Computer Science Review 57 (2025) 100748

Contents lists available at ScienceDirect

Computer Science Review

journal homepage: www.elsevier.com/locate/cosrev

Review article

Empowering multimodal analysis with visualization: A survey

Jiachen Wang [a], Zikun Deng [b,*], Dazhen Deng [c], Xingbo Wang [d], Rui Sheng [e], Yi Cai [b], Huamin Qu [e]

[a] Department of Sports Science, Zhejiang University, Hangzhou, Zhejiang, China
[b] School of Software Engineering, South China University of Technology, Guangzhou, Guangdong, China
[c] School of Software Technology, Zhejiang University, Ningbo, Zhejiang, China
[d] Weill Cornell Medical College, Cornell University, New York, NY, United States
[e] The Hong Kong University of Science and Technology, Hong Kong, China

ARTICLE INFO

Keywords:
Visualization
Multimodal data
Machine learning

ABSTRACT

Multimodal data, which encompasses text, audio, image, and other modalities, is a popular research target in the field of visualization research. Existing visualization techniques for multimodal data are scattered and categorized by application domains, such as multimodal model analysis or online education. It lacks a comprehensive review from the perspective of data that summarizes the methodologies, research gaps, and future trends for researchers and practitioners. In this study, we delve into existing visualization research, identifying their data modalities, applications, strengths, and limitations. Furthermore, we shed light on the potential challenges and opportunities for further research in this domain to advance intelligent visualizations for multimodal data.

Contents

* Corresponding author.
E-mail address: zkdeng@scut.edu.cn (Z. Deng).

https://doi.org/10.1016/j.cosrev.2025.100748
Received 1 February 2024; Received in revised form 28 January 2025; Accepted 16 March 2025
Available online 2 April 2025

1. Introduction

With the development of data collection technology (e.g., cameras and sensors), multimodal data (e.g., text, audio, and video) gains prevalence across diverse fields, such as education, environment, and healthcare. Data from a single modality often comprise limited information, and it is particularly important to consider data from multiple modalities in the aforementioned fields [1]. Take a teaching speech as an example. When analyzing a speech video, it is necessary to view the speaker's gestures and postures through a multi-frame image sequence and listen to his exciting and frustrating speech. In other words, a video is vivid and expressive only when it has both audio and images of video frames and when the audio and images are aligned.

In the domain of artificial intelligence (AI), researchers have proposed numerous approaches for multimodal data, encompassing representation [2], generation [3,4], alignment [5], and fusion [6,7] of diverse modalities. Despite these remarkable advances, the interpretability and usability of multimodal AI in data analysis remain constrained. Multimodal AI mainly focuses on the effective vectorization of multimodal data (e.g., the fusion of multimodal features) and the subsequent utilization of these vectorized representations to generate multimodal data (e.g., visual question answering [8,9] and image/video captioning [10,11]). However, end-to-end generative models exhibit limitations in interpretability, both during model training and inference stages. End-users encounter challenges in accessing nuanced insights and validating the reliability of model outcomes due to the absence of exploratory analysis tools tailored for multimodal data exploration. Take the education speech as an example again. While multimodal AI can evaluate whether a speech is inspiring based on cues like gesture, audio, and linguistic content [12], it falls short in discerning the specific cues or combinations thereof that contribute to the perceived inspiration, a concern paramount to stakeholders. Analogously, in the realm of medical diagnostics, integrating data from disparate sources such as medical records, computed tomography (CT) images, and electrocardiograms poses analogous challenges for transparent and informed disease diagnosis.

To support sufficient and efficient analysis of multimodal data, visualization researchers have developed effective multimodal visualizations to help analysts interpret and explore multimodal data and gain actionable insights with their capacity in reasoning and decision [13–15]. For example, EmoCo [16] allows users to analyze joke-telling styles by simultaneously exploring scripts, facial expressions, and voice pitch—for example, delivering funny sentences with deadpan seriousness. While multimodal visualization has been successfully applied in many fields, such as education [17], speech [18], and healthcare [19], a comprehensive review of visualizations of multimodal data is still unavailable, leaving many important questions pending. For example, what is the major methodology for applying visualization to multimodal analysis? How does visualization assist multimodal data analysis? What are the future research trends and challenges? How can AI4VIS [20,21] be applied to intelligent multimodal visualization?

To answer these questions, we conducted a survey to examine the state-of-the-art works of visualization for multimodal data systematically. We identified each work's data modality, analysis tasks, visualization techniques, application domain, strengths, and limitations. Then, we summarized a common methodology for multimodal data visualization. Moreover, we discussed potential challenges and opportunities for further research in this domain. Based on these findings, this work aims to advance the development of intelligent visualization techniques for more efficient and effective multimodal data analysis in diverse domains. The contributions are as follows.

- Collecting the state-of-the-art works of visualization for multimodal data.
- Summarizing the data modality, analysis tasks, visualization techniques, application domain, strength, and limitations in works of visualization for multimodal data.
- Discussing the potential challenges and research opportunities for multimodal data analysis.

2. Methodology and overview

In this section, we present the scope of this survey and the methods for paper collection and analysis.

