Spatial Visual Analytics for Multi-Document Summary Verification

Jiahao Xu, Wei Liu, Yang Liu, Eric Krokos, Kirsten Whitley, Xuan Wang, Rebecca Faust, and Chris North

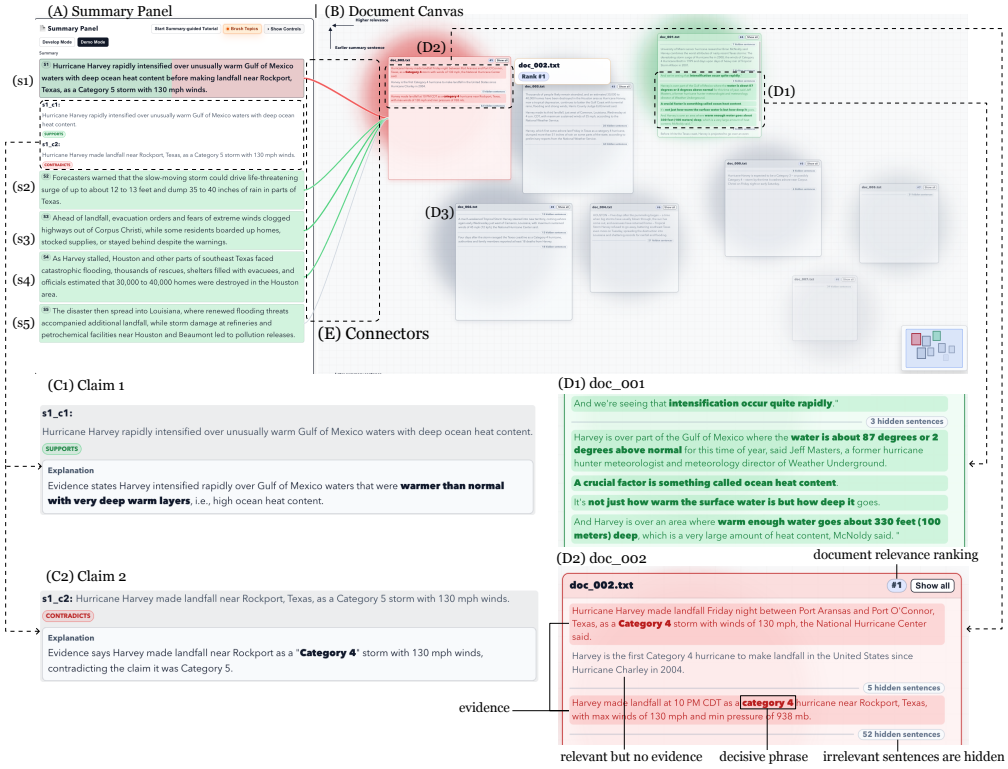

Figure 1: Summary Verification Space (SVS). The summary panel (A) presents sentences (S1–S5), decomposed claims (C1–C2), inference judgments, and explanations. The document canvas (B) arranges source cards (D1–D3), highlights supporting or contradictory evidence, compresses irrelevant content, and links summary content to evidence with connectors (E).

Abstract—Large language models increasingly generate summaries from collections of documents to support sensemaking and reporting, but verifying whether summary statements are grounded in source materials remains difficult. In multi-document summarization (MDS), evidence is distributed across many source documents and may be incomplete, conflicting, or missing. We present Summary Verification Space (SVS), a visual analytics system for verifying multi-document summaries through spatial document organization and coordinated provenance visualization. To support scalable verification, we investigate two alternative 2D canvas layouts: a SUMMARY-GUIDED layout that organizes documents by alignment with summary sentences, and a SOURCE-GUIDED layout that arranges documents by semantic similarity. Coordinated provenance visualization then makes relationships among summary content, source documents, and supporting evidence explicit, enabling users to trace support, contradiction, and missing evidence during verification. A task-driven usage scenario illustrates an auditing workflow in which users utilize the layouts to locate relevant documents and coverage gaps, then inspect linked claims and source evidence to make their own grounding judgments. In a comparative study with provenance held constant, both spatial layouts improved aggregate accuracy and reduced workload relative to a linear baseline, with the clearest gains on relevance tasks. The SUMMARY-GUIDED layout provides the strongest overall balance of accuracy, efficiency, confidence, and workload.

Index Terms—Summarization, verification, visual analytics, multi-document spatialization

1 INTRODUCTION

Large language models (LLMs) are increasingly used to generate summaries to support sensemaking and decision-making workflows [11,53]. By condensing large information spaces into concise narratives, summaries help users quickly grasp key topics and relationships without reading extensive source materials [1]. However, automatically generated summaries may contain unsupported statements, overlook important information, or present incomplete or conflicting interpretations [23]. As a result, users must determine whether summary content is factually grounded in the underlying sources.

This need makes summary verification a critical step in summary-supported analysis [16,25,33]. For example, during a literature review, a researcher using an automatically generated survey summary must verify whether key claims are supported by findings reported in individ-

- *Jiahao Xu, Wei Liu, Yang Liu, Xuan Wang, and Chris North are with Virginia Tech. E-mail: {jiahao, wliu3, yangliu07, xuanw, north}@vt.edu.*
- *Eric Krokos and Kirsten Whitley are with the Department of Defense.*
- *Rebecca Faust is with Tulane University. E-mail: rfaust1@tulane.edu.*

Manuscript received xx xxx. 202x; accepted xx xxx. 202x. Date of Publication xx xxx. 202x; date of current version xx xxx. 202x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.202x.xxxxxxx

ual research papers, identify studies that present conflicting results, and check whether important prior work or alternative viewpoints have been overlooked. Visual analytics offers a promising approach for verifying summaries by enabling users to externalize relationships between summaries and sources [7,13,20,49]. However, existing approaches are less effective in multi-document summary (MDS) scenarios, where verification requires users to identify relevant evidence across many documents. Designing effective provenance visualization for verifying summaries against source documents in MDS is fundamentally challenged by two key verification characteristics: relevance and consistency [15].

First, (**C1**) **relevance** poses a fundamental challenge in MDS verification. For any given summary, only a subset of the source collection is actually pertinent, yet that evidence may be scattered across many documents. Users must identify how relevance is distributed across the collection because the pattern of such distribution reveals if the summary covers important content from the sources. As document sets grow, this process becomes increasingly demanding, requiring repeated navigation and shifts of attention across sources. Supporting relevance is therefore a matter of helping users understand where pertinent evidence resides and how it is distributed at scale [13, 36].

Second, (**C2**) **consistency** poses a complementary challenge. Even after relevant evidence has been identified, users must still determine whether that evidence collectively supports, contradicts, or fails to substantiate the summary. In MDS, such judgments rarely depend on a single passage; instead, they often emerge from synthesizing information across multiple claims, sentences, and documents, including cases where evidence is incomplete, complementary, or in conflict. Supporting consistency verification therefore requires more than tracing isolated evidence fragments.

It requires helping users connect these evidence fragments to higher-level summary meaning [20, 37, 49].

To address these challenges, we present SVS (Fig. 1), a visual analytics system for MDS verification that combines spatial document organization for **(C1) relevance** with integrated provenance visualization for **(C2) consistency**. For **C1**, SVS centers on an interactive 2D canvas workspace that spatially organizes source documents. We identify two alternative layouts from different perspectives. The SUMMARY-GUIDED layout arranges documents by their relationships to summary sentences, making summary–document relevance immediately visible and helping users identify which parts of the collection are most pertinent to a verification target. The SOURCE-GUIDED layout instead places documents in a semantic embedding space, and positions summary sentences as visual anchors, so that users can interpret relevant documents within the broader structure of the source collection. We conduct a user study to compare these layouts, identify design tradeoffs, and measure improvement with respect to a baseline 1D layout used in existing systems.

SVS supports **C2** through an integrated provenance representation that links summary content to its evidence across the collection. Summary sentences are first decomposed into verifiable claims, and each claim is connected to localized evidence passages retrieved from the source documents. These provenance links are overlaid directly on the workspace, connecting summary and source documents within a coordinated view. This design allows users to move seamlessly between global relevance patterns and local evidence inspection, supporting judgment of whether summary content is supported, contradicted, or insufficiently grounded. Our paper offers three contributions:

- Two alternative spatial document layout strategies designed to support multi-document summary verification.
- An integrated provenance visualization design that supports reasoning about summary–source relationships across granularities.
- A usage scenario and user study demonstrating the value of spatial organization and provenance visualization for MDS verification.

2 Related Work

2.1 Visualizing Summary Verification

Early verification approaches focused on surface coverage assessment, commonly operationalized through lexical overlap metrics such as ROUGE [28]. Visualization research at this stage primarily supported similarity inspection workflows, employing ranked document views, token-level highlighting, and similarity heatmaps to help users examine correspondence between summaries and source texts at the lexical level [21, 40].

