snapshot* of a field, lacking temporal constraints to capture the evolutionary trajectory, milestone breakthroughs, and paradigm shifts of research topics over time. Although our evaluation covers four disciplines, it still excludes argumentative-heavy domains (e.g., law, philosophy) where structural conventions differ and narrative synthesis often supersedes rigid taxonomies. Furthermore, we rely on computationally expensive frontier LLMs for multi-view generation and judging, raising cost, reproducibility, and potential environmental concerns. Mitigating these dependencies by distilling these capabilities into smaller, domain-specific open-source backbones remains a critical priority for future deployments.

Ethics Statement

Data Provenance and Human Annotators. MVS-Bench utilizes publicly available corpora (arXiv) in strict compliance with their terms of service. For reference surveys, we redistribute only metadata (titles, DOIs, etc.) to respect copyright. Our human evaluation protocol was approved by our institutional ethics board. Annotators (graduate students and domain experts) participated voluntarily, provided informed consent, remained anonymous, and were compensated at or above prevailing local academic rates, ensuring fair labor practices throughout the evaluation phase.

Fidelity, Misuse, and Intended Use. A primary ethical risk of LLM-based survey generation is citation hallucination and the deceptive inflation of scholarly output. While MVSS explicitly mitigates this via strict evidence grounding (HKT) and NLI-based entailment checks, it is strictly designed as an assistive drafting tool. It is not intended for high-stakes decision-making, automated peer review, or replacing human critical analysis. End-users must treat generated artifacts as drafts requiring rigorous expert verification before any formal use or publication to prevent the unintended proliferation of pseudo-science.

Evaluator Bias and Representation. LLM-as-a-judge protocols inherently carry biases reflecting their training distributions. To mitigate this, we aggregate scores across three independent LLMs and calibrate them against human expert ratings (Pearson $r > 0.79$). Furthermore, we acknowledge that MVS-Bench is currently dominated by English-language publications and high-citation topics, which may underrepresent non-English scholarship and emerging low-resource subfields, potentially amplifying dominant research paradigms at the expense of marginalized scientific voices.

Reproducibility and Environmental Cost. We open-source the MVSS pipeline, MVS-Bench, prompts, and evaluation scripts to ensure transparent reproducibility. Acknowledging the nontrivial energy consumption and carbon footprint associated with large-scale LLM inference, we document API usage costs (Appendix C) and provide caching mechanisms for intermediate artifacts to minimize redundant computation.

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.

Yu Chao, Siyu Lin, and 1 others. 2025. LLM×MapReduce-V3: Enabling interactive in-depth survey generation through a mcp-driven hierarchically modular agent system. *Preprint*, arXiv:2510.10890.

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.

Leandro Carísio Fernandes, Gustavo Bartz Guedes, Thiago Soares Laitz, Thales Sales Almeida, Rodrigo Nogueira, Roberto Lotufo, and Jayr Pereira. 2024. SurveySum: A dataset for summarizing multiple scientific articles into a survey section. In *Brazilian Conference on Intelligent Systems (BRACIS)*.

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.

Chao-Chun Hsu, Erin Bransom, Jenna Sparks, Bailey Kuehl, Chenhao Tan, David Wadden, Lucy Lu Wang, and Aakanksha Naik. 2024. CHIME: LLM-assisted hierarchical organization of scientific studies for literature review support. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 118–132. Association for Computational Linguistics.

Yuntong Hu, Zhuofeng Li, Zheng Zhang, Chen Ling, Raasikh Kanjiani, Boxin Zhao, and Liang Zhao. 2024. HiReview: Hierarchical taxonomy-driven automatic literature review generation. *Preprint*, arXiv:2410.03761.

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.

Priyanka Kargupta, Nan Zhang, Yunyi Zhang, Rui Zhang, Prasenjit Mitra, and Jiawei Han. 2025. TaxoAdapt: Aligning LLM-based multidimensional taxonomy construction to evolving research corpora. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29834–29850. 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.

Xun Liang, Jiawei Yang, Yezhaohui Wang, Chen Tang, Zifan Zheng, Shichao Song, Zehao Lin, Yebin Yang, Simin Niu, Hanyu Wang, Bo Tang, Feiyu Xiong, Keming Mao, and Zhiyu Li. 2025. SurveyX: Academic survey automation via large language models. *Preprint*, arXiv:2502.14776.

Ran Liu, Ming Liu, Min Yu, He Zhang, Jianguo Jiang, Gang Li, and Weiqing Huang. 2024. SumSurvey: An abstractive dataset of scientific survey papers for long document summarization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9632–9651.