2.1. Scope

We aim to focus on works leveraging visualization techniques to empower multimodal analysis in this survey. We use multimodal analysis to denote analysis tasks that concentrate on multimodal data collected with multiple instruments, measurement devices, or acquisition techniques [1]. Multimodal data brings considerable challenges to analysis due to the unique data characteristics (e.g., data formats, scales, level of noise) of different modalities. Therefore, in this survey, we aim to summarize the methodology for applying visualization to multimodal analysis by revealing the relationships between data modality, analysis

tasks, visualization techniques, and application domains. We selected papers based on the following three exclusion criteria.

- **C1: Excluding multimodal interaction.** At the beginning of the paper collection, we searched papers based on the keyword, "multimodal". Therefore, we inevitably collected some papers investigating multimodal interactions in virtual reality and augmented reality [22], for example, Yuan et al.'s work [23]. These papers did not fit within our scope of multimodal data analysis. We excluded such papers during paper filtering.
- **C2: Excluding scientific data.** Scientific data has a strong inherent reference to space, time, and results from data acquisition methods [24]. Such data is usually collected in the form of flows and volumes by various simulation and imaging approaches, for example, X-ray and neutron computed tomography [25] shown in Fig. 1. We excluded works on scientific data due to the different research challenges and interests. For example, Lawonn et al. [26] conducted a comprehensive survey about multimodal medical data visualization. The multimodal medical data were collected through different medical imaging techniques, such as computed tomography and ultrasound. The overall visualization goal of such data is developing effective rendering approaches to clearly illustrate the structural anatomical information, which do not fit within our defined scope.
- **C3: Excluding "pseudo-multimodal" data.** According to [1], multimodality should be complementary, which implies that each data modality should offer unique values that cannot be inferred or acquired from other data modalities. Therefore, during filtering, we excluded works on "pseudo-multimodal" data where researchers infer data of different modalities from data of a single modality. For example, in ForVizor [27], both time series data (i.e., players' tracking data and key indicators) and event sequence data (i.e., match events and team formations) are derived from the image data (i.e., video frames). The research interests and challenges of multimodal data are weakened in such works.

2.2. Paper collection

To obtain a holistic view of state-of-the-art visualization works on multimodal data, we collected papers from the top conferences and journals of visualization (IEEE TVCG, IEEE VIS, EuroVis, etc.) and human–computer interaction (CHI, UIST, etc.). We went through three rounds to collect the papers. In the first round, we used search-based selection. We searched the keyword, "multimodal" in different corpus of journals and conference proceedings. Then, we filtered out the papers concerning multimodal data visualization by reviewing the abstracts and full texts. During this filtering, we mainly excluded papers based on **C1** and **C2**. In the second round, we used reference-based methods. We checked the references of the selected papers in the first round and collected more related papers. In this round, we mainly excluded papers based on **C2** and **C3**. In the third round, to ensure that our paper corpus is as comprehensive as possible, we repeated reference-based methods based on the selected papers in the second round. In this round, we mainly excluded papers based on **C3** After the three rounds, we collected 179 papers and filtered out 46 papers concerning multimodal data visualization.

According to the publication time displayed in Fig. 2, we found that since 2020, there has been a rapid increase in papers related to multimodal data visualization. In 2020 alone, eight related papers were published. We thought this was mainly because of the significant progress in machine learning [28]. Researchers can easily derive rich information from multimodal data, such as video [14,16,17,29], image [30–32], audio [33], and text [34,35].

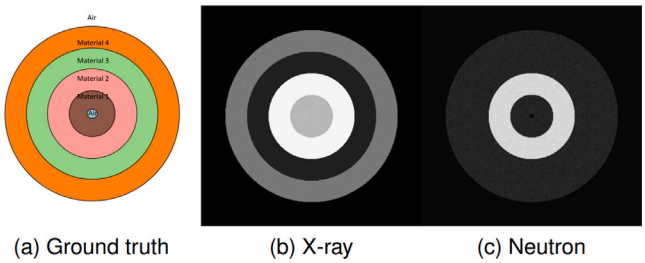

Fig. 1. An example of multimodal scientific data [25].

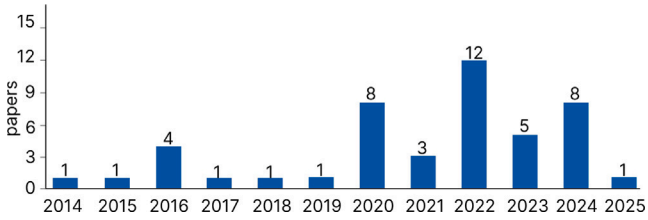

Fig. 2. The number of the surveyed papers studying multimodal data with visualization in each year.

2.3. Paper coding

We invited four experts to code the papers. All of them are researchers in visualization, studying novel visualization techniques in various domains such as e-learning, medical, and social media. They had rich experience in visualizing multimodal data for diverse analysis tasks. They were asked to code all papers based on the data modalities, analysis tasks, visualization techniques, and application domains. The detailed coding scheme is as follows.