As summarization verification research evolved toward semantic grounding, newer evaluation methods introduced question-answering reconstruction [12, 48], semantic similarity scoring [17, 54], and natural language inference (NLI) models [26, 51] to assess whether summary claims are supported by source evidence. Visualization work in this phase increasingly emphasized exploration of semantic relationships, moving beyond lexical inspection to support reasoning about evidence relevance at scale [29, 41]. For example, SummVis [47] provides coordinated lexical and semantic visualizations that align model outputs, reference summaries, and source documents, enabling detailed inspection of factual consistency and abstraction patterns in generated summaries. Similarly, Summary Explorer [43] supports comparative analysis of summarization systems through interactive relevance ranking and similarity-based grouping views that reveal relationships among summaries and source evidence. Other interpretable summarization interfaces such as RTSUM [8] visualize salience signals at multiple granularities, including sentence- and phrase-level importance, helping users understand how salient textual units contribute to summary generation. These techniques improve faithfulness by enabling users to trace how individual summary statements relate to supporting passages and navigate semantic relationships across heterogeneous document collections. More recently, the emergence of LLM-as-a-Judge paradigms [27, 31] has further expanded opportunities for visualization. By producing structured verification outputs such as entailment labels, confidence scores, and natural language rationales, LLM-based evaluation methods allow interfaces to incorporate claim-level annotations, explanation panels, and evidence highlighting views that support user interpretation of automated judgments [35].

Although these modern techniques improve faithfulness analysis for summary verification, they provide limited support for MDS verification, where evidence is distributed across many documents rather than localized within a single source or a small comparison set. In such settings, users must not only find relevant evidence, but also understand how that evidence is organized across the collection and whether multiple sources collectively support, contradict, or fail to ground the summary [10, 37, 38, 47].

These limitations motivate the development of visual analytics systems that integrate automated verification signals with coordinated provenance visualization and spatial reasoning support to enable comprehensive multi-document summary verification workflows.

2.2 Spatial Document Organization

Spatial organization is a well-established strategy for supporting sensemaking in complex analytic tasks. Prior research shows that analysts arrange documents and visual artifacts within spatial workspaces to externalize relationships, maintain contextual awareness, and reduce cognitive effort during reasoning [2, 3]. Studies of large display analytics demonstrate that flexible spatial layouts can improve recall and facilitate reasoning about relationships in large data collections [24].

Early work on spatial hypertext systems explored how informal spatial arrangements enable users to construct knowledge representations without rigid hierarchical structures [39]. These systems allowed users to group related documents, form visual clusters, and incrementally organize information in ways that reflect evolving analytic goals [32]. Subsequent visual analytics tools extended these ideas by integrating spatial layouts with coordinated views [40], filtering mechanisms [6], and interactive exploration techniques [9], enabling users to examine thematic structures, relevance relationships, and distributional patterns across document corpora.

Recent research increasingly adopts spatial workspace paradigms to support exploration of semantic similarity and topic relationships across large textual corpora [30]. These systems employ clustering layouts, embedding projection techniques, and network representations to organize documents and model outputs within flexible visual canvases that facilitate exploratory reasoning [44,45]. For example, systems such

as Sensecape [42] support spatial grouping and iterative sensemaking across heterogeneous information sources, while ChainForge [4] introduces node-based spatial workflows that enable analysts to compare and organize large language model outputs during prompt experimentation. Building on similar spatial reasoning principles, LADiCA [55] provides a 2D canvas for arranging textual artifacts according to semantic relationships, allowing users to construct spatial representations of evolving analytic insights. By enabling users to externalize semantic structures and iteratively refine spatial layouts, these approaches highlight the continued importance of spatial organization as a cognitive scaffold for reasoning over complex, large-scale text collections.

Despite these strengths, less is known about how spatial layouts can be combined with provenance visualization for systematic evidence assessment. This gap is especially important for multi-document summary verification, where users must synthesize distributed and potentially conflicting evidence across sources. We address this gap by combining spatial document layouts with coordinated provenance visualization, helping users navigate complex evidence structures and assess the faithfulness of generated summaries.

3 Design Goals and System Overview

Our analysis of MDS verification suggests two design goals.

DG1: Scalable Overview of Relevant Evidence. In MDS verification, only a subset of documents and passages is pertinent to any given summary sentence or claim, yet that evidence may be distributed across many long sources. Users need support for understanding how relevant evidence is distributed across the collection because the pattern of such distribution reveals if the summary covers important content from the source. A verification interface should therefore provide an overview through spatial organization that reduces search and navigation burden while preserving meaningful relationships among documents. This design goal primarily addresses the challenge of **(C1) relevance**.

DG2: Explicit Representation of Consistency Relationships. In MDS verification, users must do more than identify relevant passages; they must determine whether distributed evidence collectively supports, contradicts, or fails to substantiate a summary statement. These judgments are often made between source documents and individual summary sentences, or even atomic claims within a sentence, rather than solely at the level of the summary as a whole [26, 52]. The interface should therefore integrate verification outputs into a coherent provenance representation that links localized evidence passages to atomic claims, summary sentences, and source documents. This design goal primarily addresses the challenge of **(C2) consistency**.

Following these goals, we designed SVS, a visual analytics system for supporting the verification of MDS through integrated provenance visualization and spatial document organization. The system interface is organized around two coordinated components: a summary panel (Fig. 1A) that presents the generated summary together with sentence- and claim-level verification results, and a 2D canvas (Fig. 1B) that displays source documents as interactive document cards. More implementation details are provided in Appendix C. Sections 4 to 7 describe the spatial layouts, provenance visualization, usage scenarios, and comparative study.

4 Spatial Organization of Source Documents

To support **DG1**, SVS organizes source documents spatially to provide a scalable overview of relevant evidence. We investigate two alternative grounding strategies for spatially organizing source documents: a novel SUMMARY-GUIDED projection (Fig. 1), which derives document positions from their alignment with summary structure, and a SOURCE-GUIDED semantic projection (Fig. 2), which follows conventional embedding-based dimensionality reduction to preserve semantic relationships among source documents.

4.1 SUMMARY-GUIDED Layout

The SUMMARY-GUIDED layout organizes source documents according to how they contribute to the structure of the summary. When assessing whether summary statements are well supported, users need to identify which documents are important overall and determine which documents contribute evidence to specific parts of the summary. The layout therefore arranges documents using two spatial ordering signals.

4.1.1 Horizontal Ordering: Overall Document Relevance

The horizontal axis encodes overall document relevance to the summary. Intuitively, this dimension answers the question: Which documents matter more for verifying this summary? During verification, users often prioritize documents that contain strong supporting or contradicting evidence. To reflect this need, relevance is determined by the strongest sentence-level relationship observed between a document and any summary sentence.

For each document d, the system constructs a similarity matrix $M_d \in \mathbb{R}^{n_d \times K}$, where n_d is the number of sentences in the document and K is the number of summary sentences. Each entry

$$M_d[i,j] = sim(s_i^d, t_j) \tag{1}$$

represents the semantic similarity between document sentences s_i^d and summary sentence t_j. The overall relevance score for document d is then defined as

$$R_d = \max_{i,j} |M_d[i,j]|. \tag{2}$$

Documents are sorted in descending order of R_d and mapped to evenly spaced horizontal positions on the canvas. This places documents containing decisive evidence, either strong support or strong contradiction, closer to the summary panel. For instance, in Fig. 1B, D1 is positioned to the left of D2 and D3, indicating that it is more important to the summary overall. Horizontal position therefore acts as a prioritization cue for which documents users should inspect first.

4.1.2 Vertical Ordering: Alignment with Summary Structure

The vertical axis encodes where a document's strongest positive evidence occurs within the summary sequence. This dimension answers the question: Which part of the summary does this document primarily relate to? It complements horizontal ordering because two documents can be equally relevant overall but correspond to different summary sentences or claims. Without this cue, users would still need to search within a ranked set to determine where each document fits in the verification workflow.

To estimate this alignment, the system computes for each summary sentence j the strongest positive similarity observed in the document:

$$w_d(j) = \max_i \max(M_d[i,j], 0). \tag{3}$$

These values form a non-negative distribution describing how evidence from document d is distributed across summary sentences. The document's alignment center is then defined as the weighted mean summary index:

$$\mu_d = \frac{\sum_{j=1}^{K} j \cdot w_d(j)}{\sum_{j=1}^{K} w_d(j)}. \tag{4}$$

Documents are sorted by μ_d and assigned evenly spaced vertical positions. Documents whose strongest evidence relates to earlier summary content appear higher in the layout, while those aligned with later summary regions appear lower. For example, in Fig. 1B, D1 and D2 align more strongly with earlier summary content, such as S1, while D3 aligns more strongly with later summary content, such as S3. By following the top-to-bottom order of the summary, the vertical arrangement turns the layout into a sentence-indexed evidence map, helping users separate evidence for different summary parts and notice where some summary regions have little document support.

