CAPE: LLM-Powered Context-Aware Projection Explanation of Document Embeddings

Category: Research

Paper Type: Technique

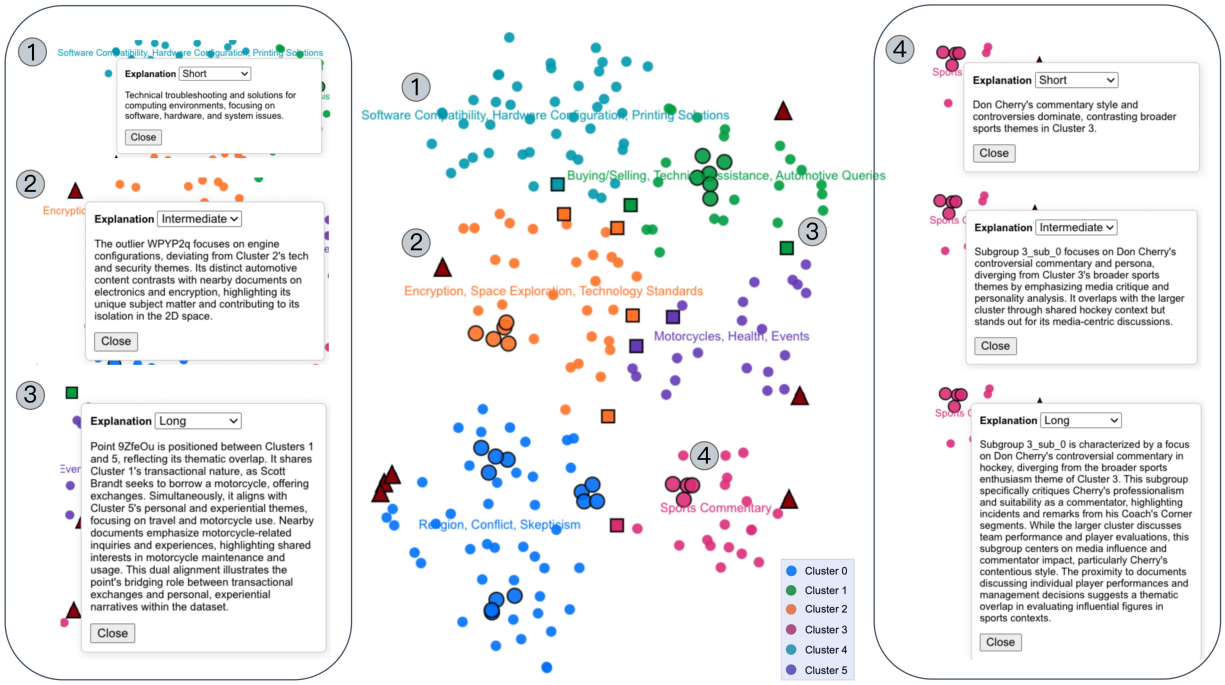

Fig. 1: CAPE in AI Explanation Mode on a t-SNE projection of the 20 Newsgroups dataset. CAPE automatically detects and explains meaningful spatial patterns—including clusters, outliers, bridging documents, and subgroups—with multi-level explanations (Short, Intermediate, Long) to support varied user needs. (1) Cluster summary on technical support topics. (2) Outlier focused on automotive content near tech-oriented documents. (3) Bridging document linking transactional and experiential themes (e.g., motorcycles, travel). (4) Subgroup centered on media critique of Don Cherry, diverging from general sports commentary.

Abstract— Dimensionality reduction is widely used to visualize high-dimensional text embeddings, enabling users to discover document clusters and relationships. However, these projections often lack interpretability, making it difficult to understand why documents are positioned in specific regions. We introduce CAPE (Context-Aware Projection Explanations), a visual analytics method that leverages Large Language Models (LLMs) to generate context-aware explanations for document embedding projections. CAPE integrates global, local, and pattern-specific contexts to explain structural patterns such as clusters, outliers, subgroups, and transitions. It synthesizes structured prompts based on both automatically identified spatial patterns and user-selected areas of interest. We demonstrate CAPE on two datasets—news articles and research papers—illustrating how it highlights thematic structures and explains document positions in the projection space. We also compare our approach to keyword-based explanation methods, showing that CAPE's context-rich explanations reveal deeper semantic relationships than purely term-centric baselines.

Index Terms—Dimension Reduction, Explainability, Text Embeddings, LLM

1 Introduction

Low-dimensional projections of high-dimensional text embeddings are widely used to explore the structure of document collections. Dimension Reduction (DR) methods project these embeddings into a two-dimensional (2D) space, enabling users to visually inspect clusters, similarities, and outliers in large text datasets [23]. Following the *proximity* $\approx$ *similarity* metaphor [4, 21], users often interpret the spatial proximity as reflecting semantic relationships between documents.

While these projections reveal patterns in unstructured text, they often lack interpretability due to the abstract nature of embeddings and the non-linear transformations used in DR methods [19, 25]. In practice, projections of large datasets can be visually complex or ambiguous. Sometimes, patterns such as clusters are distinct; other times, they are overlapping or unobvious. More importantly, it's often unclear why documents appear in specific positions. Users may ask: *What does this cluster represent? Why is this document isolated or bridging two groups? How are these regions related?*

Traditional post-hoc approaches, such as keyword-based labeling, attempt to explain documents or clusters by extracting representative terms [25]. However, these techniques focus only on content, and do not incorporate the spatial structure of the projection. Therefore, they often fail to explain how document positions relate to one another in the embedding projection space.

Recent work has explored spatial explanations in DR visualizations. For example, Bian et al. [4] introduced counterfactual-based techniques

for interpreting embedding changes in interactive DR. Liu et al. [19] proposed a gradient-based method to generate spatial word clouds, highlighting word-level impact on projection positions. Although these methods are effective for inspecting local word importance, they lack broader semantic context and do not support pattern-specific explanation, such as why clusters form, what makes a document an outlier, or how regions relate within the projection. Additionally, individual keywords can be generic or ambiguous, making it difficult to understand relationships between documents.

With the advancement of large language models (LLMs) like ChatGPT, there is increasing potential for generating rich, contextual explanations through natural language [37]. However, directly applying LLMs to projection-based visualizations is challenging. LLMs do not inherently understand spatial structure, and naïve summarization can lead to vague or misleading outputs. Furthermore, the input limitation of LLMs makes it impossible to process entire datasets. Therefore, it requires a more guided and context-aware approach to integrate LLM reasoning into visualization workflows.

In this paper, we introduce CAPE (Context-Aware Projection Explanations), a method for generating spatially grounded, multi-level explanations of document embedding projections. Instead of just summarizing documents, CAPE explains why patterns appear in the projection, based on both semantic content and spatial context.

To do this, CAPE dynamically constructs prompts for the LLM using contextual metadata at multiple levels: (1) global cluster relationships, (2) local neighborhood information, and (3) pattern-specific metadata. CAPE supports two modes of interaction: **(1) AI Explanation Mode**, which automatically detects and explains interesting patterns in the projection; **(2) User Exploration Mode**, which enables on-demand explanations for selected documents or regions.

The contributions of this paper include:

- **A context-aware explanation framework** for document embedding projections that integrates spatial patterns with semantic context.
- **A multi-level prompting strategy** that combines global, local, and pattern-specific context to guide LLM-based explanations grounded in projection space.
- **An interactive prototype** of CAPE that supports both auto-detected and user-driven explanation workflows.
- **Two usage scenarios and a comparison evaluation** with keyword-based explanation baselines, demonstrating how CAPE enables deeper and more contextually grounded interpretation of embedding projections.