Shuaiqi Liu, Jiannong Cao, Ruosong Yang, and Zhiyuan Wen. 2022. Generating a structured summary of numerous academic papers: Dataset and method. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4259–4265.

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*.

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.

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)*.

Minh-Anh Nguyen, Minh-Duc Nguyen, Ha Lan N.T., Kieu Hai Dang, Nguyen Tien Dong, and Dung D. Le. 2025. SurveyG: A multi-agent llm framework with hierarchical citation graph for automated survey generation. *Preprint*, arXiv:2510.07733.

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.

Zhiyuan Wen, Jiannong Cao, Zian Wang, Beichen Guo, Ruosong Yang, and Shuaiqi Liu. 2025. InteractiveSurvey: An llm-based personalized and interactive survey paper generation system. *Preprint*, arXiv:2504.08762.

Xiangchao Yan, Shiyang Feng, Lei Bai, and 1 others. 2025. SurveyForge: On the outline heuristics, memory-driven generation, and multi-dimensional evaluation for automated survey writing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*. ACL 2025, 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.

Guo-Biao Zhang, Ding-Yuan Liu, Da-Yi Wu, Tian Lan, Heyan Huang, Zhijing Wu, and Xian-Ling Mao. 2026. Deepsurvey-bench: Evaluating academic value of automatically generated scientific survey. *arXiv preprint arXiv:2601.15307*.

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*.

Kun Zhu, Lizi Liao, Yuxuan Gu, Lei Huang, Xiaocheng Feng, and Bing Qin. 2025. Context-aware hierarchical taxonomy generation for scientific papers via LLM-guided multi-aspect clustering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15616–15634. Association for Computational Linguistics.

A MVS-Bench: Construction and Human Evaluation

We document the full benchmark-construction protocol underlying §4.1, including annotator qualifications, screening procedure, inter-annotator agreement, and a side-by-side comparison against prior evaluation setups.

Annotator qualifications. For the 75 CS topics, the annotator pool consisted of M.S. and Ph.D. students from four top-tier universities, all specializing in CS subfields directly relevant to the evaluated topics (e.g., deep learning, NLP, computer vision). For the 25 non-CS topics, we additionally recruited **domain-matched annotators**: economics Ph.D. students for economics topics, EE Ph.D. students for EE topics, and systems-science researchers for systems-science topics. All evaluators possess the domain expertise to assess complex scientific literature and hierarchical taxonomies in their respective fields.

Screening and consensus mechanism. Each candidate paper and generated survey structure was independently reviewed by four expert annotators. To mitigate subjective bias and guarantee a high-consistency "gold standard," we enforced a strict consensus rule: a sample or judgment was retained only if at least three of four annotators reached definitive agreement (a $\geq 75\%$ consensus requirement). Any sample failing this threshold was excluded.

Inter-annotator agreement (IAA). We measured IAA using Fleiss' Kappa (κ). Table 4 reports the breakdown across screening and fine-grained qualitative dimensions. Our expert panel achieved a robust average $\kappa = 0.81$ (substantial-to-almost-perfect); objective dimensions such as Relevance and Screening reached "Almost Perfect" agreement. This procedure ensures the reproducibility

and credibility of the human-grounded evidence in our experiments.

Table 4: Inter-Annotator Agreement (Fleiss' κ) across evaluation tasks. Agreement-level interpretation follows Landis & Koch (1977).

Evaluation Task / Dimension	Fleiss' κ	Agreement
Survey Screening (Inclusion/Exclusion)	0.85	Almost Perfect
Coverage (Cov)	0.77	Substantial
Structure (Str)	0.79	Substantial
Relevance (Rel)	0.83	Almost Perfect
Average Overall	**0.81**	**Almost Perfect**

Position relative to prior evaluations. Table 5 contrasts MVS-Bench with prior evaluation setups along five dimensions necessary for reproducible, discipline-general evaluation: a fixed citation-anchored topic set, explicit cross-domain coverage, paired expert reference surveys, a unified retrieval corpus, and a corpus-agnostic evidence-quality metric. To our knowledge, MVS-Bench is the first benchmark satisfying all five.

Table 5: Comparison of MVS-Bench with prior evaluation setups for automated survey generation.