- **Data modality**: We formulated the initial taxonomy of data modality based on the survey of multimodal machine learning [36] which provides ten data modalities. We merged "force sensors" into "time-series" since sensor data are continuous signals that can also be treated as time-series data. Additionally, during the coding, we added three additional labels, namely, "event sequence", "geolocation", and "graph" to ensure the comprehensiveness of data modality. Finally, we identified nine data modalities as Fig. 3 shows. To be mentioned, we coded both the source modality and analysis modality of each paper. The source modality refers to the modality of the source data before processing. The analysis modality refers to the modality of the data processed and visualized.
- **Analysis task**: At the beginning, we referred to the low-level visualization tasks proposed by Amar et al. [37] to code the analysis task for each paper. However, after the first round of paper collection, we found the granularity of this task taxonomy was too fine to provide effective insights. Therefore, we went through the "design requirements" or "analysis tasks" of the papers and collected the high-level analysis tasks. Moreover, we referred to the design space of visualization tasks [38] and summarized the five analysis tasks in multimodal data visualization as Fig. 3 shows.
- **Visualization technique**: To obtain in-depth insights, we coded two types of visualization techniques. First, we coded the visualization techniques of each single data modality separately. Here, we only coded the analysis modality. We referred to the taxonomy proposed by Deng et al. [39] for coding since this taxonomy is one of the most comprehensive taxonomy about visualization techniques in research. Second, we coded the techniques for visualizing multiple data modalities. We summarized four kinds of

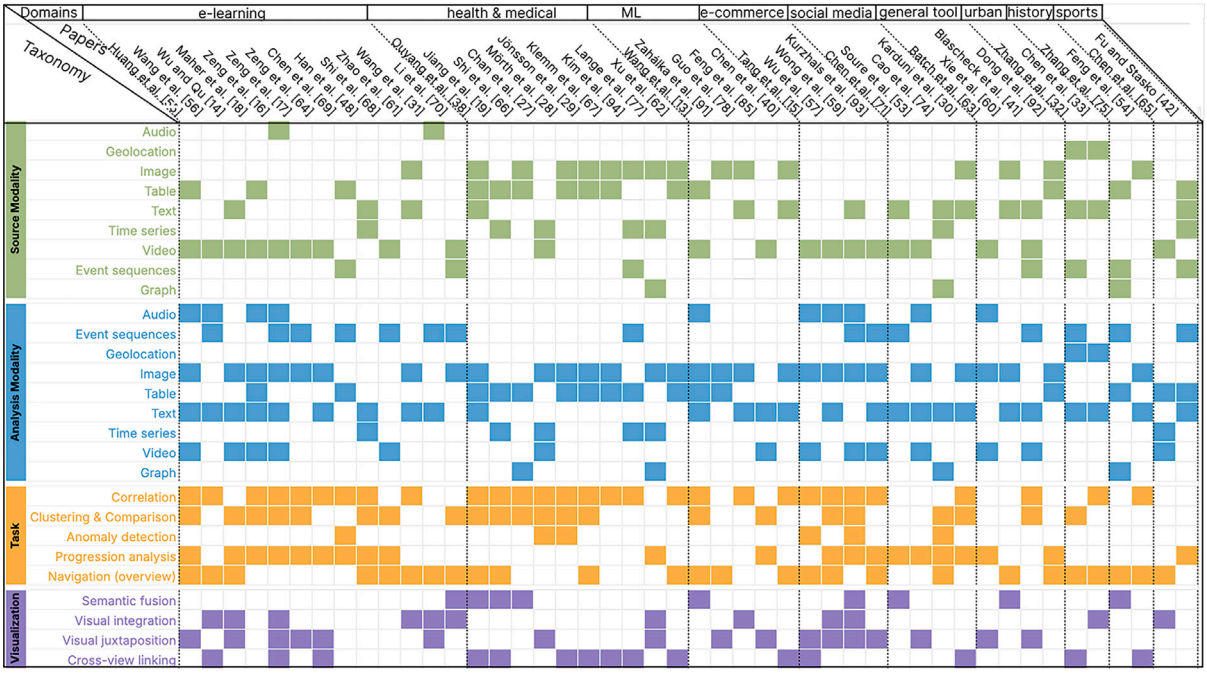

Fig. 3. The codes (i.e., data modality, analysis tasks, visualization techniques, and application domains) of the surveyed papers.