2 Related Work

2.1 Dimensionality Reduction in Text Visualization

DR techniques are often used to visualize and explore high-dimensional data, including text embeddings. These techniques can be broadly classified into linear and non-linear methods. Linear methods, such as PCA, are typically used for datasets with linear relationships [30]. However, textual data often exhibit complex, non-linear relationships that linear methods cannot effectively capture [5]. To address this, nonlinear algorithms such as t-distributed Stochastic Neighbor Embedding (t-SNE) and UMAP have become popular [22, 34]. These methods efficiently maintain local data structures and excel in revealing clusters of text data [33]. They enable analysts to identify meaningful patterns in the projection. However, the resulting projections are often difficult to interpret: it is not clear why documents appear in specific locations.

Recent work has explored interactive visual analytics methods to help users explore projections. INSPIRE was one of the first text projection systems to include interactive features such as filtering [36]. For more advanced examples, Intent Radar introduces interactive intent modeling to support exploratory search by visualizing search intents in a radial layout [27]. Typograph visualizes topics, keywords, and snippets to support multi-scale exploration within a single spatial projection [9]. StarSPIRE uses semantic interaction to let users adjust the document positions, and update the layout accordingly to reflect inferred intent [6]. Similarly, Cosmos combines spatial projection and a computational model for human-in-the-loop sensemaking [8]. DeepSI further incorporates deep learning into the sensenmaking loop to support semantic interaction inference [3]. Although these interactive visual methods support exploration, they lack explainability: they do not provide explanations for the document position. Our work aims to bridge the gap by introducing a framework that generates spatially grounded explanations of document embedding projections.

2.2 Explainability in Text Embedding Projections

Prior work has used a variety of document representation methods to explain the results of document clustering or classification [28]. Many of these approaches rely on term-based representations, including Term Frequency and Inverse Document Frequency(TF-IDF), Latent Dirichlet Allocation (LDA), and n-gram language models [28]. They use representative keywords or topics to characterize document clusters, providing high-level summaries of document groups.

Post-hoc explanation approaches, including word clouds, keyphrase labeling, and term frequency analysis, are commonly used to summarize the document collections [15, 24]. For example, TopicPanorama [18] shows the document topics for the large corpus in a full picture. Scattertext [16] visualizes words and phrases in a 2D scatterplot to support comparative text analysis. Although these visual representations provide useful summaries, they rely primarily on keyword frequency and do not account for the spatial position of documents in projection spaces. Therefore, they do not explain the semantic or structural relationships between documents in the layout.

In explainable AI, counterfactual explanations and feature contribution methods are popular approaches. Counterfactual explanation, such as What-IF Tool [35], ViCE [13], AdViCE [14], allows users to explore the change of input features to the model's output. Feature contribution methods, including LIME [26] and SHAP [20], estimate the importance of input features to the output. Although these approaches are useful for explaining individual document classifications, they are not designed to interpret embedding projections.

Recent work has also explored gradients for understanding projection layouts. The methods evaluate how sensitive a projection point is to its input features. For example, DimReader [11] analyzes how input feature perturbations affect point positions, and uses generalized axes to show the influences of individual dimensions. Similarly, Liu et al. proposed a gradient-based method to assess the sensitivity of projected document positions to specific words in the input text [19]. This approach provides insight into the local impact of words on DR outcomes. However, it focuses on word-level impact and does not provide explanations for some patterns in projection space, such as bridging points or subgroups. Instead of focusing on terms or single documents, CAPE grounds the projection space and explains why documents or regions are located where they are.

2.3 LLM-based Summarization and Explanation

Recent advances in large language models (LLMs), such as GPT-3, GPT-4, have shown big potential in text summarization, semantic analysis, and natural language explanation in a wide range of tasks [31, 37]. These models have been integrated into interactive systems to support storytelling, sensemaking, and visual analytics [32].

In visual analytics, LLMs have been used to support data analysis tasks, such as summarizing trends, generating comparisons, or answering user queries about structured datasets [2]. For example, recent tools have been using LLMs to help users interpret the charts, plots, and tables [7, 29]. However, using LLMs for interpreting spatial data, such as 2D projections of high-dimensional embedding remains limited. Existing work often focuses on small-scale datasets or predefined tasks, and few explore how to guide LLMs in understanding large, complex spatial layouts produced by DR methods [12].

Different than structured tables or plots, embedding projections do not have clear axes, labels, or feature names, making it difficult for LLMs to directly interpret them without additional context. Therefore, it needs a method to incorporate both spatial and semantic context to guide the LLM to generate the meaningful response. CAPE addresses

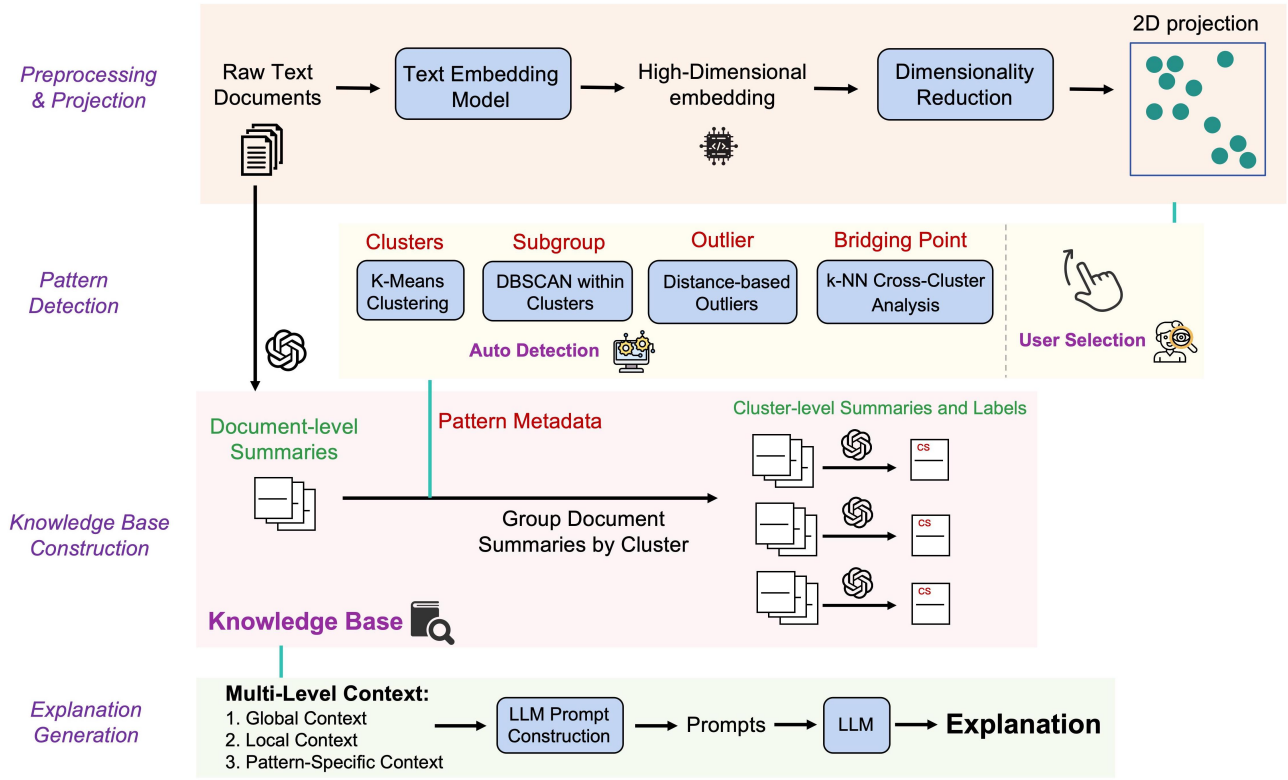

Fig. 2: CAPE explanation pipeline. (1) Raw text documents are embedded and projected into 2D using DR. (2) CAPE automatically detects spatial patterns such as clusters, subgroups, outliers, and bridging documents. Users can also manually select regions of interest. (3) Document summaries, cluster summaries, and pattern metadata are organized into a structured knowledge base. (4) CAPE retrieves relevant content and synthesizes multi-level context (global, local, pattern-specific) to construct prompts and guide the LLM in generating context-aware explanations.

this by constructing multi-level, context-aware prompts that incorporate global structure, local neighborhoods, and pattern-specific details. This enables LLMs to generate explanations that are grounded in both spatial and semantic context.