4.2 SOURCE-GUIDED Layout

The SOURCE-GUIDED layout organizes source documents according to their semantic relationships within the collection rather than their relevance to summary order. During verification, users often need to understand how documents cluster around shared topics, where evidence is concentrated across the source set, and how individual summary sentences relate to these semantic neighborhoods. The layout

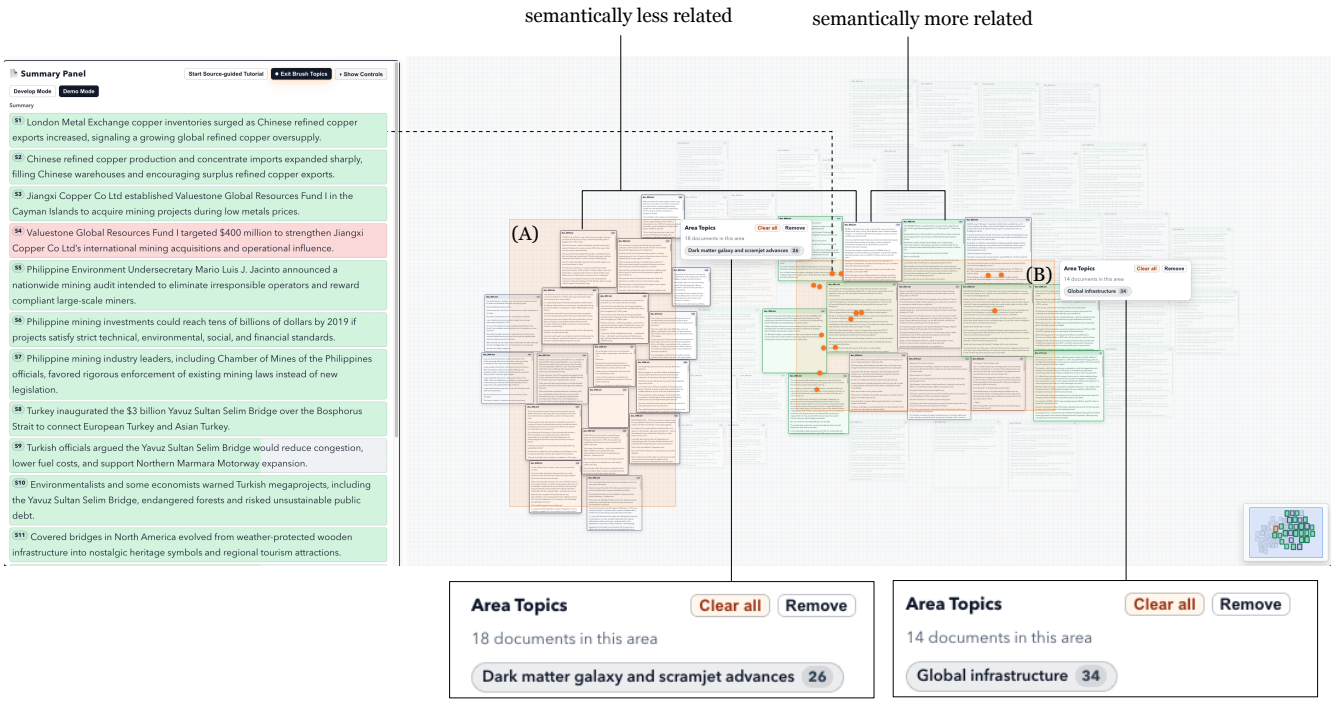

Figure 2: SOURCE-GUIDED layout. Documents form semantic clusters, while summary anchors show where evidence is concentrated. Brushed regions expose local topics, revealing that region A is largely uncovered by the summary.

therefore proceeds in two stages: it first projects documents into a space, and then places summary sentences within this document space using semantic signals.

4.2.1 Document Organization: Semantic Projection of Sources

The SOURCE-GUIDED layout begins by organizing documents according to overall semantic similarity. It represents each document using a whole-document embedding and projects the document collection into two dimensions via UMAP [34] (or any other commonly used projection algorithm such as t-SNE [46]). Documents that are similar in overall content are therefore placed near one another, while documents that are less relevant are positioned farther apart. This allows document clusters to reveal coherent topical neighborhoods within the source set.

Unlike the SUMMARY-GUIDED layout, the spatial axes in this mode do not carry explicit meaning. Instead, the layout is interpreted relationally: document position is meaningful primarily with respect to nearby documents. The distance between two projected documents,

$$\|\mathbf{z}_i - \mathbf{z}_j\|, \tag{5}$$

approximates their semantic similarity. As a result, the projection forms a semantic map of the source collection, where users read proximity as document similarity and clusters as source-level topic structure.

4.2.2 Summary Sentence Placement: Evidence Anchoring

After document positions are established, summary sentences are placed into the projection using the same sentence-level similarity matrices M_d from Eq. (1). For each summary sentence t_j, the system uses the document-level contribution signal $w_d(j)$ to represent how strongly document d contributes evidence to that sentence. These contribution strengths are then normalized into weights, and each summary sentence is placed at the weighted average of document positions:

$$\mathbf{p}_j = \sum_d \alpha_d(j)\mathbf{z}_d, \tag{6}$$

where $\alpha_d(j)$ denotes the normalized contribution weight of document d for summary sentence t_j. This process places each summary sentence near the documents that contribute most strongly to it. The resulting sentence anchors indicate where evidence for each summary statement is concentrated in the semantic document map. Anchors located between clusters suggest that evidence is distributed across multiple document groups rather than concentrated in a single source region. For example, as shown in Fig. 2, summary sentence anchors are concentrated in region B, corresponding to documents related to global infrastructure, while region A, associated with dark matter and scramjet advances, remains largely separated from the anchors. This spatial pattern suggests that evidence for the summary is concentrated within one topical region of the collection, whereas another coherent topic cluster receives little or no coverage.

5 COORDINATED PROVENANCE VISUALIZATION

To support MDS verification, SVS represents provenance as coordinated relationships among summary and source documents at multiple levels. These relationships are derived through a pipeline that decomposes summary content, retrieves evidence, and assigns claim-level inference judgments. The coordinated provenance is then visualized from the pipeline outputs. We first describe the verification pipeline, then the visual encodings and coordination mechanisms.

5.1 Verification Pipeline for Provenance Analysis

As shown in Fig. 3, SVS derives the provenance shown in the interface through four stages: summary structuring, claim decomposition, evidence retrieval, and claim-level inference judgment.

Summary structuring. Given a summary, the backend first segments it into an ordered list of summary sentences using spaCy [22] sentence segmentation. Source documents are segmented into sentences in the same way, and all summary sentences and source sentences are embedded with the same embedding model. Using these embeddings, the system computes the sentence-level similarity matrices M_d defined in Eq. (1), where each entry measures the semantic similarity between a source sentence and a summary sentence.

Claim decomposition. Each summary sentence is then decomposed into a set of atomic claims using an LLM. An atomic claim is a single verifiable proposition extracted from the sentence. The decomposition process enforces that claims remain self-contained and preserves attri-

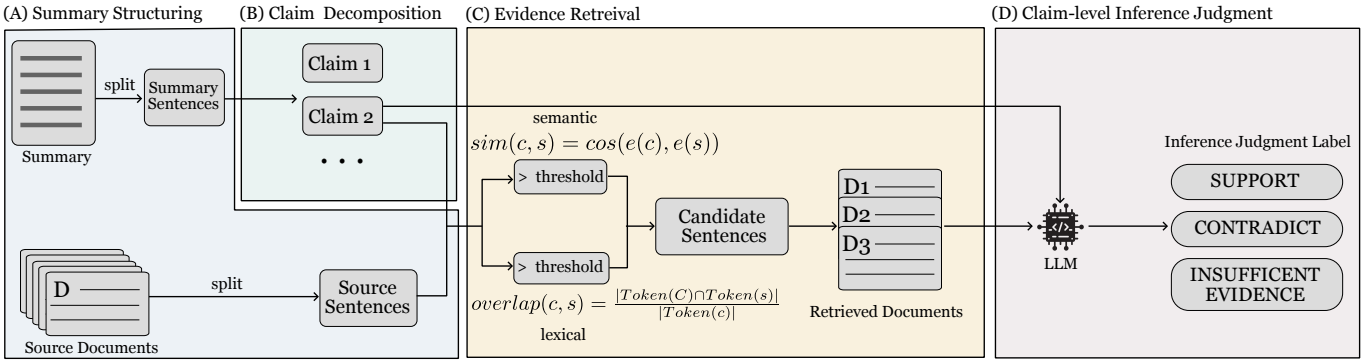

Figure 3: Verification pipeline for provenance analysis. SVS transforms summary content into structured provenance through four stages: summary structuring, claim decomposition, evidence retrieval, and claim-level inference judgment. Candidate evidence is identified using a hybrid semantic and lexical filter, and an LLM-based judge assigns SUPPORTS, CONTRADICTS, or INSUFFICIENT EVIDENCE labels.

bution and hedging, so the resulting claims can be judged independently without changing the meaning of the original sentence.