Benchmark	Fixed topics	Cross-domain	Paired ref.	Unified corpus	Cross-sys. evidence
Multi-XScience (Lu et al., 2020)	✓	✗	✗	✗	✗
BigSurvey (Liu et al., 2022)	✓	✗	✗	✗	✗
SciReviewGen (Kasanishi et al., 2023)	✓	✗	✗	✗	✗
SumSurvey (Liu et al., 2024)	✓	✗	✗	✗	✗
SurveySum (Fernandes et al., 2024)	✓	✗	✗	✗	✗
AutoSurvey-eval (Wang et al., 2024)	✗	✗	✗	✓	✗
MVS-Bench (ours)	✓	✓	✓	✓	✓

B Evaluation Metric Details

Survey-level Likert anchors. Table 6 lists representative 1/5 anchors for the survey- and structure-level Likert criteria used throughout MVS-Bench.

Table 6: Evaluation criteria of MVS-Bench. All dimensions use a 1–5 Likert scale; representative anchors (1/5) shown.

Criterion	Anchors of 1–5 scale
Coverage	1: Very limited; misses most key areas. 5: Fully comprehensive; covers key and peripheral topics in depth.
Structure	1: No clear logic between sections. 5: Tightly structured, clear logic, smooth transitions, no redundancy.
Relevance	1: Outdated/unrelated to the topic. 5: Exceptionally focused; every detail supports the topic.
TreeQuality	1: No meaningful tree or wrong hierarchy. 5: Comprehensive, correct, clear grouping, useful abstraction.
TableQuality	1: No usable table or misleading. 5: Comprehensive comparisons with consistent formatting.

Fine-grained HKT tree metrics. While *TreeQuality* captures macro-level utility within the end-to-end pipeline, it does not fully capture the intrinsic structural and content nuances of the taxonomy. We therefore developed a specialized rubric of four dedicated dimensions (Table 7).

Table 7: Detailed evaluation criteria for Hierarchical Knowledge Trees (HKT), with anchors at extremes (1 and 5).

Criterion	Anchors of 1–5 Scale
Coverage	**1:** Very limited coverage, citing few retrieved papers, omitting key areas. **5:** Comprehensive coverage of central and peripheral aspects.
Structure	**1:** Lacks clear organization; sections arranged arbitrarily. **5:** Tightly organized, logically coherent at all levels, balanced hierarchical flow.
Relevance	**1:** Largely irrelevant keywords/citations. **5:** Every keyword and citation tightly aligned to the topic.
Salience Alignment	**1:** Keyword ordering inconsistent with importance; trivial items early. **5:** Major keywords consistently placed in prominent positions.

LLM-judge details. We use **GPT-4o**, **Gemini-2.5-Pro**, and **DeepSeek-chat** for the main benchmark, all supporting 128K+ token inputs so full expert surveys can be ingested without truncation. Prompts are aligned with the human-written guidelines (Appendix E, `CRITERIA_BASED_JUDGING_PROMPT`), and a small expert-annotated set is used for scale calibra-

tion (Kocmi and Federmann, 2023; Fabbri et al., 2021). For ablations (Table 2, Figure 4) we additionally use **GPT-5.1** with a stricter rubric calibrated to amplify inter-variant differences; absolute scores under this protocol are systematically lower than under the lenient multi-judge panel, but relative rankings remain the meaningful quantity.

LID validity: correlation with human judgments of evidence quality. LID is designed as a corpus-agnostic proxy for the diversity of distinct logical contributions in a retrieved paper set. To validate that LID tracks human judgments, we conducted a focused study on 30 randomly sampled CS topics. For each topic, four annotators (separately recruited, blind to system identity) rated the top-60 retrieved set of each system on a 5-point "logical diversity" scale (*1: highly redundant*; *5: each paper contributes a distinct idea*). Across the 210 (topic, system) pairs, Pearson correlation between LID and averaged human diversity is $r = \mathbf{0.84}$ ($p < 10^{-19}$), Spearman $\rho = \mathbf{0.81}$; LID reproduces the human rank order on 27 of 30 topics (Table 8). This confirms LID is a meaningful, human-aligned, cross-system-comparable measure of evidence-pool quality.

Table 8: Correlation between LID and human-judged logical diversity of retrieved paper sets (30 CS topics $\times$ 7 systems).