3 Formative Study & Design Goals

To better understand how users explore and interpret 2D projections of text embeddings, and what challenges or needs arise in this process, we conducted a formative study. Our goal was to observe how users naturally identify and annotate patterns in text embedding projections, and to gather their feedback on the interpretability of such visualizations.

3.1 Formative Study Setup

We conducted the study with three graduate students: two in computer science and one in statistics. The participants had different levels of familiarity with text embedding-based visualizations, ranging from occasional experience to frequent use in prior research.

3.1.1 Materials and Interface

We used a subset of 216 documents from the 20 Newsgroups dataset [17], ensuring that each subcategory had sufficient documents to maintain topical diversity. Documents were embedded using OpenAI's `text-embedding-3-small` model and projected into a 2D space using t-SNE. Each document was represented as a single point in the scatterplot. In the interface, users could hover over a point to view the document title, click a point to open the full document content, and use annotation tools (free drawing, arrows, text labels) to record observations.

3.1.2 Procedure

The study included three phases:

Introduction (10 min) Participants were introduced to the interface and informed that each point represented a document. They were asked to explore the arrangement, and observe their natural interpretation of the space.

Exploration & Annotation (30 min) Participants freely explored the projection, identifying interesting patterns, relationships, or structures, and providing explanations for these patterns based on the document content. The process used a think-aloud protocol, where participants verbalized their reasoning while navigating the projection.

Discussion & Feedback (20 min) After exploration, we asked participants about difficulties, desired support tools, and thoughts on potential use of AI-driven explanations.

3.2 Challenges & Needs

We qualitatively coded participants' verbalizations and feedback to extract themes. These were then grouped into four key categories of challenges and needs.

3.2.1 Challenges Navigating and Interpreting the Projection

All participants expressed difficulty analyzing a large number of documents in the projections space. The process of reading individual documents, connecting them to others, and reasoning about relationships was time-consuming and overwhelming. One participant noted:

> "It's not easy… I'm getting lost. I don't know what I am looking for." "I have to check them one by one... (otherwise I might miss a subtle difference)."

Because of this high cognitive load, participants spent nearly the entire exploration time identifying broad clusters and groupings. They needed to first understand these topics to get a general sense of the dataset before they could delve into deeper insights about individual

documents, even though they expressed strong interest in doing so. These types of patterns were also difficult to interpret without additional support.

Additionally, participants expressed a need for "hints" such as cluster labels or subgroup characteristics, to help them understand why certain regions formed in the projection. This emphasizes the need for structured guidance, such as auto-generated labels and visual highlights, to help users navigate and interpret the projection more effectively.

3.2.2 Need for Concise & Expandable Information

Participants wanted concise, one-sentence summaries for each document, rather than relying solely on titles or full document content. They felt this would help them quickly grasp each document's topic, while still wanting the option to expand for more details when needed. These findings indicate that explanation should be multi-level, providing concise summaries by default but allowing users to expand for more details as needed.

3.2.3 Need for On-Demand Explanations

While participants considered broad topic keywords helpful for orientation, they also wanted interactive explanations on demand. They sought an interactive approach, for example, where selecting a region or set of dots would generate short explanations describing the characteristics of the selected group(s). These findings suggest the need for an interactive explanation mode that allows users to dynamically request insights, which aligns with their different needs.

3.2.4 Need for Contextually Relevant Explanations

Participants wanted explanations that not only described what a document was about, but also helped them understand why it appeared in a particular position and and how it related to other documents in the projection. This shows that explanation with context—both spatial and semantic—is important for interpretation.

For example, participants often pointed to outliers or bridging points as particularly interesting, but hard to make sense of without additional support. One participant noted:

> "Outliers are important." "It is hard to see the connections (especially when the topic boundaries were blurry or overlapping)."

While these points were considered meaningful, their relationship to nearby groups was often unclear. This highlights the importance of generating explanations that are grounded in the actual context of the projection space. Rather than just generic description, the explanation should meaningfully connect to the data and structure.

3.3 Design Goals

Based on the challenges and user needs identified in the formative study, and the insights from prior literature, we define four design goals for CAPE (Context-Aware Projection Explanations) to help users understand, interpret, and explore text embedding projections.

DG1: Highlight and Explain Key Patterns in the Projection Participants are interested in understanding why certain patterns form in the 2D projection, including clusters, subgroups, outliers, and bridging documents. However, they found it difficult to determine why documents were positioned in specific locations and how different groups or individual points relate to each other (Sec. 3.2.1). CAPE should automatically detect, highlight, and explain these key patterns to enhance interpretability. This also reduces the cognitive burden of manually discovering and interpreting relationships throughout the projection.

DG2: Support Different Levels of Explanation Detail Participants sought explanations that were both concise and expandable - they wanted quick insights at a glance but also the option to dive deeper when needed (Sec. 3.2.2). Therefore, CAPE should provide explanations at different levels of detail, allowing users to explore insights at varying depths based on their preferences or goals.

DG3: Enable Interactive and On-Demand Explanations Participants expressed a strong need for on-demand, user-driven exploration (Sec. 3.2.3). CAPE should allow users to request explanations for individual documents, a selected group of documents, or multiple regions. This supports flexible exploration and lets users tailor the explanation process to their own analysis needs.

DG4: Generate Spatially and Semantically Grounded Explanations Since CAPE uses LLMs to generate explanations, it is important to guide the model outputs so they are grounded in actual document content and spatial context. Explanations should aim to reflect the semantic and structural reality of the projection space, avoiding vague or generic responses (Sec. 3.2.4). CAPE should construct prompts using targeted, context-aware information to guide the LLM toward meaningful responses.

These design goals guided the development of CAPE, shaping both the explanation generation pipeline (Sec. 4) and interactive visual design (Sec. 5).

4 CAPE Method

CAPE is a context-aware explanation framework for 2D text embedding visualizations. It combines embedding projection, automatic pattern detection, and LLM-powered explanations to help users interpret semantic structures in the projection space. Fig. 2 illustrates CAPE explanation pipeline, which includes four main stages: (1) projection of text embeddings into 2D space, (2) detection of spatial patterns, (3) construction of a structured knowledge base, and (4) generation of context-aware explanations.

4.1 Embedding & Projection

As in Sec. 3.1.1, we use ChatGPT's `text-embedding-3-small` to generate high-dimensional embeddings for each document. This model, based on GPT architecture, captures the documents' semantic content, producing embeddings that reflect high-level conceptual meaning. It has been widely used for various NLP tasks, making it suitable for visual-semantic analysis.

To visualize these embedding, we use DR techniques, such as t-SNE or UMAP to project the high-dimensional embeddings into a 2D space. These DR methods are known for preserving local structures, making them effective for revealing neighborhood patterns and clusters in document collections. Each document is represented as a single point in the 2D projection. The resulting projection serves as the basis for visual pattern detection, and explanation generation. The choice of embedding model and DR technique can be adjusted based on users' needs or domain preferences.

4.2 Pattern Detection

To reduce cognitive load and support sensemaking (**DG1**), CAPE automatically detects meaningful spatial patterns: clusters, outliers, bridging (in-between) documents, and subgroups. These are identified using machine learning techniques that approximate how a human might naturally notice structures in visualizations. While these techniques are not guaranteed to be perfectly accurate, they serve as starting points and can be replaced or adjusted based on the user's needs.

Cluster Assignment Documents are grouped using K-means clustering based on their 2D positions. Each cluster is treated as a high-level semantic group. We compute the centroid of each cluster by averaging the positions of all documents within it. These clusters form the basis for understanding thematic regions in the projection.