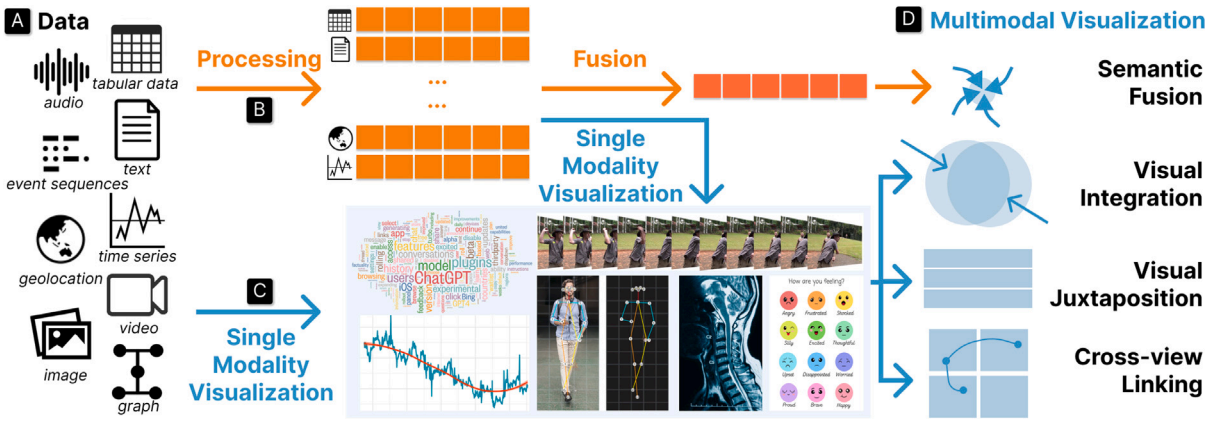

Fig. 4. The general framework of multimodal visualization. Orange indicates the computational processing. Every modality is processed into a vectorized representation, respectively, which can be followed by the fusion that generates the unified representation for multiple modalities. Blue indicates the visual design. The raw modality, the representation, and the fusion result can be visualized, forming multimodal visualizations and visual analytics systems.

techniques based on the degree of integration, namely, semantic fusion, visual integration, visual juxtaposition, and cross-view linking (Fig. 3).

- **Application domain** At first, we mainly coded the application domain of each paper based on the keywords. After we coded all papers, we merged similar domains such as "healthcare" and "medical". Finally, we had nine application domains as Fig. 3 shows.

In each paper collection round (Section 2.2), one paper was coded by three experts separately. After the initial coding, we held meetings to discuss the controversial papers. The final coding result was not decided until we all agreed.

2.4. Overview

We finally concluded a general multimodal data visualization framework shown in Fig. 4. The orange indicates the computational processing, while the blue indicates the visual design. On the processing side, every modality is processed into a vectorized representation, respectively, which can be followed by the fusion that generates the unified representation for multiple modalities. On the visualization side, the raw modality, the representation, and the fusion result can be visualized, forming multimodal visualizations and visual analytics systems.

This survey will be organized following this framework. Section 3 introduces the data modalities (Fig. 4A) in the paper collections and how these modalities are processed for in-depth analysis (Fig. 4B). Section 4 introduces common tasks of multimodal data analysis. Section 5 introduces single modality visualizations (Fig. 4C). Based on single modality visualizations, Section 6 further introduces multimodal visualizations (Fig. 4D).

3. Data modality and processing

In this section, we introduce the data modality and the corresponding processing that used in the visualization community. Fig. 5 shows the data modalities commonly used in the papers we surveyed, sorted

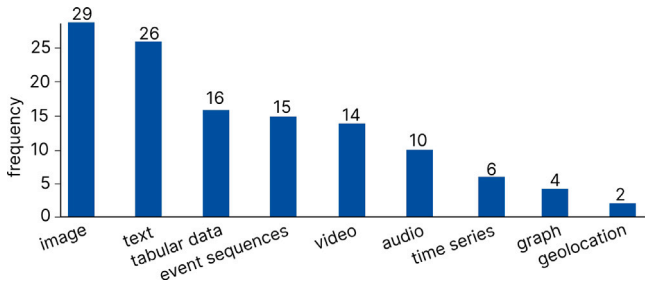

Fig. 5. The frequency of the different modalities used in the surveyed papers.

in descending order of the number of times they appear. Below we introduce them one by one with concrete examples.

3.1. Image

Images often come from a camera or video camera. For example, on social media, users tend to attach pictures when writing tweets [40]; keyframes in videos are often extracted as images for analysis [16]; and CT/MRI data are often rendered as images [31,41]. The macro features in a natural picture are easy to understand. For example, people can easily distinguish whether the picture is of a cat or a dog. However, to quickly understand microscopic features in images, such as the speaker's emotion at a certain moment [16,17,42] and abnormal parts in medical images [30,41], it is necessary to introduce advanced image processing techniques to provide visual cues.

- *Feature Extraction.* For example Xie et al. [43] used a neural network to transform images and their captions into high-dimensional representations, which were subsequently used to obtain semantic-preserving projections. Feng et al. [44] also generated semantic-preserving representations for images in a similar way for clustering and retrieval. Image feature extraction also has applications in the medical field [45]. Morth et al. [30] also attempted to create medical images. They generated high-dimensional feature vectors of tumors based on medical images and tumor masks (image segments) in a parametric imaging sequence.
- *Sentiment Analysis.* The most commonly used image processing technique is sentiment analysis [16,17,42,46,47], which is especially useful for images containing faces in the fields of e-learning and electronic commerce. For example, to figure out how live commerce streamers' personalities and selling skills impact viewers to buy the selling product, Wu et al. [48] first detected the faces in live commerce videos and extracted emotional expressions based on these faces. As emotion is critical to the effectiveness of speeches, Maher et al. [18] extract the time-varying emotion of speeches, also following the pipeline of face detection and then emotion extraction.
- *Content Detection.* Content in the image, such as objects [49], gestures [50], and postures [51], is often detected to facilitate subsequent analysis and augmentation. For example, Huang et al. [46] extracted and used the gestures of speakers to analyze the speaking effectiveness since body language is especially important for audience-speaker interactions. Chen et al. [52] detected players in videos of basketball games and overlaid visualizations of player statistics onto the videos. Chen et al. [53] leveraged the objects detected from the image to guide image captioning.