Evidence retrieval. For each atomic claim, the system retrieves candidate evidence from source-document sentences using a hybrid lexical-semantic filter. Let c denote an atomic claim, s denote a source-document sentence, and $e(\cdot)$ denote the sentence embedding. Each source sentence receives a semantic similarity score:

$$\text{sim}(c,s) = \cos(e(c), e(s)),$$

and a lexical overlap score:

$$\text{overlap}(c,s) = \frac{|\text{Token}(c) \cap \text{Token}(s)|}{|\text{Token}(c)|}.$$

A candidate sentence is retained if either its semantic similarity exceeds a semantic threshold or its lexical overlap exceeds a lexical threshold. The retained candidates are then grouped by document, producing the retrieved documents.

Claim-level inference judgment. Finally, each atomic claim is passed to an LLM-based judge together with the retrieved evidence packets. For each claim, the model outputs one of three labels [26]:

- SUPPORTS: the retrieved evidence jointly establishes the claim.
- CONTRADICTS: the retrieved evidence shows that at least one material part of the claim is false.
- INSUFFICIENT EVIDENCE: the retrieved evidence is insufficient to support or contradict the full claim.

The model also returns reason phrases, retrieved documents, and cited support or contradiction spans.

Overall, this pipeline converts summary content into structured provenance that can be visualized throughout the interface, providing the computational basis for the provenance encodings described next.

5.2 Provenance Visualization

SVS represents provenance across two coordinated views. The summary panel organizes provenance around summary sentences and their atomic claims, providing the primary entry point for verification. The document canvas grounds claim-level inference judgments in the source collection by showing how relevant evidence is distributed across documents and how each document relates to the current verification target.

5.2.1 Summary Panel Provenance

The summary panel serves as the primary entry point for provenance inspection. Provenance focus narrows progressively as users move from no active selection (Fig. 1S2–S5), to a selected summary sentence (Fig. 1S1), and then to an atomic claim (Fig. 1C). This progression shifts the panel from summary-level overview to sentence-level inspection and finally to claim-level evidence assessment.

Each summary sentence is shown as its own sentence card. At the overview level, the card background encodes the sentence's provenance state. When the sentence contains claims with different inference judgments, the background becomes a left-to-right mixed gradient whose proportions reflect the distribution of SUPPORTS, CONTRADICTS, and INSUFFICIENT EVIDENCE claims within that sentence. This makes sentence-level verification patterns visible without expanding the sentence. For example, in Fig. 1, the background of the second sentence (S2) contains only green, indicating full support from source documents.

When a sentence is activated, the panel expands to reveal its atomic claims (Fig. 1S1). Each claim is shown as a separate row with a status badge indicating one of three claim-level inference judgments: SUPPORTS, CONTRADICTS, and INSUFFICIENT EVIDENCE. This decomposition exposes provenance at the level where individual verification decisions are made, rather than forcing judgment at the sentence as a whole.

Selecting a claim opens a compact explanation panel that presents the textual rationale for the assigned inference judgment (Fig. 1C). Within this explanation, key phrases are bolded to isolate the most decisive parts of the justification. The summary panel therefore provides both a compact provenance overview and a claim-level explanation layer, while also establishing the provenance focus that drives the document-side encodings.

5.2.2 Document Canvas Provenance

Where the summary panel organizes provenance around verification targets, the document canvas grounds those targets in the source collection. Each source document is represented as a document card positioned within the spatial layout. Provenance at the document level is encoded through card color, which summarizes how the document relates to the current provenance focus. In addition, rank badges attached to card headers expose relevance ordering.

Within each card, provenance is further localized to source sentences (Fig. 1D2). Retrieved evidence sentences are colored by evidence role so that supporting and contradictory evidence is visible directly in document context. Decisive phrases inside evidence sentences are bolded. Sentences that are relevant but not direct evidence remain visible with softer styling, and non-evidence regions are compressed into compact placeholders.

The canvas also provides provenance cues that operate across the collection. Connector lines (Fig. 1E) link documents to summary sentences, with color indicating whether the document supports, contradicts, or insufficiently grounds each summary sentence.

To support localized thematic inspection, users can brush regions of the canvas to open topic summaries for the selected neighborhoods (Fig. 2A and B). These summaries are derived by matching brushed documents to a precomputed topic corpus, allowing local topic composition to be inspected quickly and compared across regions.

Taken together, these encodings turn provenance into a spatial representation of evidence across documents. The canvas therefore supports not only local evidence reading within a document card, but also cross-

document inspection of how support, contradiction, and insufficiency are distributed across the source collection.

6 Usage Scenario

6.1 Relevance Task: Assess Summary Coverage

In this section, we illustrate how SVS supports the relevance task of assessing summary coverage over a source collection [15]. We use a 60-document cluster from the WCEP dataset [18], containing 30 Business articles and 30 Science articles. The goal is to determine whether the summary covers the major topics represented in the source collection, a coverage question that is difficult to verify through prompting alone.

For example, a user could ask an LLM whether the summary covers the major topics in the source collection, but the resulting textual answer would still need to be audited against the collection. The key evidence is not only in individual cited passages, but also in how documents are distributed across topics, whether omitted content forms a substantial cluster rather than isolated examples, and whether the summary is concentrated around only part of the collection.

Because this evidence is distributed across 60 documents, manually comparing the summary against every source document is impractical. Instead, the SOURCE-GUIDED layout provides an overview of how the summary is positioned within the semantic structure of the collection. As shown in Fig. 2, the summary sentence anchors are concentrated in one region of the canvas rather than distributed across the full set of documents. This pattern suggests that the summary is aligned with only part of the source collection.

To examine which topics are covered and which may be omitted, users can further inspect two regions of the canvas (Fig. 2A and B). Topic analysis shows that documents in Area B focus on "global infrastructure", whereas documents in Area A focus on "Dark matter galaxy and scramjet advances". The summary anchors lie primarily within Area B, while Area A remains disconnected from the summary. SVS further indicates that 26 documents belong to the Area A topic group, making it a substantial portion of the collection. Taken together, these cues suggest that the summary emphasizes the business topic while largely omitting a major science-related topic cluster. The SOURCE-GUIDED layout therefore helps users identify important parts of the collection that are not adequately covered by the summary.

6.2 Consistency Task: Verify Summary Sentence

In this section, we illustrate how SVS supports the consistency task of verifying a summary sentence through coordinated provenance visualization. The scenario uses an eight-document hurricane cluster from the Multi-News dataset [14], where the documents describe the same event but provide overlapping and sometimes conflicting evidence across sources. This walkthrough shows how provenance visualization in SVS helps trace support and contradiction, while the value of spatial document organization is evaluated later in Sec. 7 through a comparative user study. The goal is to verify whether the following summary sentence is supported by the source documents, a consistency question that is difficult to audit through prompting alone:

(S1) *"Hurricane Harvey rapidly intensified over unusually warm Gulf of Mexico waters with deep ocean heat content before making landfall near Rockport, Texas, as a Category 5 storm with 130 mph winds."*

For example, a user could ask an LLM whether S1 is supported by the sources, but a textual answer would still need to be audited against the evidence. S1 contains multiple factual claims, and the same source collection may support one part while contradicting another. The user would still need to check which claim each source addresses, whether the cited evidence justifies the assigned inference judgment, and where conflicting evidence appears across documents.

To make this mixed evidence structure inspectable, the verification begins from the summary panel. The background gradient of S1 immediately indicates that the sentence is only partially consistent with the source collection: one part is supported, while another part is contradicted. This sentence-level provenance cue tells the user that S1 should be inspected more closely rather than accepted as fully grounded.

Selecting S1 activates its associated provenance across the interface. Two source documents become emphasized on the canvas: `doc_001` and `doc_002`. Their contrasting encodings reveal different verification roles. `doc_001` is surrounded by a green heatfield, indicating supporting evidence, while `doc_002` is surrounded by a red heatfield, indicating contradicting evidence. Expanding S1 further reveals that it contains two atomic claims. At this stage, the provenance visualization already shows that the evidence for S1 is distributed across multiple documents rather than concentrated in one source.