Outlier Detection Outliers are defined as documents whose 2D positions are significantly distant from their assigned cluster's centroid. For each cluster, we calculate the Euclidean distance between each document and the cluster centroid, and label the top-$k\%$ (e.g., 3%) farthest document as outliers. These often represent unique content or documents with weak semantic ties to the rest of the group.

Bridging Documents We define bridging points (or "in-between documents") as those that spatially lie between two or more clusters. Using k-nearest neighbor (k-NN) analysis in the 2D space (e.g., $k = 3$), we check if a document's nearest neighbors belong to more than one cluster. If so, the document is marked as a bridge, often indicating topic transitions or shared themes.

Subgroup Detection To identify more fine-grained patterns within clusters, we apply the density-based clustering algorithms (e.g. DBSCAN) within each K-means cluster to detect subgroups (smaller, dense regions). Subgroup centroids are computed by averaging the positions of documents within each subgroup. These subgroups often reflect subtopics or more focused variations within a broader theme.

Nearest Neighbors (for Contextual Summaries) To support local context construction, we retrieve the n nearest documents (e.g. $n = 8$) for each identified pattern using a KD-tree over the 2D projection. These are used later to generate context-aware explanations.

Users can also manually select patterns or regions of interest within the projection space. Each identified pattern is stored with metadata, including text summaries, projection positions, distances, cluster affiliations, and surrounding documents. This pattern-level metadata supports the explanation generation pipeline described next.

4.3 Context-Aware Explanation Generation

To generate meaningful and semantically grounded explanations, we introduce a context-aware prompting strategy for LLMs (e.g. GPT-4o). While LLMs like ChatGPT are powerful at synthesizing textual insights, they do not inherently understand spatial structures such as 2D embeddings. In addition, feeding the entire text dataset into the model is impractical due to input length limits and scalability concerns.

To address these challenges, our method dynamically constructs multi-level contextual prompts that reflect both the semantic and spatial characteristics of identified projection patterns (**DG4**). As illustrated in Fig. 2, the explanation generation process leverages a structured knowledge base of summaries and pattern-level metadata to compose targeted prompts that guide the LLM toward producing contextually grounded explanations.

4.3.1 Knowledge Base Construction

Before generating explanations, we build a knowledge base that includes:

Document-level summaries, generated using LLM with prompts like: *"Please generate a single concise sentence that represents the main idea of the following text"*

Cluster-level summaries and 1–3 representative keywords for each cluster, to capture global semantics.

These summaries are used to avoid overloading the model with full document texts while still preserving key semantics. They also support user needs observed in our formative study (Sec. 3.2.2).

In addition, we record **metadata for projection patterns** including clusters, outliers, bridging points, subgroups, and individual documents. For each pattern, we store associated document summaries, 2D projection coordinates, nearest neighbors (along with their summaries), and cluster properties (size, centroid position, summary, and label).

This structured information forms an extensible knowledge base that serves as a foundation for generating prompts dynamically in response to auto-identified patterns or user-selected regions.

4.3.2 Dynamic Prompt Construction

When a pattern is detected or when a user selects documents in the projection space, CAPE dynamically constructs a prompt by integrating relevant information from three contextual levels:

(1) Global context includes high-level summaries and statistics that position the pattern within the broader projection space. It includes cluster labels, sizes, centroid positions, and summaries - both for the target cluster and its neighboring clusters (if applicable).

(2) Local context draws from nearby documents (using k-NN search, Sec. 4.2) in the 2D projection, along with their corresponding summaries.

(3) Pattern-specific context is tailored to the type of pattern. In addition to the general projection metadata described in Sec. 4.3.1, it includes metrics such as the outlier distance from its cluster centroid, bridging points' proximity to neighboring clusters, and subgroup-to-cluster relationships.

By combining these contextual layers, CAPE ensures that each prompt provides the LLM with sufficient semantic grounding while staying within input limits and avoiding irrelevant noise.

4.3.3 Explanation Prompting and Output

CAPE formulates a task-specific prompt based on either auto-detected patterns or user-selected regions, supporting **DG1** and **DG3**:

For auto-detected patterns:

a. Clusters: Explain what defines the cluster, including common themes, distinguishing characteristics, and how it relates to nearby clusters.

b. Subgroups: Describe how the subgroup differs from its parent cluster while still being related.

c. Outliers: Provide insights into why a document is positioned far from clusters, such as whether due to a unique topic, or weak semantic ties to other documents.

d. Bridging Documents: Explain how a document serves as a "bridge" between multiple clusters, indicating potential shared themes or conceptual links.

For user-selected patterns:

a. Single document: Clarifying why a document is positioned in a certain area.

b. Group of Documents: Summarizing the characteristics or dominant themes of a selected group.

c. Multi-group Comparison: Highlighting key similarities and differences when users select multiple regions.

4.3.4 Explanation Output Levels

To support different user needs and interaction scenarios (**DG2**), CAPE produces explanations at three levels of detail:

(1) Concise Overview: A short, high-level summary of the identified pattern, providing an immediate understanding of its core concept or important character in a single sentence.

(2) Intermediate Explanation: A more descriptive explanation that includes additional context, such as nearby clusters, document relationships, or key differences.

(3) In-Depth Analysis: A more comprehensive explanation that examines both semantic and spatial details. It may include how the pattern connects to broader themes, why it appears in its specific location, and what insights it reveals about the dataset.

5 Visualization Design

To help users understand and explore complex document projections, we designed an interactive interface that enables both guided and exploratory explanation workflows: **AI Explanation Mode**, which automatically detects and explains spatial patterns in the projection; and **User Exploration Mode**, which allows users to select documents or regions and request on-demand explanations.

All explanation outputs are grounded in the multi-level context described in Sec. 4.3, incorporating global context, local neighbors, and pattern-specific metadata. This helps ensure that explanations are relevant, and meaningful connected to the projection space, supporting **DG4**.

To support different user preferences and analysis needs, each explanation is offered at three levels of detail—Short, Intermediate, and Long, in line with **DG2**.

While both modes share the same 2D projection layout, they differ in interaction mechanics, visual encodings, and the types of explanations provided.

5.1 AI Explanation Mode

This mode automatically detects and highlights key projection patterns, including clusters, outliers, bridging points, and subgroups (Sec. 4.2). It provides a structured overview of the embedding space by presenting visually the semantically meaningful regions, helping reduce cognitive effort and supporting **DG1**.

As shown in Fig. 1, the 2D scatterplot displays each document as a point, projected from high-dimensional ChatGPT embeddings using

the DR method (Sec. 4.1). Patterns are visually differentiated using color, shape, and styling encodings:

- **Clusters**: color-coded circular markers based on K-means cluster; keyword labels placed at cluster's centroid, matching the cluster color.
- **Outliers**: triangular markers in dark red, regardless of cluster color.
- **Bridging points**: square markers with the same color as their assigned cluster.
- **Subgroups**: larger circular markers with a bold black outline.

Users can hover over any point to see a tooltip with document's title and a one-sentence summary. Clicking on any of the pattern markers (e.g., outlier, bridge, or subgroup point, or cluster label) opens an explanation panel with a context-aware explanation for that pattern. For subgroups, clicking on any document in the group shows the shared explanation for the entire subgroup.

AI Explanation Mode provides a guided overview of the projection space, helping users understand the structure through LLM-enhanced semantic explanations.

5.2 User Exploration Mode

The user exploration mode enables on demand interaction, allowing users to select specific documents or regions of interest and request explanations. The projection layout remains the same as in AI explanation mode, but all points are colored in neutral blue, indicating that no auto-detected pattern highlighting is active (Fig. 3). This supports **DG3** by enabling user-driven inquiry.