3.2. Text

Text is one of the easiest data modalities to generate and collect without relying on specialized equipment but with a keyboard or pen. It frequently appears in many multimodal applications because it can deliver descriptive information in a condensed but accurate manner. For example, medical history used in health and medical diagnosis is a type of text [41]; subtitles in videos are also a type of text [51]; the text is often used to describe an image [53]. People can easily understand a word and a sentence. Thus, in some multimodal studies, authors choose to present textual content directly without processing it [54,55]. When the text corpus is large, such as an article or even multiple articles, textual information often needs to be processed by computational methods before it can be analyzed by users. Below are some popular methods:

- *Content Summary*. Researchers usually leverage computational methods to summarize textual content in multimodal data. Such that, users can quickly obtain an overview of multimodal data through summaries in the form of keywords [56,57] and topics [58,59]. For example, Zhang et al. [34] applied latent Dirichlet allocation (LDA) [60] to extract the theme and keywords of tourist posts. Sung et al. [61] utilized TF-IDF [62] to obtain the topic of the comments posted by online learners on e-learning platforms. Chen et al. [35] estimated which aspect of a city the citizen comments were talking about.
- *Sentiment Analysis*. Besides the overall textual content, advanced computational methods also allow for the analysis of individual sentences [63–65]. The most common task at the sentence level is sentiment analysis, which is often used in video analysis. For example, Zeng et al. [16] extracted emotion tones, such as anger, happiness, and sadness, for every sentence spoken by the speakers of TED talks, followed by the narrative emotion analysis of videos. Soure et al. [66] detected user emotions based on what they said during user studies to help review the user studies. In addition to the video, sentiment analysis can be found in image understanding [49] and comment review [61].
- *Similarity Analysis*. Capturing the similarity between texts is the fundamental step of exploring large-scale multimodal data. To do that, the first step is usually the vectorization of text, for example, based on the RNN [64] or TF-IDF [61]. Then, the text similarity can be computed based on their vectors. For example, Sung et al. [61] adopted the cosine similarity to group multiple comments of e-learning learners. Wang et al. [33] leveraged the similarity of speech video for video recommendation based on the speech content. Zhang et al. [34] used topic coherence measures to cluster tourism themes. These similarity analyses allow users to retrieve and view a lot of multimodal information efficiently [46].

In addition to the aforementioned popular text processing methods, some studies presented interesting ones for special requirements to help analyze multimodal data from another perspective. For example, Wang et al. [67] identified and revealed the relationship among build-ups of a punchline based on the humor script. Karduni et al. [32] used feature engineering to analyze which languages (e.g., those words with fear, anger, and negative sentiment) suspicious social media accounts tended to use. Wu et al. [48] segmented live commerce videos into multiple selling strategies based on the video scripts, which can help navigate the videos.

3.3. Tabular data

Tabular data is organized into a format where information is arranged in rows and columns. In this structure, each row typically represents a record or observation, while each column corresponds to a specific attribute or variable related to these records. This type of data is most common in multimodal medical diagnosis, such as the patient's demographic information [68], medical history [41], and the results of various medical tests [45,68]. For video analytics, tabular data is often meta information about videos used to classify and retrieve them [18,67].

The straightforward way for tabular data is multidimensional visualization, such as parallel coordinate plots [30,31,69], and even table directly [70]. To obtain in-depth insights, further features can be extracted from them to enhance subsequent analysis. For example, Ouyang et al. [41] extracted vectorized features based on patient's laboratory test results and demographic information. These features are subsequently used to predict the diagnostic result. Jiang et al. [19] transformed children's living behaviors and demographics into vectorized features and used these features to model children's health profiles.

3.4. Event sequence

An event sequence consists of a series of time-stamped events. The intervals are usually varied. The event sequences appearing in multimodal analysis are mainly generated during the people's interaction with the multimedia, such as the user's action on the video (pause and fast forward, etc.) [71,72] and the user's operating system behavior (click and drag and drop, etc.) [73], the user's streaming comments (i.e., Dammu or Danmaku [74]) while watching the video, and the customer's interactions with customer service system [42].

Although event sequences are not the primary form of media, making good use of them can enhance multimodal data analysis. For example, in the scenario of an online exam, it is difficult to detect cheating based on video alone accurately. Thus, Li et al. [73] took the participants' mouse events on the exam system into account. In the scenario of user study review, the mouse events on the test system can also be utilized to analyze the user's special behavior (e.g., confusion) in addition to the recorded video and audio [54].