To understand why S1 is partially verified, the claims are inspected individually. Selecting the first claim reveals its explanation in the summary panel and highlights the corresponding supporting evidence in `doc_001`. Zooming into the document exposes the relevant sentences. These localized evidence cues confirm that Harvey intensified rapidly over Gulf of Mexico waters that were warmer than normal and had deep warm-water layers. In this way, the provenance visualization connects the SUPPORTS inference judgment for the claim to the specific source phrases.

The second claim can then be inspected in the same manner. Its explanation indicates that the inconsistency concerns Harvey's landfall strength. The summary states that Harvey made landfall as a *Category 5* storm, whereas the source evidence identifies it as *Category 4*. Inspecting `doc_002` reveals two highlighted sentences that explicitly describe Harvey as a Category 4 hurricane at landfall near Rockport. These passages provide direct contradictory evidence and explain why S1 is not fully supported by the source collection.

The final judgment for S1 is therefore mixed. The claim about rapid intensification over unusually warm Gulf waters with deep ocean heat content is supported, while the claim about landfall as a Category 5 storm with 130 mph winds is contradicted by the source evidence. This scenario shows how provenance visualization in SVS supports consistency analysis by linking sentence-level verification status, claim-level inference judgments, document-level polarity, and localized evidence within a single workflow.

7 Comparative User Study

We conducted a comparative user study with 18 participants to evaluate how different document organization strategies affect user performance in MDS verification. Because provenance visualization was held constant across conditions, the study isolates the effect of spatial document organization. We investigate the following questions:

- **Q1**: Does spatial document organization improve user performance over a LINEAR baseline for multi-document summary verification?
- **Q2**: How do SUMMARY-GUIDED and SOURCE-GUIDED layouts differ in supporting different verification tasks?
- **Q3**: How do users perceive and use the three organization conditions during verification?

7.1 Study Design

7.1.1 Conditions

All three conditions used the same summary panel and provenance visualization. They differed only in how source documents were organized within the workspace:

1. LINEAR: Documents are arranged in a fixed sequential order, without grouping based on summary structure or source relationships.
   - It represents a common provenance pattern in mainstream commercial LLM assistants (e.g., NotebookLM [19]) for summary verification: the summary is presented alongside a linear list of source materials that users can inspect.
2. SUMMARY-GUIDED: Documents are organized with respect to the summary, such that their positions reflect how they relate to different parts of the summary content.
3. SOURCE-GUIDED: Documents are organized according to inter-document relationships, such that documents with semantically similar or related source content are placed near one another.

7.1.2 Tasks

Participants completed ten tasks spanning two categories: **relevance** (Tasks 1–7) and **consistency** (Tasks 8–10). Relevance tasks focused on

identifying important topics and relevant documents, while consistency tasks focused on determining whether summary content was supported, contradicted, or unsupported by the source documents.

Relevance Tasks

1. **Topic identification**: Identify which topics are discussed in the source documents.
2. **Topic coverage estimation**: Estimate how many source documents discuss a given topic.
3. **Summary emphasis comparison**: Determine whether the summary discusses one topic more than others.
4. **Most relevant document to the summary**: Identify the source document that is most relevant to the summary as a whole.
5. **Least relevant document to the summary**: Identify the source document that is least relevant to the summary as a whole.
6. **Most relevant documents to a summary sentence**: For a given summary sentence, identify the top two documents that are most relevant to that sentence.
7. **Most relevant summary sentence for a document**: For a given source document, identify the most relevant summary sentence.

Consistency Tasks

8. **Supporting documents for a summary sentence**: For a summary sentence, select all documents that contribute to its inference judgment label.
9. **Contradicted claim identification**: Identify which claim is contradicted by the source documents.
10. **Unsupported claim identification**: Identify which claim has neither supporting nor contradicting evidence (insufficient evidence) across the source documents.

7.1.3 Datasets

We used three 60-document collections derived from the WCEP dataset introduced in Sec. 6. Each collection combined documents from two topical categories to create mixed-topic verification settings:

1. **Arts and Environment**: 30 Arts and Culture documents and 30 Environment documents.
2. **Disasters and Health**: 30 Disasters and Accidents documents and 30 Health and Medicine documents.
3. **Crime and Sports**: 30 Crime and Law documents and 30 Sports documents.

7.1.4 Participants

We recruited 18 participants from a university mailing list. The study protocol was approved by the Virginia Tech Institutional Review Board (Protocol No. 25-1176), and all participants provided informed consent before participating. All participants were able to read and understand English. During the study, participants completed questionnaires on a laptop while interacting with SVS on a 27-inch monitor.

7.1.5 Procedure

We used a within-subject design in which each participant completed all three interface conditions: LINEAR, SUMMARY-GUIDED, and SOURCE-GUIDED. To mitigate condition-order effects, we counterbalanced the presentation order using all six permutations of the three conditions.

The dataset order was fixed as Dataset 1, Dataset 2, and Dataset 3 for all participants. As described in Sec. 7.1.3, the three collections were carefully constructed to provide comparable mixed-topic verification settings, so dataset progression did not require additional counterbalancing. Each participant completed three rounds of tasks, one under each condition and on one dataset per round, with the mapping between condition and dataset determined by the assigned condition order.

Before starting each condition, participants were given a tutorial introducing the corresponding interface organization and interaction workflow. The tutorial was presented as an in-place guided overlay on top of the live workspace, allowing participants to learn the interface in the same environment used for the study tasks. After the tutorial, participants completed the assigned tasks for that condition. After all three conditions were completed, participants provided feedback on their experience using the system.

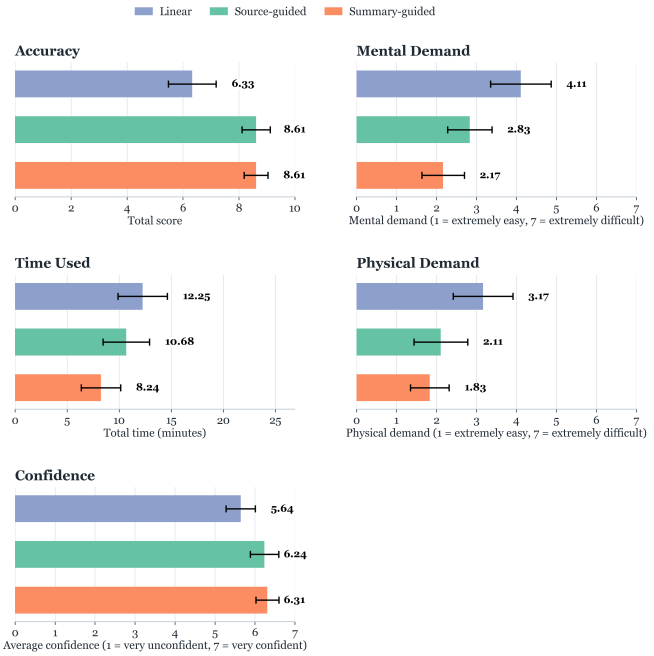

Figure 4: Overall performance across the three document organization conditions, including accuracy, completion time, confidence, mental demand, and physical demand (error bars indicate 95% confidence intervals).

7.1.6 Measures

We collected both task-level and condition-level measures. At the task level, we recorded the time required to complete each task and participants' self-reported confidence in their answer on a 7-point Likert scale: 1 indicated *very unconfident* and 7 indicated *very confident*.

After completing each condition, participants reported their perceived **mental demand**, reflecting the required cognitive effort, and **physical demand**, reflecting the required interaction effort. In addition, they provided free-text feedback describing which features of the condition they found helpful for completing the tasks. These responses were used to capture subjective workload and condition-specific impressions that were not reflected in task completion data alone.

7.2 Quantitative Results

We analyzed each metric using linear mixed-effects [5] models with *condition* as the main fixed effect, *dataset* as a control variable, and participant (*pid*) as a random intercept. For Fig. 4, models were fit across all tasks; for Fig. 5, the same model was fit separately within each task category. Reported omnibus p-values test the overall effect of *condition*, and pairwise p-values come from post hoc comparisons of the fitted models.

Overall performance. Figure 4 shows the overall effect of interface condition across all tasks.

For *accuracy*, mean scores were 6.33 for LINEAR, 8.61 for SOURCE-GUIDED, and 8.61 for SUMMARY-GUIDED, with a significant condition effect ($p = 8.07 \times 10^{-8}$). Both guided layouts outperformed LINEAR by 2.28 points (both $p = 2.70 \times 10^{-10}$), with no difference between them ($p = 1.00$). For *completion time*, means were 12.25, 10.68, and 8.24 minutes, respectively, with a significant condition effect ($p = 2.15 \times 10^{-4}$). SUMMARY-GUIDED was faster than LINEAR by 4.01 minutes ($p = 4.10 \times 10^{-6}$) and faster than SOURCE-GUIDED by 2.44 minutes ($p = 0.0050$). SOURCE-GUIDED was numerically faster than LINEAR, but not significantly ($p = 0.0716$).