Comparison	Pearson r	Spearman ρ
LID vs. human diversity	0.84	0.81
LID vs. Coverage (LLM)	0.71	0.69
LID vs. TableQual. (LLM)	0.66	0.63

C Implementation Details

Hyperparameters. Unless otherwise stated, we retrieve 1,200 topic-relevant papers from the underlying corpus (arXiv for CS, EE, and systems-science topics; RePEc/SSRN for economics topics), using abstracts and metadata for initial tree and outline generation. For fine-grained text synthesis in MVSS, we retrieve the top-$k = 60$ most relevant papers per tree node or table row. The reflection loop in the HKT module runs for three iterations ($R = 3$), dynamically selecting the best structure candidate via a joint assessment of TreeQuality and citation metrics. The temperature is fixed at 1.0 for all LLM API calls.

Cost and runtime. Table 9 reports average per-topic API cost (standard public pricing) and end-to-end wall-clock time.

Table 9: Average cost and runtime per topic.

Method	API Cost ($)	Time (min)
HKT generation only	0.55	11.30
Full MVSS pipeline	0.94	30.48

D Additional Experimental Results

We organize additional results by relevance to the main claims: (§D.1) validation that LLM judges agree with humans, (§D.2) statistical significance of MVSS gains, (§D.3) detailed cross-domain breakdowns, (§D.3) format-agnostic content analysis, (§D.4) tree-level comparison, (§D.5) section-granularity analysis, and (§D.6) a reader-preference pilot.

D.1 LLM–Human Agreement

Pearson (r) and Spearman (ρ) correlations between automated LLM ratings and human expert Likert ratings over all 75 CS topics (Table 10) consistently exceed 0.77, confirming LLM judges are reliable proxies for expert assessment in this domain.

Table 10: Correlation between LLM scores and human ratings over 75 CS topics.

Score Pair	Pearson r	Spearman ρ
Q_{survey} vs. Human Overall	0.81	0.83
LLM Cov vs. Human Cov	0.84	0.85
LLM Str vs. Human Str	0.81	0.77
LLM Rel vs. Human Rel	0.79	0.82

D.2 Statistical Significance

Paired t-tests of MVSS against each baseline on overall survey quality across the 75 CS topics confirm that all improvements reported in Table 1 are statistically significant.

Table 11: Statistical significance (paired t-test) of MVSS against core baselines on overall quality.

Comparison Pair	t-stat	p-value	Sig.
MVSS vs. AutoSurvey	3.86	0.0002	Yes (***)
MVSS vs. SurveyG	5.12	$< 10^{-5}$	Yes (***)
MVSS vs. LLMxMapReduce	7.35	$< 10^{-9}$	Yes (***)
MVSS vs. SurveyForge	6.88	$< 10^{-8}$	Yes (***)
MVSS vs. SurveyX	9.45	$< 10^{-12}$	Yes (***)
MVSS vs. InteractiveSurvey	8.64	$< 10^{-10}$	Yes (***)
MVSS vs. HiReview	11.23	$< 10^{-14}$	Yes (***)

D.3 Cross-Domain: Per-Judge, Per-Discipline, and Human Evaluation Detail

This appendix provides the detailed breakdowns underlying the aggregate cross-domain results in §4.6. All four fairness controls from §4.3 are preserved, and LLM-judge and human protocols mirror §4.7.

Per-judge breakdown. Table 12 shows MVSS's lead is consistent across all three judges rather than driven by any single evaluator.

Table 12: Per-judge overall (Avg) scores on the 25 non-CS topics, for MVSS and the two strongest automated baselines.

Method	GPT-4o	Gemini	DeepSeek	Avg
LLMxMapReduce	4.18	4.30	3.91	4.13
AutoSurvey	4.40	4.45	4.11	4.32
Human writing(Ref.)	*4.80*	*4.86*	*4.71*	*4.79*
MVSS (Ours)	**4.70**	**4.75**	**4.53**	**4.66**

Per-discipline breakdown. MVSS leads in every discipline (Table 13), with the largest margin in economics (Δ=+0.40 over AutoSurvey)—the discipline whose retrieval corpus (RePEc/SSRN) and citation conventions differ most from CS. The structure-first paradigm is robust to substantial shifts in literature distribution.

Table 13: Per-discipline overall scores (Avg) on the cross-domain subset.

Method	Economics	EE	Sys. Sci.
AutoSurvey	4.38	4.35	4.23
LLMxMapReduce	4.20	4.15	4.04
MVSS (Ours)	**4.78**	**4.68**	**4.52**

Human evaluation on non-CS topics. Following the same double-blind protocol as §4.7, domain-matched annotators rated MVSS at $\mathbf{4.45 \pm 0.36}$ overall on the non-CS subset, compared with $\mathbf{4.12 \pm 0.42}$ for AutoSurvey and $\mathbf{4.81 \pm 0.17}$ for human-expert references. IAA was Fleiss' $\kappa = 0.79$, consistent with CS results. These ratings reproduce the LLM-judge ordering, confirming cross-domain gains are not an artifact of automated evaluation.