3.5. Video

Video, as a commonly used data type in multimodal analysis, is rich in information and multidimensional features. It encompasses not only visual content but also audio and text (such as subtitles), making it particularly advantageous in applications like emotion analysis, behavior recognition, and scene understanding. To facilitate analysis, video data is often segmented into individual frames (**images**), transformed into **audio** data, or transcribed into **text**. Additionally, video data is frequently used in various scenarios such as surveillance (e.g., online exam monitoring), education (e.g., MOOCs), and presentations (e.g., TED Talks). For more detailed information, please refer to the other subsections.

3.6. Audio

Audio data in multimodal analysis are mainly those recording how and what people speak instead of natural sounds, such as city noise. These audios are sometimes transcribed into text (subtitles) through translation before they are analyzed. However, beyond text alone, the content of the speech with audio features like tone, pitch, and pauses will be more informative [33,46,48]. For example, Soure et al. [66] considered, in user studies, users tend to change their pitch when they encounter a problem while thinking aloud. Wang et al. [67] extracted the volume, pitch, speed, and pause from audio and visually encoded them into textual sentences, which facilitates understanding what voice skills to use to tell a joke.

Furthermore, audio is also leveraged to enhance sentiment analysis [18,42], particularly when scripts cannot accurately reflect people's emotions, For example, Zeng et al. [16] applied emotional analysis on audio data of presentation videos, classifying every spoken sentence into seven emotions. They also developed EmoCo to analyzed the coherence and conflicts of the emotions detected from audio, textual, and facial data, respectively.

3.7. Time series

A time series is a series of numeric values ordered by time, often collected at (nearly) fixed time intervals. A typical time series data is the monitoring data of the human body, for example, human motion data [19,75], which can help analyze the human health status. Electrocardiogram is also another common kind of time series data [76], but it rarely appears in current multimodal visualization and visual analysis.

3.8. Graph

A graph structure with nodes and edges can also be considered as a data modality since it can record the relationships (by edges) between discrete entities (by nodes). This modality appears mainly in social media-related analysis. For example, the detection of abnormal social media accounts involves analyzing the content of the posts, the behavior of the posts, and the relationships between the accounts [77]. Even in historical "social media", the relationships of historical celebrities, together with celebrities' demographics and historical events, can enhance the analysis of cohorts [78].

3.9. Geolocation

Location sensors record another media modality, geolocation, from a geographical perspective. For example, the comments for user-gener ated points of interest (POIs) in urban spaces are inherently associated with geolocation. Chen et al. [35] applied topic modeling to such comments and considered their spatial contexts to analyze the user-perceived performance of urban space. Similarly, Zhang et al. [34] leveraged such comments to assist tourists in planning their itineraries.

4. Multimodal tasks

In this section, we introduce the analysis tasks conducted with multimodal data.

4.1. Navigation

Navigation plays an important role at the beginning of multimodal data analysis, enabling analysts to efficiently identify the specific objects of interest. This task is often supported by a data selection panel or an overview. A data selection panel enables analysts to filter out the data of interest efficiently [41,67]. For example, DeHumor [67] presents metadata (e.g., name and views) and the temporal laughter occurrences for humor experts to locate a speech that contains interesting patterns (e.g., strong opening and ending with dense audience laughter). After selecting a speech, users can utilize multimodal humor feature sliders to filter out specific humor snippets with certain verbal and vocal styles.

An overview provides summative information of the entire dataset, guiding in-depth analysis [13,19]. As the mantra from Shneiderman [79] says, "*Overview first, zoom and filter, then details-on-demand*". It enables analysts to identify significant objects for analysis. For example, HealthPrism [19] has a Summary View that helps experts gain an overall understanding of context (e.g., demographics) and motion features (e.g., sensor data for physical activities) in childern's mental health profiles. It displays the statistics of features and allows experts to assess overall importance and influences of features (e.g., sleep patterns) on the health profiles. SpeechMirror [46] has a Factor Panel that supports public speakers' evaluation and understanding of presentation techniques by summarizing different speech factors (e.g., facial expression and eye contact) and their effectiveness trends.

4.2. Progression analysis

Progression analysis aims to discover the evolutional patterns of particular data features over time. This task is often conducted on video data since video data has intrinsic temporal features (Fig. 3). For example, targeting presentation coaches, EmoCo [16] uses a line and a bar code chart to show the overall presenters' emotion coherence and emotion status of individual communication channels, respectively. This helps coaches infer the presentation styles. Anchorage [42] adopts a lateral buoy chart to show the customer satisfaction progression for a service. The emotions inferred from visual and audio channels of service videos are summed over the agent operation. Each operation is represented by three dots, including satisfaction scores derived from visual, audio, and events. The dot size encodes the temporal deviation of a score to draw attention to an operation. Analysts can use the chart to conduct fine-grained evaluation of agent performance. Lange et al.'s work [80] is a special example of this task. They focused on how a cell in the biomedical field, instead of the individuals mentioned above, behaves (such as growing, dividing, proliferating, and dying).