Takeaway 1. Spatial document organization improves the effectiveness of MDS verification, and SUMMARY-GUIDED provides the best overall operating point by matching the accuracy gains of both spatial layouts while delivering the largest efficiency benefit.

For *confidence*, mean ratings were 5.64, 6.24, and 6.31, respectively, with a significant condition effect ($p = 3.47 \times 10^{-6}$). Both SOURCE-GUIDED and SUMMARY-GUIDED were higher than LINEAR ($p = 9.58 \times 10^{-7}$ and $p = 3.90 \times 10^{-8}$), with no difference between the two guided layouts ($p = 0.552$).

Workload ratings showed the same overall pattern. For *mental demand*, means were 4.11 for LINEAR, 2.83 for SOURCE-GUIDED, and 2.17 for SUMMARY-GUIDED, with a significant condition effect ($p = 1.05 \times 10^{-6}$). Both guided layouts were lower than LINEAR, and SUMMARY-GUIDED was also lower than SOURCE-GUIDED ($p = 0.030$). For *physical demand*, means were 3.17, 2.11, and 1.83, with a significant condition effect ($p = 8.97 \times 10^{-4}$). Both guided layouts were lower than LINEAR, while the difference between the two guided layouts was not significant ($p = 0.413$).

Takeaway 2. Guided spatial layouts improve the user experience of MDS verification by making users more confident while reducing perceived workload, with SUMMARY-GUIDED providing the lowest mental demand overall.

Task-specific results. Figure 5 shows task-level performance across the ten study tasks (T1–T10) for the three conditions.

For *T1* (topic identification), all three metrics showed significant condition effects. Both guided layouts outperformed LINEAR in *accuracy*, with gains of 77.78 percent correct for SOURCE-GUIDED and 72.22 percent correct for SUMMARY-GUIDED (both $p < 0.001$). Both guided layouts were also faster than LINEAR in *completion time*, by 1.57 and 1.36 minutes, respectively (both $p < 0.001$), and both produced higher *confidence*, by 1.93 and 1.72 points (both $p < 0.001$). *T2* (topic coverage estimation) showed the same pattern. Relative to LINEAR, SOURCE-GUIDED and SUMMARY-GUIDED improved *accuracy* by 61.11 and 55.56 points (both $p < 0.001$), reduced *completion time* by 1.68 and 1.81 minutes (both $p < 0.001$), and increased *confidence* by 2.22 and 2.00 points (both $p < 0.001$).

The remaining relevance tasks showed more selective effects. In *T3* (summary emphasis comparison), only *confidence* had a significant omnibus effect ($p = 0.0064$), with pairwise gains for SOURCE-GUIDED in *accuracy* (+27.78%, $p = 0.024$) and for SUMMARY-GUIDED in *confidence* (+1.06, $p = 0.0017$). In *T4* (most relevant document to the summary), SUMMARY-GUIDED improved *accuracy* (+33.33%, $p = 0.0039$) and *confidence* (+0.78, $p = 0.0072$), while SOURCE-GUIDED was slower (+0.50 minutes, $p = 0.048$). In *T5* (least relevant document to the summary), SUMMARY-GUIDED improved *accuracy* (+27.78%, $p = 0.0048$), whereas SOURCE-GUIDED was slower (+0.51 minutes, $p = 0.020$).

For sentence-level relevance, effects were mostly temporal. In *T6* (most relevant documents to a summary sentence), only *completion time* differed: SOURCE-GUIDED was slower than LINEAR (+0.64 minutes, $p < 0.001$), while SUMMARY-GUIDED did not differ ($p = 0.549$). In *T7* (most relevant summary sentence for a document), both guided layouts were faster than LINEAR (-0.53 and -0.65 minutes; $p = 0.0016$ and $p = 1.19 \times 10^{-4}$), with no significant effects for *accuracy* or *confidence*.

Takeaway 3. Guided layouts are most beneficial for relevance judgments in MDS verification: both layouts improve topic identification and topic coverage estimation, while SUMMARY-GUIDED provides the clearest advantage for judging document relevance to the summary.

The consistency tasks showed fewer significant condition effects. For *T8* (supporting documents for a summary sentence), there was no condition effect for *accuracy* ($p = 0.821$) or *completion time* ($p = 0.162$), but there was a significant effect for *confidence* ($p = 0.0135$). Only SUMMARY-GUIDED produced higher *confidence* than LINEAR, by 0.50 points ($p = 0.0052$). For *T9* (contradicted claim identification), the omnibus condition effects were not significant for any metric, although SOURCE-GUIDED showed a pairwise *accuracy* gain of 16.67 points over LINEAR ($p = 0.046$). For *T10* (unsupported claim identification), no significant condition effects were observed for *accuracy* ($p = 0.332$), *completion time* ($p = 0.485$), or *confidence* ($p = 0.873$).

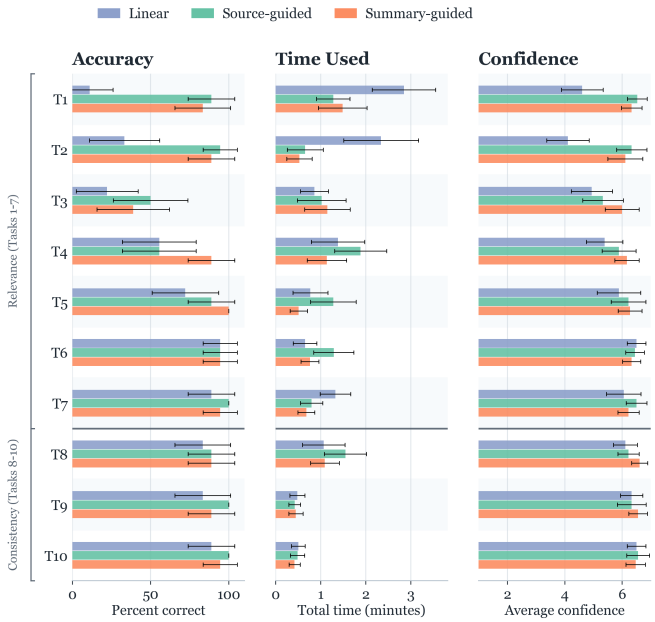

Figure 5: Task-level performance across the ten study tasks (T1–T10) for the three document organization conditions. Accuracy, completion time, and confidence are shown with 95% confidence intervals.

Takeaway 4. Consistency tasks are less sensitive to layout condition: unlike the relevance tasks, they show few significant differences across interfaces, with only isolated improvements such as higher confidence for SUMMARY-GUIDED in T8 and a modest accuracy gain for SOURCE-GUIDED in T9.

7.3 Qualitative Feedback

Participant feedback broadly aligned with the quantitative results, with many participants expressing a clear preference for SUMMARY-GUIDED. A recurring reason was that the layout made relevance to the summary easy to perceive directly from document position, while still supporting both overview and detailed inspection. Participants repeatedly described this condition as intuitive and easy to navigate.

"Top-left is most relevant . . . easy to look."

"The order from top to bottom correspond[s] [to] the summary sentence . . . [it] makes more sense."

"It has flexibility to let me see the whole and the parts in detail."

Takeaway 5. Participants preferred SUMMARY-GUIDED because document–summary relevance was immediately visible and supported both overview and detailed inspection in a single workflow.

Feedback on SOURCE-GUIDED revealed a different strength. Participants who preferred this layout often focused on its ability to reveal relationships within the source collection and provide additional spatial reference points for navigation.

"Quickly locate relevant document by the layout."

"It has summary sentence dot, so it give[s] us another reference point to see which documents . . . are close to the summary."

At the same time, several participants found SOURCE-GUIDED less intuitive for ranking documents by summary relevance.

"It was not intuitive for me to see which is the most relevant."

"The document layout was abstract for me to understand."

Takeaway 6. SOURCE-GUIDED was useful for understanding relationships among source documents, but less clear for judging which documents were most relevant to the summary.

In contrast, feedback on LINEAR emphasized the extra effort required to search and compare documents manually. Although some participants appreciated its simple ranked order, comments focused on the additional work needed to navigate the document set. A few

participants also explicitly contrasted the absence of richer interaction support in LINEAR with the guided layouts.

"For linear mode(layout), I have to do a lot of searching."

"The only good thing for the linear one is the linear order of documents that ranked by the relevance."

Takeaway 7. Compared with LINEAR, the guided layouts reduced manual search and made cross-document relevance relationships easier to understand.