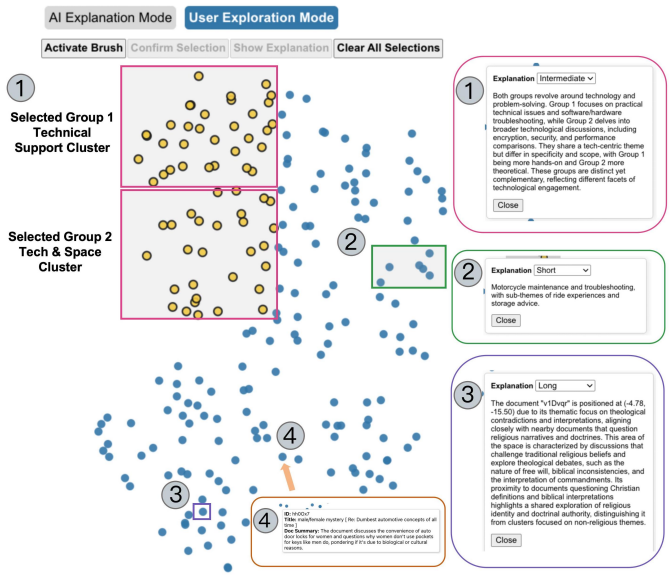

Fig. 3: CAPE in User Exploration Mode on a t-SNE projection of the 20 Newsgroups dataset. Users interactively select documents or regions to request explanations. (1) Comparison between two tech-related clusters, contrasting practical support topics with broader technological themes. (2) A group-level explanation summarizing motorcycle-related discussions. (3) A long explanation for a single document focused on theological contradictions, aligning with nearby religious content. (4) Tooltip showing the title and summary of a selected document on hover.

Users can explore the projection through two primary types of interaction:

Single-document explanation: Hovering shows a tooltip (title and summary); clicking generates an explanation of why that document appears in its current position in the projection space.

Region-based exploration and comparison:

1. Clicking *Activate Brush* enables brush selection
2. After selecting a group of documents, the user clicks *Confirm Selection* to finalize the group
3. Users may repeat this process to select additional groups.
4. Once all groups are selected, clicking *Show Explanation* triggers tailored explanations.
5. Clicking *Clear All Selections* resets the interface to its initial state.

These interactions support flexible exploration and hypothesis-driven analysis. Depending on the selection, CAPE generates one of three explanations types (See Fig. 3): a single document's positioning; the characteristics of a document group; and key similarities and differences between multiple selected regions.

6 Usage Scenarios

To illustrate how CAPE supports real-world exploratory analysis, we present two usage scenarios and a comparison with the keyword-based baseline.

6.1 Exploring Themes in 20 Newsgroups Dataset

In this user scenario, we illustrate the application of CAPE to a subset of 216 documents from the 20 Newsgroups dataset (also used in Sec. 3.1.1). Using AI explanation mode (Fig. 4 (1)), we began by examining a region of the projection space labeled around religious and ideological contradictions. CAPE automatically identified three subgroups within this cluster: one on societal biases against homosexuality, another on theological contradictions within Christianity, and a third on governmental overreach and the Waco siege. The subgroups were spatially localized and semantically differentiated. The explanations provided by CAPE aligned with the content of the documents: discussions ranged from theological interpretation and scriptural contradiction to critiques of social norms and institutional authority. These insights help us understand not only the general theme of the group, but also the diverse perspectives within it, which may be harder to distinguish using simpler clustering approaches.

Next, we explored Clusters 4 and 2, which appeared spatially close but represented semantically different topics. CAPE's cluster labels revealed that Cluster 4 emphasized technical support and troubleshooting, while Cluster 2 focus on technological advancement, encryption, and space exploration. To better understand the relationship between them, we used User Exploration Mode to select both clusters for comparison. As shown in Fig. 3 (1), CAPE explained that while both clusters involved technology, they differ in scope: Cluster 4 centered on immediate, practical issues, such as legacy system compatibility and troubleshooting hardware, whereas Cluster 2 engaged with broader technological themes, including encryption security, multimedia standards, and space exploration. Sampling document summaries in each cluster confirmed this distinction (Fig. 4 (2)): Cluster 4 included "troubleshooting Windows 3.1" and "configuration issues." while Cluster 2 featured topics like "Clipper chip vulnerabilities" and "DES keysearch." This difference was further reinforced by bridging documents. One example, `13t78w`, discussed source code acquisition and software modifications, positioned between two clusters (Fig. 4 (3)). CAPE's explanation highlighted the shared themes of technical development and practical support, and the document summary confirmed the interpretation.

Toward the right side of the projection, we examined Clusters 1 and 5, which CAPE described as reflecting more personal and practical themes: product inquiries, travel, event planning, motorcycle maintenance, and health concerns. These clusters share a soft boundary in some regions. In AI explanation mode, we explored document along this boundary to examine their ambiguity. One example, `9Zfe0u`, discussed borrowing a motorcycle for a trip (Fig. 4 (4)). CAPE's explanation indicated a combination of Cluster 1's transaction and Cluster 5's personal experience themes. The summary supported this interpretation by highlighting elements of travel planning, vehicle use, and informal exchange.

Using User Exploration Mode, we then selected a small group of documents below this bridging point Fig. 3 (2). CAPE revealed a