4.3. Clustering & comparison

Clustering & comparison aims to investigate the similarities among data points. The data points here can be textual comments [61], videos [51], and even patients [68]. For example, Sung et al. [61] presented a system that integrates ToPIN and ThemeRiver to help course lecturers gain insights to improve content delivery and engage with students. ToPIN shows topic clustering and evolution, as well as multiple attributes of comments and topics. ThemeRiver reveals the variations in the pool of time-anchored comments, reflecting how topics and the volume of discussion change over time. MV2Net facilitates comparison tasks by enabling neuroscientists to group brain regions and compare brain network features between subject groups, such as patients and healthy controls, to identify disease biomarkers. The system visual designs have coordinated multiple views that allow users to perform side-by-side comparisons of brain networks, using explicit-coding and juxtaposition to highlight differences in connectivity features across different subject groups. The composite views, like the brain necklace visualization, aggregate multiple features into a single representation, facilitating the correlation analysis of complex patterns across various brain connections.

4.4. Cross-modal correlation

Cross-modal correlation aims to uncover the relationships among different modalities for pattern interpretation and explanation. For example, Zeng et al. [16] proposed an augmented Sankey diagram design that summarizes the coherence across three communication channels and provides extracted features for explanation. For the visual channel, a treemap-based design shows a quick overview of representative detected faces in the video. For the textual channel, a word cloud highlights some important words for context understanding. For the audio channel, histograms show acoustic feature distribution. The coaches can explore how the presenters convey their emotions in different modalities. Wang et al. [13] designed a three-layer augmented tree-like visualization to help model experts to gain an overview of the intra- and inter-modal interactions that are learned by a model for multimodal sentiment analysis. The first layer displays distributions of the ground truths and model prediction errors. The second layer reveals the importance of individual modalities in bee swarm plots. The last layer summarizes the information about the pre-defined interactions, including dominance, conflict, and complement, learned by the model.

Notably, recent advances in large multimodal models enable the generation of images with textual prompts. The effectiveness of prompts significantly affects the quality of generated images. Thus, it is also important to correlate the prompts and images to reason the text-to-image generation process. To support such a task, Guo et al. [81] and Feng et al. [44] developed an interactive interface with text and image visualizations.

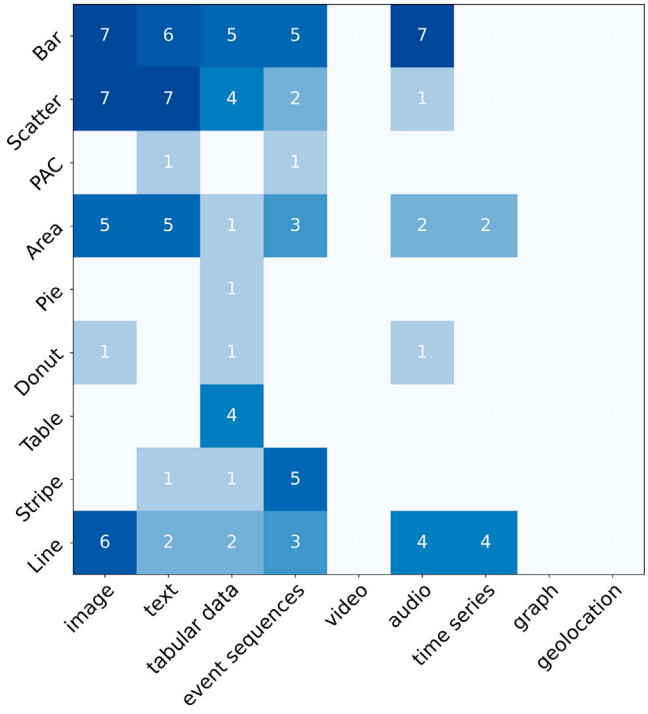

Fig. 6. The frequency of the basic visualizations used for different modalities.

4.5. Anomaly detection

Anomaly detection aims to identify unusual patterns, outliers, or deviations from the norm within multimodal data. Researchers often combine anomaly detection algorithms with visualization techniques to support this task. This task is not common in multimodal data visualization. For example, Li et al. [73] built a suspected cheating case detection engine based on head and mouse movement. Based on the computation results, they used a radar chart-based glyph design to display the risk of cheating from different aspects. The system also enabled fine-grained locations of time periods of abnormal head and mouse movement in Behavior View. Tang et al. [15] trained a binary classifier to prioritize high-risk videos for human moderators' review. Given a video, visualizations further display the risk distribution with storyline-based word visualization and video frame summary as context information.

5. Single modality visualization

In this section, we summarized the visualization techniques for each single data modality. Many surveys have summarized visualization techniques for different data modalities, such as text [82,83], image and video [84], event sequence [85], and time series [86]. While these surveys are comprehensive for each single data modality, they fail to provide valuable insights for multimodal data visualization. Therefore, in this section, we will re-summarize the visualization techniques from the perspective of multimodal data.

Based on the literacy required for each visualization technique, we classified visualization techniques into two types, basic ones and advanced ones. Basic ones contain the visualizations covered by the K-12 curriculum [87] (Fig. 6). Advanced ones contain other visualizations that require better literacy for understanding (Table 1). For advanced ones, we further categorized them into seven classes based on their main visual elements (Fig. 7).