8 DISCUSSION

Reliable Use of LLM-Generated Summary SVS supports more reliable use of generated summaries. With SVS, users can inspect whether a summary covers relevant sources and whether its claims are grounded in the source collection. Guided spatial layouts support relevance verification by showing which documents and topic regions matter for the summary and where source content may be underrepresented, reducing the risk that users miss coverage gaps or misplaced emphasis. Claim-level provenance supports consistency verification by decomposing summary sentences into claims and linking each claim to evidence spans, inference judgment labels, and source context.

Together, these views turn summary verification from open-ended reading into targeted inspection: spatial relevance guidance tells users where to look, while provenance tells them what evidence to judge once they get there. The usage scenarios illustrate how this workflow supports inspection of coverage and evidence, and the user study shows that guided layouts improved verification accuracy, efficiency, confidence, and workload over the LINEAR baseline. Thus, SVS does not make an LLM summary automatically reliable; instead, it supports more reliable use of generated summaries by making their relevance and grounding easier to verify.

Human-in-the-loop Auditing of LLM Outputs SVS supports human-in-the-loop auditing by placing users at key decision points rather than treating automated outputs as final answers. Users choose what to inspect, expand sentences into claims, compare claims with the original text, and examine each inference judgment label with its rationale, highlighted evidence, and source context. Model outputs serve as cues for further inspection, while users decide whether the summary is supported, contradicted, or under-supported.

SVS also helps users audit failure cases for LLM-as-a-judge in the verification pipeline. LLM-as-a-judge failures can occur at two stages: claim decomposition and claim-level inference judgment. For claim decomposition, the original summary sentence is shown alongside its generated atomic claims in the summary panel, allowing users to detect decomposition errors by checking whether the claims preserve the sentence's meaning (as shown in Fig. 1S1). For the claim-level inference judgment stage, each inference judgment label is linked to the judge's rationale, highlighted evidence spans, and source-document context (as shown in Fig. 1C). Users can therefore detect inference-judgment errors by checking whether the cited evidence actually justifies the assigned SUPPORTS, CONTRADICTS, or INSUFFICIENT EVIDENCE label, and whether the LLM-generated explanation is coherent and aligned with both the label and the visible source evidence. Thus, SVS does not treat LLM outputs as final ground truth. It presents decomposition and inference-judgment outputs as inspectable and contestable cues while final reliability judgments remain with the user.

Comparison with Alternatives Table 1 compares SVS with representative alternatives for summary verification. Compared with existing visualization systems designed for summary verification such as SummVis [47] and RTSUM [8], SVS extends the design space by combining spatial guidance and multi-document support with multi-level provenance visualization.

We also compare SVS with broader content-generation tools that also provide verification support. Direct LLM prompting can synthesize multiple sources, but lacks spatial guidance and persistent provenance tracking, making relevance and consistency issues hard to detect without repeated prompting and manual cross-checking. Commercial LLM tools such as NotebookLM [19] add citations and side-panel source access, but still require users to manually connect summary statements to evidence across documents. We use LINEAR as a controlled baseline for this common commercial provenance pattern, and make it a fair comparison by adding basic highlights for evidence sentences. The user study shows that SVS outperforms this baseline for relevance and consistency verification, suggesting that spatial organization adds support beyond source access and citations. More details are provided in Appendices A and B.

Table 1: Comparison of SVS with representative alternatives.

Method	Spatial guidance	Provenance visualization	Multi-document support
SummVis [47]	○	●	○
RTSUM [8]	○	●	○
Direct LLM prompting	○	○	●
NotebookLM [19]	○	○	●
AwesomeLit [50]	◐	◐	●
LINEAR baseline	○	◐	●
SVS	●	●	●

Note. Ratings summarize each method's degree of support for verification workflows: ○ = low, non-primary support; ◐ = medium, partial support; ● = high, explicit support.

AwesomeLit [50] represents recent research on agent-supported content generation across multiple sources, but provides limited spatial guidance and provenance visualization for verifying generated content. Although it supports efficient and grounded literature review generation, verification of generated content remains less guided than in SVS.

Across these comparisons, the distinction of SVS is the integration of spatial guidance, claim-level provenance visualization, and multi-document evidence inspection within one verification workflow.

Limitations & Future Work The methods in SVS have three limitations. First, the verification pipeline partly depends on the underlying LLM, since claim-level inference judgments and evidence links can vary with model quality. However, our goal is not to validate the LLM itself, but to make its verification outputs inspectable so users can check them against source evidence. Second, SVS focuses on evidence-centered verification through document cards and sentence-level highlighting, and future work could integrate richer document viewers and annotation support. Third, the workflow is primarily user-directed. Future systems could add transparent, contestable agent guidance for summary verification.

9 CONCLUSION

We presented SVS, a visual analytics system for verifying LLM-generated summaries over multi-document sources. SVS organizes evidence as a spatial provenance workspace. Its contribution lies in integrating spatial document organization and visual provenance tracking into a coherent workflow for summary verification.

Our study shows that this spatial framing provides measurable benefits over the LINEAR baseline. Both guided layouts increased overall accuracy by 36.0%, and SUMMARY-GUIDED provided the strongest overall performance: it reduced completion time by 32.7%, increased confidence by 11.9%, and lowered mental and physical demand by 47.2% and 42.3%, respectively. These results suggest that reliable use of generated summaries requires more than access to cited sources, and that verification benefits from visual structures that organize provenance around relevance and consistency checks.

SUPPLEMENTAL MATERIALS

Supplemental materials are available through our OSF project, including the supplemental video and source code. The source code is also available on GitHub, and an interactive demo is available at the project website.

ACKNOWLEDGMENTS

This material is based upon work supported in whole or in part with funding from the Department of Defense (DoD). Any opinions, findings,

conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DoD and/or any agency or entity of the United States Government.

REFERENCES

[1] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez et al. A brief survey of text mining: Classification, clustering and extraction techniques, 2017. doi: 10.48550/arXiv.1707.02919 1

[2] C. Andrews, A. Endert, and C. North. Space to think: large high-resolution displays for sensemaking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pp. 55–64. Association for Computing Machinery, New York, NY, USA, 2010. doi: 10.1145/1753326.1753336 2

[3] C. Andrews and C. North. The impact of physical navigation on spatial organization for sensemaking. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2207–2216, 2013. doi: 10.1109/TVCG.2013.205 2

[4] I. Arawjo, C. Swoopes, P. Vaithilingam, M. Wattenberg, and E. L. Glassman. ChainForge: A visual toolkit for prompt engineering and LLM hypothesis testing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, art. no. 304, 18 pp. Association for Computing Machinery, New York, NY, USA, 2024. doi: 10.1145/3613904.3642016 3

[5] D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67:1–48, 2015. doi: 10.18637/jss.v067.i01 7

[6] N. Cao, J. Sun, Y.-R. Lin, D. Gotz, S. Liu, and H. Qu. FacetAtlas: Multifaceted visualization for rich text corpora. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1172–1181, 2010. doi: 10.1109/TVCG.2010.154 2

[7] G. Y.-Y. Chan, E. Bertini, L. G. Nonato, B. Barr, and C. T. Silva. Melody: Generating and visualizing machine learning model summary to understand data and classifiers together, 2020. doi: 10.48550/arXiv.2007.10614 2

[8] S. Cho, M. Jang, J. Yeo, and D. Lee. RTSUM: Relation triple-based interpretable summarization with multi-level salience visualization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pp. 53–60. Association for Computational Linguistics, Mexico City, Mexico, 2024. doi: 10.18653/v1/2024.naacl-demo.5 2, 9

[9] C. Collins, S. Carpendale, and G. Penn. DocuBurst: Visualizing document content using language structure. *Computer Graphics Forum*, 28(3):1039–1046, 2009. doi: 10.1111/j.1467-8659.2009.01439.x 2

[10] J. DeYoung, S. C. Martinez, I. J. Marshall, and B. C. Wallace. Do multi-document summarization models synthesize? *Transactions of the Association for Computational Linguistics*, 12:1043–1062, 2024. doi: 10.1162/tacl_a_00687 2

[11] M. Dhaini, E. Erdogan, S. Bakshi, and G. Kasneci. Explainability meets text summarization: A survey. In *Proceedings of the 17th International Natural Language Generation Conference*, pp. 631–645. Association for Computational Linguistics, Tokyo, Japan, 2024. doi: 10.18653/v1/2024.inlg-main.49 1

[12] E. Durmus, H. He, and M. Diab. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5055–5070. Association for Computational Linguistics, Online, 2020. doi: 10.18653/v1/2020.acl-main.454 2

[13] A. Endert, P. Fiaux, and C. North. Semantic interaction for sensemaking: Inferring analytical reasoning for model steering. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2879–2888, 2012. doi: 10.1109/TVCG.2012.260 2

[14] A. Fabbri, I. Li, T. She, S. Li, and D. Radev. Multi-News: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1074–1084. Association for Computational Linguistics, Florence, Italy, 2019. doi: 10.18653/v1/P19-1102 6