Format-Agnostic Content Analysis To eliminate formatting bias and address concerns that LLM judges might favor explicit visual structures, we conducted a format-agnostic evaluation on the Introduction sections, where all systems use pure linear text narratives. We adopt five fine-grained academic-value metrics from DeepSurvey-Bench (Zhang et al., 2026): **OC** (Objective Clarity—clarity of stated research objectives), **CEC** (Classification–Evolution Coherence—clarity of method classification and the coherence of technological evolution), **DMC** (Dataset & Metric Coverage—completeness of datasets and rationality of evaluation metrics covered), **IC** (In-depth Comparison—systematic comparison of methods with explicit advantages and disadvantages), and **CA** (Critical Analysis—explanation of the underlying reasons for differences between methods). These five metrics cover the Information Value and Scholarly Communication Value dimensions defined therein; we omit the Research Guidance Value metrics (RG, FW) as our format-agnostic test targets the analytical core rather than forward-looking discussion.

Table 14: Fine-grained semantic evaluation of generated Introduction sections, on pure linear text to eliminate structural-formatting bias.

Method	OC	CEC	DMC	IC	CA
SurveyG	2.13	2.07	2.11	1.08	1.31
InteractiveSurvey	3.65	2.72	3.28	2.09	1.87
LLMxMapReduce	4.29	**3.84**	4.31	3.05	2.93
SurveyForge	3.73	3.49	3.86	2.71	2.68
SurveyX	2.68	1.85	2.29	1.11	1.07
AutoSurvey	3.67	3.52	3.89	2.47	2.32
HiReview	3.11	2.65	3.06	2.12	2.05
MVSS (Ours)	**4.32**	3.29	**4.45**	**3.17**	**3.08**

Excluding formatting bias, MVSS still leads on OC (4.32), DMC (4.45), and the reasoning-intensive IC (3.17) and CA (3.08). This confirms the structure-first paradigm genuinely enhances

logical density and analytical depth, not merely formatting bias.

D.4 Tree-Level Comparison

Table 15 isolates the structural-generation phase, showing HKT comprehensively outperforms naive RAG-based generation across all tree-level metrics, with the largest gain on Salience alignment (+1.10).

Table 15: Tree-level comparison between naive RAG-based 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**

D.5 Section Granularity Analysis

Table 16 shows content quality improves from 3 to 4 sections and saturates at 4–5, providing empirical guidance for optimal tree-depth constraints.

Table 16: 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**

D.6 Reader Preference Pilot Study

We surveyed 30 graduate students in CS subfields, asking which artifact they first consult when opening a survey: hierarchy/figure, table, or raw prose. 29 of 30 (96.7%) reported consulting a hierarchy or table first; we report this as supporting evidence for the structure-first design motivating MVSS.

E Prompt Templates

Table 17: Summary of key prompts and their functionalities in MVSS.

Prompt Name & Description
CRITERIA_BASED_JUDGING_PROMPT Given a survey and criteria with Score 1–5 descriptions, evaluate 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/No`.
ROUGH_OUTLINE_PROMPT Given `[PAPER LIST]`, `[PRIOR KNOWLEDGE MD]`, `[PRIOR KNOWLEDGE JSON]`, draft a comprehensive outline with `[SECTION NUM]` sections.
MERGING_OUTLINE_PROMPT Given outline candidates `[OUTLINE LIST]`, merge them into a single logical final outline.
SUBSECTION_OUTLINE_PROMPT Given an overall outline, prior knowledge, and a section description, generate subsections using `[PAPER LIST]`.
EDIT_FINAL_OUTLINE_PROMPT Refine a draft outline to remove duplicates and improve coherence. Return LaTeX-style format.
CHECK_CITATION_PROMPT Verify whether citations in a subsection are supported by papers in `[PAPER LIST]`. Fix or remove incorrect citations.
SUBSECTION_WRITING_PROMPT Write content ($>$ `[WORD NUM]` words) for a subsection. Cite using only `[Title]` format. Do not repeat subsection titles or output prior-knowledge trees.
LOCAL_TABLE_REFLECT_PROMPT Analyze the subsection and source papers. If ≥ 3 distinct methods are discussed, generate a comparison table in raw Markdown using exact paper titles.