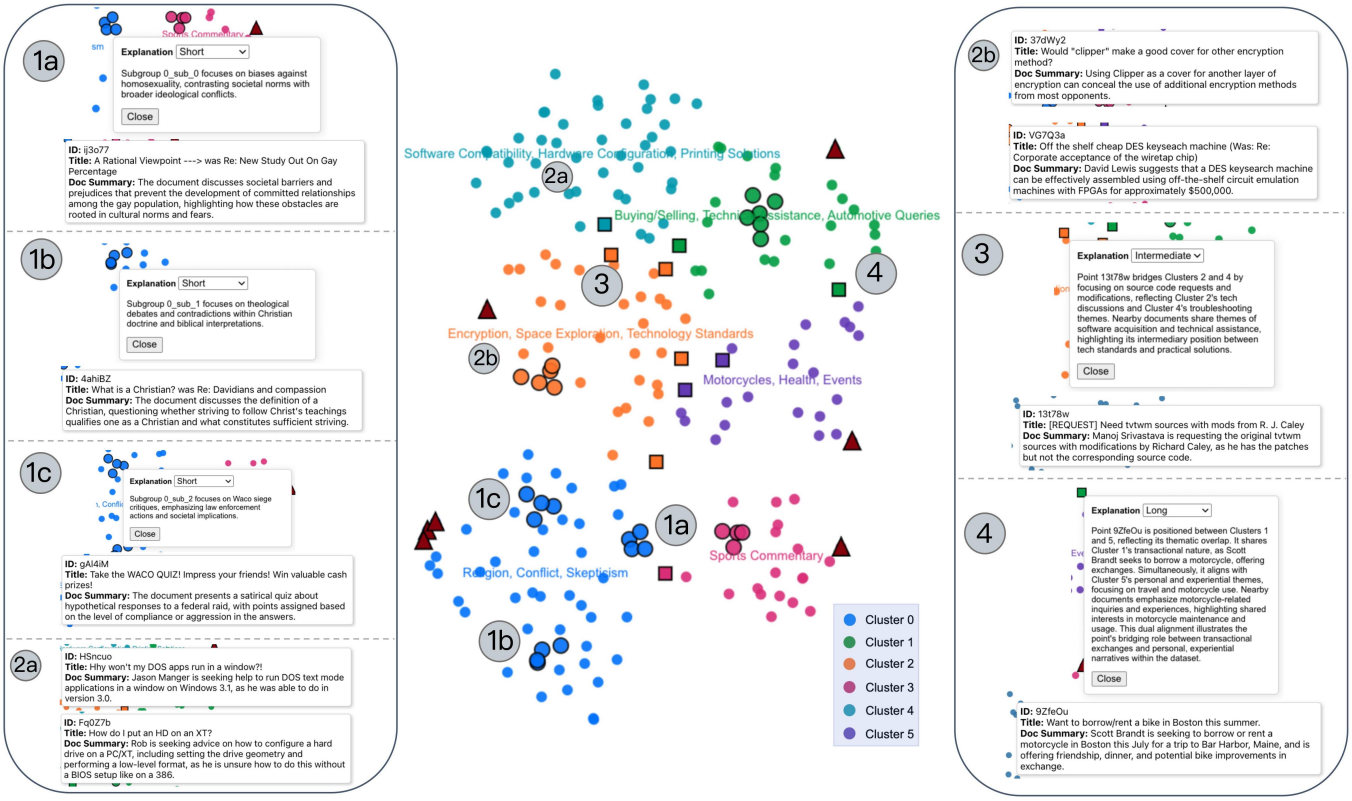

Fig. 4: Interpreting thematic structure in 20 Newsgroups Dataset with CAPE. CAPE in AI Explanation Mode highlights key spatial patterns in the projection space. (1) Three subgroups within the Religion cluster focus on different subtopics, each illustrated by one representative document. (2a) Example documents from the Technical Support cluster cover system troubleshooting and configuration issues. (2b) Example documents from the Tech & Space cluster discuss encryption techniques. (3) A bridging document positioned between Technical Support and Tech & Space Clusters, linking software modification with technical support. (4) Another bridging document connects Automotive & Transactions cluster and Lifestyle & Health cluster, combining transactional intent with personal travel and motorcycle usage.

distinct subgroup focused on motorcycle-related topics, which stood apart from nearby health-realted documents. Its group-level explanation referenced motorcycle experiences and maintenance, which aligned closely with the underlying document summaries and validating the subgroup's thematic focus.

6.2 Exploring Research Trends in IEEE VIS 2022–2023

In this scenario, we used CAPE to explore 252 papers published in IEEE VIS 2022 and 2023 with UMAP-projected text embeddings, aiming to understand emerging research directions and thematic relationships across subfields.

Under AI Explanation Mode, we began in the lower right region of the projection space, where CAPE identified Cluster 4 as focused on scientific visualization. Its explanation highlighted topics deep learning methods, topological analysis, and parallel processing techniques (Fig. 5 (1a)). Reviewing document summaries confirmed this: papers discussed neural super-resolution for volume data, real-time rendering frameworks, and topological methods for visualizing complex structures. CAPE further identified two subgroups within the cluster (Fig. 5 (1b, 1c)), corresponding to different methodology focuses—one emphasizing neural optimization, the other topological analysis.

We then moved toward Cluster 0, a nearby region CAPE labeled as focusing on innovative visualization frameworks. Unlike Cluster 4's computational emphasis, Cluster 0 focused on tools for understanding complex data and models—e.g., urban analytics, neural networks, and interactive model explainability (Fig. 5 (2a)). The subgroup `0_sub_0` in Cluster 0 centered specifically on tools enhancing interpretability in neural models, reinforcing the focus (Fig. 5 (2b)). While the two clusters were close spatially, CAPE's explanations revealed their thematic divergence (Fig. 6 (2)): one grounded in innovative visualization tools, the other in performance-driven rendering. A bridging document, `hDNxVf`, linked these clusters, discussing domain-specific latent representations for efficient scientific visualization. CAPE's explanation and the document summary confirmed this dual relevance (Fig. 5 (2c)).

To explore broader user-centric topics, we moved to the lower-left of the projection, examining Cluster 6, which CAPE described as emphasizing inclusivity, emotional engagement, and educational tools (Fig. 5 (3a)). This cluster shared boundaries with Cluster 3 (focused on cognitive and affective considerations), and Cluster 8 (focused on interactive systems and user-centered design). Subgroup `6_sub_0` in cluster 6 highlighted themes around aesthetics and emotional engagement, reinforcing Cluster 6's character (Fig. 5 (3b)).

Using User Exploration Mode, we selected documents from Clusters 6 and 3 for comparison. CAPE's explanation identified a thematic transition from emotional design and educational tools (Cluster 6) to trust, bias, and perceptual impact (Cluster 3) (Fig. 6 (2)). For example, one bridging point `xF9jPX` investigated how integrating text with charts affects reader interpretation. CAPE's explanation suggested shared themes around accessibility and cognition (Fig. 5 (4)). This insight was not obvious from cluster labels alone but became clear through the projection highlights and CAPE's contextual explanation.

Finally, we explored Cluster 2, which focused on domain-specific applications of immersive and interactive visual analytics, particularly in sports, e-commerce, and public speaking (Fig. 5 (5a)). A subgroup (`2_sub_0`) specifically discussed sports analytics (Fig. 5 (5a)), distinguishing it from neighboring clusters that also dealt with real-time interaction and immersive systems. Compared with Cluster 10, which emphasized physical and immersive visualization, Cluster 2 tended toward domain applications.