[15] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021. doi: 10.1162/tacl_a_00373 2, 6

[16] S. Gabriel, A. Celikyilmaz, R. Jha, Y. Choi, and J. Gao. GO FIGURE: A meta evaluation of factuality in summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 478–487. Association for Computational Linguistics, Online, 2021. doi: 10.18653/v1/2021.findings-acl.42 1

[17] Y. Gao, W. Zhao, and S. Eger. SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1347–1354. Association for Computational Linguistics, Online, 2020. doi: 10.18653/v1/2020.acl-main.124 2

[18] D. Gholipour Ghalandari, C. Hokamp, N. T. Pham, J. Glover, and G. Ifrim. A large-scale multi-document summarization dataset from the Wikipedia current events portal. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1302–1308. Association for Computational Linguistics, Online, 2020. doi: 10.18653/v1/2020.acl-main.120 6

[19] Google. NotebookLM, 2026. `https://notebooklm.google/`. 6, 9

[20] D. Gotz and M. X. Zhou. Characterizing users' visual analytic activity for insight provenance. *Information Visualization*, 8(1):42–55, 2009. doi: 10.1057/ivs.2008.31 2

[21] M. A. Hearst. TileBars: visualization of term distribution information in full text information access. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '95, pp. 59–66. ACM Press/Addison-Wesley Publishing Co., USA, 1995. doi: 10.1145/223904.223912 2

[22] M. Honnibal, I. Montani, S. V. Landeghem, and A. Boyd. spaCy. Software, 2020. doi: 10.5281/zenodo.1212303 4

[23] Y. Huang, X. Feng, X. Feng, and B. Qin. The factual inconsistency problem in abstractive text summarization: A survey, 2023. doi: 10.48550/arXiv.2104.14839 1

[24] P. Isenberg and T. Isenberg. Visualization on interactive surfaces: A research overview. *i-com*, 12(3):10–17, 2013. doi: 10.1524/icom.2013.0020 2

[25] W. Kryscinski, B. McCann, C. Xiong, and R. Socher. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9332–9346. Association for Computational Linguistics, Online, 2020. doi: 10.18653/v1/2020.emnlp-main.750 1

[26] P. Laban, T. Schnabel, P. N. Bennett, and M. A. Hearst. SummaC: Revisiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177, 2022. doi: 10.1162/tacl_a_00453 2, 3, 5

[27] H. Li, Q. Dong, J. Chen, H. Su, Y. Zhou, Q. Ai et al. LLMs-as-Judges: A comprehensive survey on LLM-based evaluation methods, 2024. doi: 10.48550/arXiv.2412.05579 2

[28] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81. Association for Computational Linguistics, Barcelona, Spain, 2004. 2

[29] S. Liu, Z. Li, T. Li, V. Srikumar, V. Pascucci, and P.-T. Bremer. NLIZE: A perturbation-driven visual interrogation tool for analyzing and interpreting natural language inference models. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):651–660, 2019. doi: 10.1109/TVCG.2018.2865230 2

[30] W. Liu, C. North, and R. Faust. Visualizing spatial semantics of dimensionally reduced text embeddings, 2024. doi: 10.48550/arXiv.2409.03949 2

[31] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522. Association for Computational Linguistics, Singapore, 2023. doi: 10.18653/v1/2023.emnlp-main.153 2

[32] C. C. Marshall and F. M. Shipman. Spatial hypertext: designing for change. *Commun. ACM*, 38(8):88–97, 1995. doi: 10.1145/208344.208350 2

[33] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919. Association for Computational Linguistics, Online, 2020. doi: 10.18653/v1/2020.acl-main.173 1

[34] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform manifold approximation and projection for dimension reduction, 2020. doi: 10.48550/arXiv.1802.03426 4

[35] B. Pan, Y. Fu, K. Wang, J. Lu, L. Pan, Z. Qian et al. VIS-Shepherd: Constructing critic for LLM-based data visualization generation, 2025. doi: 10.48550/arXiv.2506.13326 2

[36] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of the 2005 International Conference on Intelligence Analysis*, 6 pp. The MITRE Corporation, McLean, VA, USA, May 2005. 2

[37] E. D. Ragan, A. Endert, J. Sanyal, and J. Chen. Characterizing provenance in visualization and data analysis: An organizational framework of provenance types and purposes. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):31–40, 2016. doi: 10.1109/TVCG.2015.2467551 2

[38] A. Sarikaya, M. Gleicher, and D. A. Szafir. Design factors for summary visualization in visual analytics. *Computer Graphics Forum*, 37(3):145–156, 2018. doi: 10.1111/cgf.13408 2

[39] F. M. Shipman III and C. C. Marshall. Formality considered harmful: Experiences, emerging themes, and directions on the use of formal representations in interactive systems. *Computer Supported Cooperative Work (CSCW)*, 8(4):333–352, 1999. doi: 10.1023/A:1008716330212 2

[40] J. Stasko, C. Görg, and Z. Liu. Jigsaw: Supporting investigative analysis through interactive visualization. *Information Visualization*, 7(2):118–132, 2008. doi: 10.1057/palgrave.ivs.9500180 2

[41] H. Strobelt, S. Gehrmann, M. Behrisch, A. Perer, H. Pfister, and A. M. Rush. Seq2Seq-Vis: A visual debugging tool for sequence-to-sequence models. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):353–363, 2019. doi: 10.1109/TVCG.2018.2865044 2

[42] S. Suh, B. Min, S. Palani, and H. Xia. Sensecape: Enabling multilevel exploration and sensemaking with large language models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, art. no. 1, 18 pp. Association for Computing Machinery, New York, NY, USA, 2023. doi: 10.1145/3586183.3606756 3

[43] S. Syed, T. Yousef, K. Al Khatib, S. Jänicke, and M. Potthast. Summary Explorer: Visualizing the state of the art in text summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 185–194. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021. doi: 10.18653/v1/2021.emnlp-demo.22 2

[44] X. Tang, E. Krokos, C. Liu, K. Davidson, K. Whitley, N. Ramakrishnan et al. Steering LLM summarization with visual workspaces for sensemaking, 2024. doi: 10.48550/arXiv.2409.17289 2

[45] I. Tenney, J. Wexler, J. Bastings, T. Bolukbasi, A. Coenen, S. Gehrmann et al. The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 107–118. Association for Computational Linguistics, Online, 2020. doi: 10.18653/v1/2020.emnlp-demos.15 2

[46] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. 4

[47] J. Vig, W. Kryscinski, K. Goel, and N. Rajani. SummVis: Interactive visual analysis of models, data, and evaluation for text summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pp. 150–158. Association for Computational Linguistics, Online, 2021. doi: 10.18653/v1/2021.acl-demo.18 2, 9

[48] A. Wang, K. Cho, and M. Lewis. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5008–5020. Association for Computational Linguistics, Online, 2020. doi: 10.18653/v1/2020.acl-main.450 2

[49] C. Wang, A. Jain, D. Chen, and J. Gu. VizSeq: a visual analysis toolkit for text generation tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pp. 253–258. Association for Computational Linguistics, Hong Kong, China, 2019. doi: 10.18653/v1/D19-3043 2

[50] Z. Xie, Y. Guo, and K. Xu. AwesomeLit: Towards hypothesis generation with agent-supported literature research. In J. Byska, A. Ottley, and M. Waldner, eds., *EuroVis 2026 – Short Papers*, 5 pp. The Eurographics Association, 2026. doi: 10.2312/evs.20261000 9

[51] Y. Zha, Y. Yang, R. Li, and Z. Hu. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11328–11348. Association for Computational Linguistics, Toronto, Canada, 2023. doi: 10.18653/v1/2023.acl-long.634 2

[52] H. Zhang, Y. Xu, and L. Perez-Beltrachini. Fine-grained natural language inference based faithfulness evaluation for diverse summarisation tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1701–1722. Association for Computational Linguistics, St. Julian's, Malta, 2024. doi: 10.18653/v1/2024.eacl-long.102 3

[53] H. Zhang, P. S. Yu, and J. Zhang. A systematic survey of text summarization: From statistical methods to large language models. *ACM Comput. Surv.*, 57(11), art. no. 277, 41 pp., 2025. doi: 10.1145/3731445 1

[54] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. BERTScore: Evaluating text generation with BERT, 2020. doi: 10.48550/arXiv.1904.09675 2

[55] Z. Zhang, W. Peng, X. Chen, L. Cao, and T. J.-J. Li. LADICA: A large shared display interface for generative AI cognitive assistance in co-located team collaboration. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, art. no. 147, 22 pp. Association for Computing Machinery, New York, NY, USA, 2025. doi: 10.1145/3706598.3713289 3