F Reference Survey Papers

From Google Scholar, we strategically selected **100 highly influential surveys**—75 spanning CS domains and 25 spanning economics, EE, and systems science—to ensure balanced evaluation across citation counts and topic coverage. Table 18 lists the top-20 surveys from the CS subset by reference count; the full list (including the 25 non-CS surveys) is released with our code repository.

G Examples of Generated Structures

Table 18: Survey papers used for evaluation (top 20 from the CS subset, ranked by number of references). The full 100-survey list including economics, EE, and systems-science entries is released with our anonymous code repository.

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 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

```
A Survey of Blockchain and Artificial Intelligence for 6G Wireless Communications
 Section 1: Integration of Blockchain and AI in 6G Wireless Networks
    Subsection 1.1: Blockchain for AI Systems
       Blockchain consensus mechanisms: [63][64][65]
       Sharding techniques: [66][67][68]
       Decentralized ledger frameworks: [69][70][71]
       Privacy-enhancing tools: [72][53][73]
    Subsection 1.2: AI Integration in Blockchain
       AI-driven ledger optimization: [74][75][76]
       Federated learning frameworks: [77][78][79]
       Zero-knowledge proofs for security: [53][80][81]
       Blockchain scalability enhancements: [82][83][25]
    Subsection 1.3: Blockchain and AI Use Cases
        IoT frameworks: [84][85][86]
        Healthcare innovations: [87][88][89]
        Autonomous vehicles: [90][91][92]
        Smart city deployments: [93][94][95]
 Section 2: AI Applications in 6G Wireless Networks
    Subsection 2.1: AI Techniques for Optimization
       Reinforcement learning: [96][97][98]
       Dynamic resource allocation: [9][99][100]
       Generative AI for semantic communication: [101][102][103]
... (Hierarchy logically continues to cover the entire domain)
```

Figure 5: Qualitative example of a generated Hierarchical Knowledge Tree (HKT). Each subtopic is anchored to citation evidence, and the structured output aligns with downstream comparison tables.

Table 19: Comprehensive taxonomy and comparison of blockchain paradigms and consensus mechanisms for 6G wireless networks, synthesized from representative literature in the evidence pool.

Paradigm / Consensus	Key Characteristics	Suitable 6G Use Cases	Performance & Security Considerations	Refs
Public Blockchain	Permissionless, fully decentralized, transparent.	Spectrum auctioning, public data marketplaces.	High energy consumption (PoW); scalability bottlenecks; less suitable for latency-sensitive applications.	[135, 137]
Private Blockchain	Permissioned, single-entity governed, high throughput.	Internal network management, private industrial IoT.	High performance for trusted domains; faster federation than public chains.	[148]
Consortium Blockchain	Permissioned, pre-selected group governance.	Infrastructure sharing, smart-city trust management.	Balances decentralization and control; meets fine-grained 6G scalability needs.	[38, 136, 137]
Proof-of-Work (PoW)	Computational puzzles; high security but energy-heavy.	Legacy public applications, E-PoW for AI training.	Slow; misaligned with green 6G goals. E-PoW can repurpose power for AI tasks.	[135, 138, 139]
Proof-of-Stake (PoS)	Staked economic value; energy-efficient.	Energy-efficient public/consortium ledgers.	Potential wealth centralization; long-range-attack risk. Hybrids improve security.	[139]
BFT (e.g., PBFT)	High throughput, low-latency finality.	Real-time IoT data protection within consortiums.	Sensitive to network conditions; reliability drops beyond an "active distance" in THz channels.	[140, 141, 149]
RAFT	Crash-fault-tolerant; simple and efficient.	Trusted operator domains without malicious nodes.	Simplified consensus; impacted by wireless channel stability and active distance.	[140]
Proof-of-Reputation	Historical behavior-based leader selection.	Dynamic 6G networks, MEC trust evaluation.	Highest-reputation nodes form consensus group; critical for blockchain-based trust.	[142, 143, 150]
Proof-of-Auth (PoAh)	Cryptographic authentication based.	Resource-constrained edge IoT devices.	Lightweight; low latency ($\sim$3 secs) on hardware like Raspberry Pi.	[144]
DAG-based	Parallel transaction processing via graph.	Massive IoT micro-transactions, scalable systems.	Overcomes linear bottlenecks; security (double-spending) highly impacted by network load.	[145, 146]
Symbiotic (SBC)	Mutualistic transmission via cognitive backscatter.	Wireless PBFT/RAFT networks combating instability.	Increases consensus success and reduces energy via mutualistic node relationships. reduces energy via mutualistic node relationships.	[147]