Throughout our analysis, CAPE's explanations helped us interpret

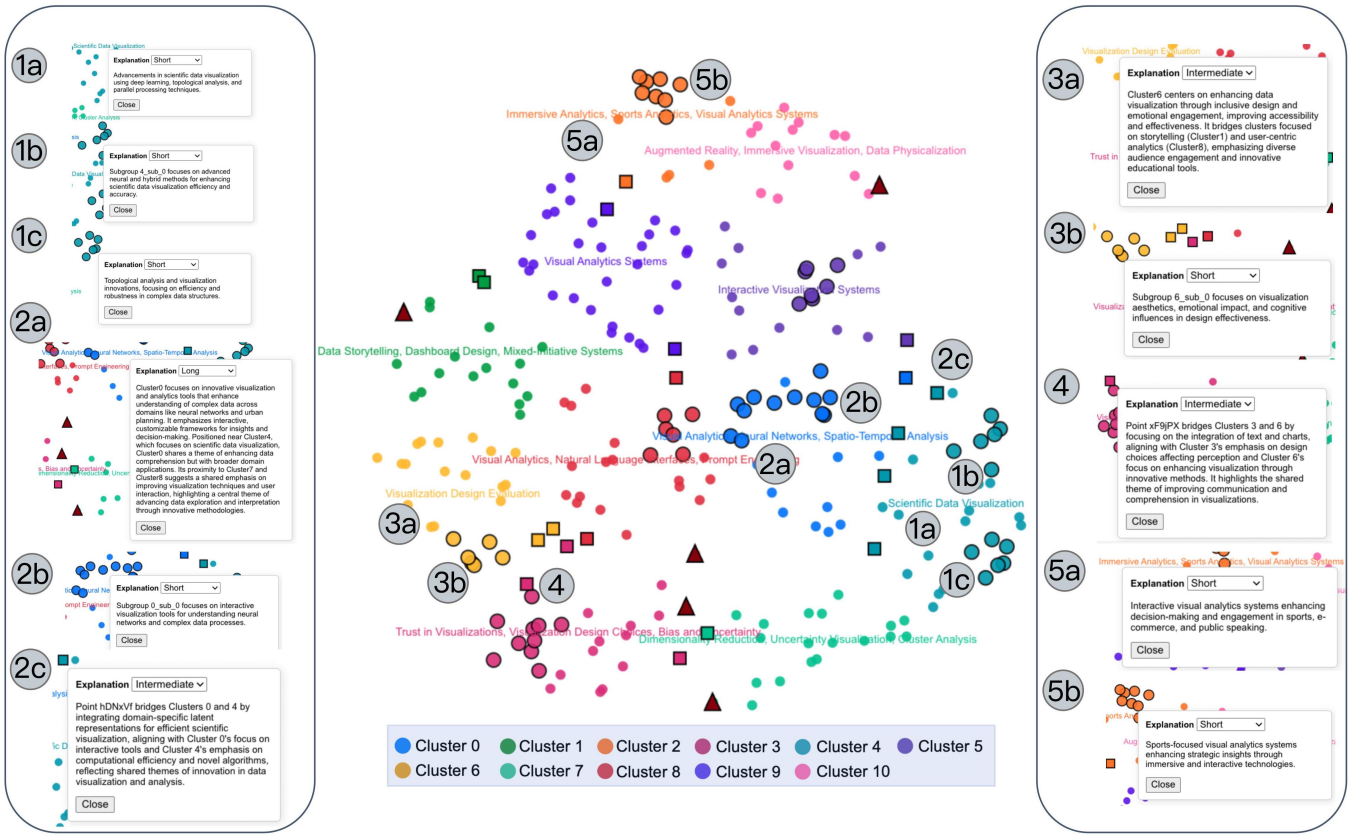

Fig. 5: Exploring IEEE VIS 2022–2023 research trends with CAPE in AI Explanation Mode using a UMAP projection of 252 papers. (1a) A cluster focusing on scientific data visualization, featuring two subgroups: (1b) neural optimization and (1c) topological analysis. (2a) A cluster centered on innovative visualization frameworks, including (2b) a subgroup on tools enhancing interpretability in neural models, and (2c) a bridging paper linking innovative visualization tools and computational efficiency. (3a) A cluster emphasizing inclusivity, emotional engagement, and educational tools, with (3b) highlighting aesthetics and cognitive aspects. (4) One bridging paper connecting cognitive design (Cluster3) and emotional design (Cluster6). (5a) A cluster focusing on immersive, domain-specific visual analytics (sports, e-commerce, public speaking), including (5b) a subgroup related to sports analytics.

local structure in the projection space, revealing both distinct subfields and interdisciplinary connections. Rather than relying on predefined categories or metadata, we were able to move fluidly through the space, discovering and understanding patterns through CAPE's semantically grounded, LLM-generated explanations.

6.3 Comparison with Keyword-Based Explanation Methods

To evaluate the CAPE's context-aware explanation approach, we compared it with a traditional keyword-based explanation baseline method. In this baseline, we used the 20 Newsgroups dataset. After removing stopwords, we extracted the top 20 TF-IDF keywords from each document, and analyzed the keyword lists for all documents involved in a selected pattern (e.g., a cluster or outlier). These keyword sets were used to approximate explanations. We observed several key differences between the keyword-based approach and CAPE:

(1) Semantic Depth While keyword-based explanations captured frequently occurring terms, they often failed to reflect deeper semantic relationships or internal structure within a cluster. For example, a cluster centered on religious and ideological discussions produced common keywords such as "god," "church," and "Islam." While these terms were topically relevant, they lacked structure and did not reveal the range of themes present in the cluster, such as theological contradictions, social attitudes, and historical conflicts, which CAPE's explanations clearly identified.

In contrast, CAPE provided natural language explanations that captured the subtopics present within the cluster. It also suited each cluster within the projection space, explaining how they related to neighboring clusters or connected, something the flat keyword list could not do.

(2) Pattern-Specificity Keyword lists offered no insight into why a document or pattern appeared in a specific position in the projection space, especially for ambiguous points such as outliers and bridging documents.

For example, one outlier document was located near a technology-focused cluster (labels: "encryption," "space exploration," and "technology standards"). The keyword summary for the outlier included "flat," "crank," and "engine," which were too generic and failed to explain their relevance or difference to nearby documents.

CAPE's explanation (Fig. 1 (2)), by contrast, recognized that the document was about automotive engineering, a topic semantically different from the surrounding documents about encryption and space systems. CAPE also explained why the document appeared as an outlier: it focuses on engine configurations, deviating from Cluster 2's tech and security themes. This pattern-level explanation was entirely missing from the keyword baseline.

(3) Human Interpretability CAPE's explanations were consistently easier to understand and more informative of the embedding space. Additionally, keyword lists offered no ability to control the level of detail other than list length, whereas CAPE supports multi-level explanations (concise, intermediate, in-depth) tailored to the user's needs. It is especially useful during exploratory tasks where users shift between overview and detail.

While keyword-based methods can offer a quick thematic snapshot, they lack structure, context, and interpretability. CAPE, by contrast, generates semantic narratives grounded in spatial and document context, providing a more meaningful explanation.

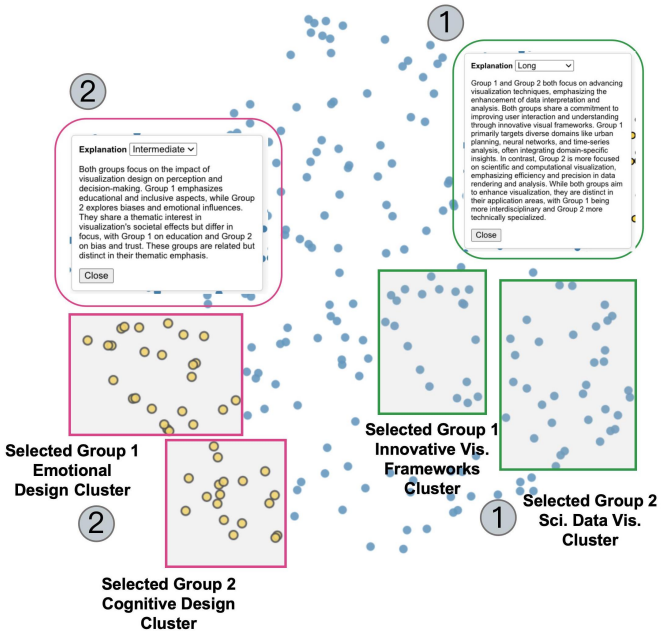

Fig. 6: User Exploration Mode applied to the IEEE VIS 2022–2023 dataset (UMAP projection). (1) Comparing two spatially adjacent clusters shows different focuses on innovative visualization tools vs. performance-driven rendering. (2) Selecting clusters centered on emotional/educational design and trust/bias highlights a thematic transition from inclusivity to cognitive considerations.

7 Discussion

Explaining Projection Spaces, Not Just Clusters CAPE addresses a big limitation in traditional document projection tools: the lack of meaningful, interpretable explanations grounded in both semantics and spatial structure. While DR techniques like t-SNE and UMAP reveal patterns in high-dimensional embeddings, they do not explain why those patterns exist. CAPE bridges this gap by generating natural language explanations that are grounded not only in the document contents, but also in their position and relationships in the projection space. Compared with static labels or keyword lists, these contextual explanations enable richer, more human-centered interpretation. Rather than treating projections as visual groupings alone, CAPE treats them as semantic landscapes, where users can understand both broad topics and local relationships.

Beyond Keywords: Context-Aware Explanation Compared to traditional keyword-based or static label approaches, CAPE produces explanations that are richer, multi-layered, and contextually grounded. Keyword lists often reflect frequent terms, but fail to explain why certain documents are positioned where they are. CAPE's explanations incorporate projection context, such as cluster proximity, local neighborhood structure, or document role within a pattern, allowing users to understand relationships, not just topics. This context-aware strategy helps bridge the gap between machine-generated embeddings and human-centered interpretation by making abstract spatial arrangements more explainable and semantically meaningful.

Generalization While CAPE was designed for explaining projected text embeddings, the core idea—using LLMs to generate contextual, multi-level explanations grounded in spatial structure—is generalizable to other types of large-scale, spatially arranged data. Example domains could include geospatial analysis and multimodal projections. Our approach provides the potential of combining LLMs and spatial analytics for interpretable sensemaking in different big data scenarios. Additionally, CAPE provides a modular framework that allows its components, such as the embedding model, projection technique, pattern detection, or explanation approach, to be customized or extended for specific domains.

Future Directions CAPE generates explanations automatically based on detected patterns and user selections. Future versions could allow users to refine or edit explanations. This would facilitate a more collaborative approach where users can contribute domain knowledge, and guide the explanation process. Incorporating user feedback could enable model refinement that improves output quality, reduces hallucinations, and better aligns with specific analytical needs. Even further, this approach could be adapted to explain the results of semantic interactions in the projection space [4]. Ultimately, this could enable more advanced human-AI sensemaking processes that incrementally formalize the projection as a form of "space to think" [1, 10, 31].

In addition, future work could include user studies to evaluate effectiveness and usability. Understanding how users engage with CAPE in exploratory tasks can help improve explanation quality and visual design. Studies could help to further enlighten cognitive process and explanation needs involved in human interpretation of embedding projections.

8 Conclusion

In this work, we introduced CAPE, a context-aware explanations framework for generating semantically and spatially grounded explanations for document embedding visualizations. CAPE supports both automated pattern detection and user-driven exploration, providing multi-level explanations for clusters, outliers, subgroups, bridging documents, individual documents, and user-selected regions. By grounding explanations in global, local, and pattern-specific context, CAPE enables users to understand not only what documents are about, but why they appear where they are. Through usage scenarios and comparison with keyword-based baselines, we demonstrate how CAPE reveals deeper semantic structure and enables more insightful interpretation of text embeddings projections. While currently focused on text, CAPE's approach can generalize to other spatialized data domains where LLM-based explanations can enhance interpretability.

Supplemental Materials

All supplemental materials are available on OSF at https://osf.io/xp5eg/?view_only=9444a6147224456aac20d5a8568fce7d, released under a CC BY 4.0 license.

References

[1] 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, 10 pages, p. 55–64. Association for Computing Machinery, New York, NY, USA, 2010. doi: 10.1145/1753326.1753336 9

[2] A. Bendeck and J. Stasko. An empirical evaluation of the gpt-4 multimodal language model on visualization literacy tasks. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 2

[3] Y. Bian and C. North. Deepsi: Interactive deep learning for semantic interaction. In *26th International Conference on Intelligent User Interfaces*, pp. 197–207, 2021. 2

[4] Y. Bian, C. North, E. Krokos, and S. Joseph. Semantic explanation of interactive dimensionality reduction. In *2021 IEEE Visualization Conference (VIS)*, pp. 26–30. IEEE, 2021. 1, 9

[5] Z. Bitvai and T. Cohn. Non-linear text regression with a deep convolutional neural network. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 180–185, 2015. 2

[6] L. Bradel, C. North, and L. House. Multi-model semantic interaction for text analytics. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 163–172. IEEE, 2014. 2

[7] K. Choe, C. Lee, S. Lee, J. Song, A. Cho, N. W. Kim, and J. Seo. Enhancing data literacy on-demand: Llms as guides for novices in chart interpretation. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 2

[8] M. Dowling, N. Wycoff, B. Mayer, J. Wenskovitch, L. House, N. Polys, C. North, and P. Hauck. Interactive visual analytics for sensemaking with big text. *Big Data Research*, 16:49–58, 2019. 2

[9] A. Endert, R. Burtner, N. Cramer, R. Perko, S. Hampton, and K. Cook. Typograph: Multiscale spatial exploration of text documents. In *2013 IEEE International Conference on Big Data*, pp. 17–24. IEEE, 2013. 2

[10] A. Endert, P. Fiaux, and C. North. Semantic interaction for visual text analytics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, 10 pages, p. 473–482. Association for Computing Machinery, New York, NY, USA, 2012. doi: 10.1145/2207676.2207741 9

[11] R. Faust, D. Glickenstein, and C. Scheidegger. Dimreader: Axis lines that explain non-linear projections. *IEEE transactions on visualization and computer graphics*, 25(1):481–490, 2018. 2

[12] R. Fu, J. Liu, X. Chen, Y. Nie, and W. Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning. *arXiv preprint arXiv:2403.11401*, 2024. 2

[13] O. Gomez, S. Holter, J. Yuan, and E. Bertini. Vice: Visual counterfactual explanations for machine learning models. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pp. 531–535, 2020. 2

[14] O. Gomez, S. Holter, J. Yuan, and E. Bertini. Advice: Aggregated visual counterfactual explanations for machine learning model validation. In *2021 IEEE Visualization Conference (VIS)*, pp. 31–35. IEEE, 2021. 2

[15] F. Heimerl, S. Lohmann, S. Lange, and T. Ertl. Word cloud explorer: Text analytics based on word clouds. In *2014 47th Hawaii international conference on system sciences*, pp. 1833–1842. IEEE, 2014. 2

[16] J. S. Kessler. Scattertext: a browser-based tool for visualizing how corpora differ. *arXiv preprint arXiv:1703.00565*, 2017. 2

[17] K. Lang. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 331–339, 1995. 3

[18] S. Liu, X. Wang, J. Chen, J. Zhu, and B. Guo. Topicpanorama: A full picture of relevant topics. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 183–192. IEEE, 2014. 2

[19] W. Liu, C. North, and R. Faust. Visualizing spatial semantics of dimensionally reduced text embeddings. *arXiv preprint arXiv:2409.03949*, 2024. 1, 2

[20] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017. 2

[21] C. C. Marshall, F. M. Shipman III, and J. H. Coombs. Viki: Spatial hypertext supporting emergent structure. In *Proceedings of the 1994 ACM European conference on Hypermedia technology*, pp. 13–23, 1994. 1

[22] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 2

[23] L. H. Nguyen and S. Holmes. Ten quick tips for effective dimensionality reduction. *PLoS computational biology*, 15(6):e1006907, 2019. 1

[24] K. Nokkaew and R. Kongkachandra. Keyphrase extraction as topic identification using term frequency and synonymous term grouping. In *2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, pp. 1–6. IEEE, 2018. 2

[25] R. Patil, S. Boit, V. Gudivada, and J. Nandigam. A survey of text representation and embedding techniques in nlp. *IEEE Access*, 11:36120–36146, 2023. 1

[26] M. T. Ribeiro, S. Singh, and C. Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016. 2

[27] T. Ruotsalo, J. Peltonen, M. Eugster, D. Głowacka, K. Konyushkova, K. Athukorala, I. Kosunen, A. Reijonen, P. Myllymäki, G. Jacucci, et al. Directing exploratory search with interactive intent modeling. In *Proceedings of the 22nd ACM international conference on information & knowledge management*, pp. 1759–1764, 2013. 2

[28] K. Singh, H. M. Devi, A. K. Mahanta, et al. Document representation techniques and their effect on the document clustering and classification: A review. *International Journal of Advanced Research in Computer Science*, 8(5), 2017. 2

[29] Y. Sui, M. Zhou, M. Zhou, S. Han, and D. Zhang. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pp. 645–654, 2024. 2

[30] V. Sumithra and S. Surendran. A review of various linear and non linear dimensionality reduction techniques. *Int J Comput Sci Inf Technol*, 6(3):2354–60, 2015. 2

[31] X. Tang, E. Krokos, C. Liu, K. Davidson, K. Whitley, N. Ramakrishnan, and C. North. Steering llm summarization with visual workspaces for sensemaking, 2024. 2, 9

[32] X. Tang, S. Wong, M. Huynh, Z. He, Y. Yang, and Y. Chen. Sphere: Scaling personalized feedback in programming classrooms with structured review of llm outputs. *arXiv preprint arXiv:2410.16513*, 2024. 2

[33] F. S. Tsai. Dimensionality reduction techniques for blog visualization. *Expert Systems with Applications*, 38(3):2766–2773, 2011. 2

[34] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 2

[35] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics*, 26(1):56–65, 2019. 2

[36] J. Wise, J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. Visualizing the non-visual: spatial analysis and interaction with information from text documents. In *Proceedings of Visualization 1995 Conference*, pp. 51–58, 1995. doi: 10.1109/INFVIS.1995.528686 2

[37] Y. Zhang, H. Jin, D. Meng, J. Wang, and J. Tan. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *arXiv preprint arXiv:2403.02901*, 2024. 2