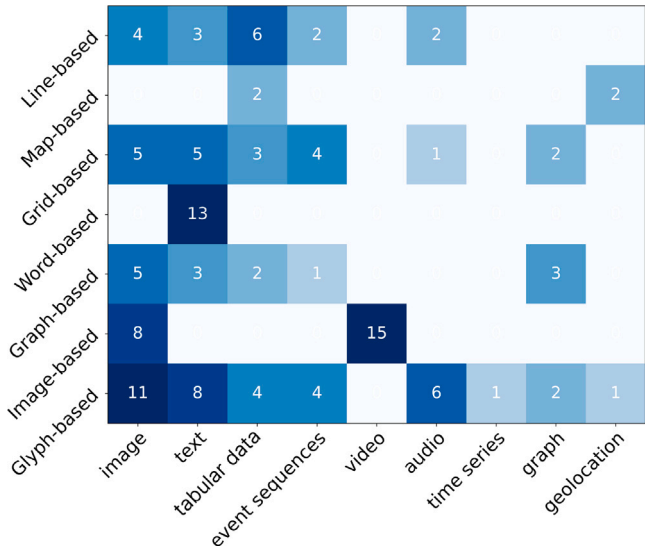

Fig. 7. The frequency of the advanced visualizations used for different modalities.

Table 1
The classification of advanced visualizations.

Class name	Visualization technique
Line-based	PCP, Polar, Arc, Storyline, Sankey, Flow
Map-based	Map
Grid-based	Heatmap, Matrix
Word-based	Word cloud, Text
Graph-based	Tree, Treemap, Sunburst/Icicle, Graph
Image-based	Video player, Augmented video, Image
Glyph-based	Unit, Small multiples, Glyph

5.1. Image

Images contain rich semantic information that people can easily understand when seeing them. This kind of information represents the macroscopic features of images. While macroscopic features are easy to obtain, microscopic features such as speakers' emotions [17,42] and patients' lesion locations [30,41] are difficult to discover. Therefore, researchers need to employ data processing algorithms and intuitive visualization techniques to extract and visualize such features to facilitate multimodal analysis.

Basic: With features extracted, images are most frequently visualized by Bar Chart [42,45] and Scatterplot [43,44] (Fig. 6). The Bar Chart supports efficient comparisons of extracted features. For example, Li et al. [73] conducted face detection and head pose estimation on video frames to identify students' suspected cheating cases. Based on the identified cases, they computed and visualized students' risk levels of cheating by bar charts (Fig. 8). The bar charts could help analysts identify students at high risk of cheating and find questions where high-risk cheating behaviors occurred by comparison. Scatterplot supports efficient navigation of large amounts of image data. For example, Mörth [30] projected all patients on a scatterplot based on their magnetic resonance image (MRI) sequences. The x-axis presents the homogeneity measurement and the y-axis presents the dimension reduction result of one-dimensional t-SNE on selected MRI sequences. Analysts can flexibly explore the MRI sequences of all patients and discover patients of interest. Other popular visualizations including the Area Chart and the Line Chart are often employed to present the temporal trend of extracted features such as customers' satisfaction with services [42], students' head movements during online exams, and users' scrolling speed [66].

Advanced: Due to the rich features images contain, Glyph-based visualization is often used to facilitate the analysis of the multi-dimensional

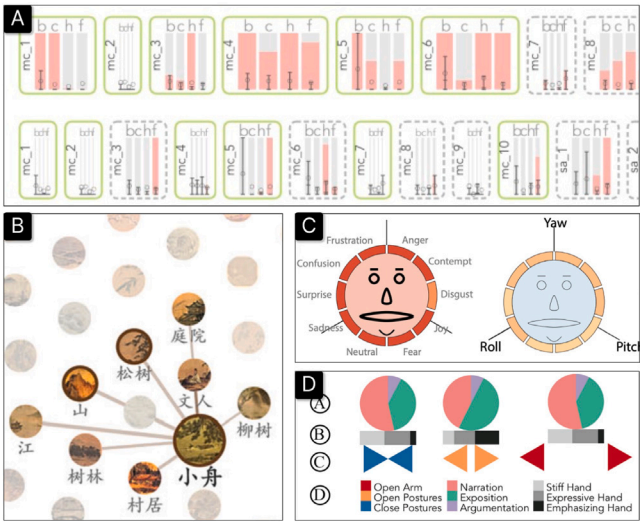

Fig. 8. Examples of image visualization. (A) Li et al. [73] leveraged bar charts to help identify students at high risk of cheating. (B) Feng rt al. [49] leveraged a graph to display images. (C) Wang et al. [13] designed glyphs to display facial expression and sentiment. (D) Wu and Qu [51] visualized postures and gestures through a designed glyph.

feature data extracted from images [88] (Fig. 7). For example, both Wang et al. [13] and Huang et al. [46] designed glyphs based on Chernoff faces [89] to visualize high-dimensional facial expression and sentiment (Fig. 8C). In addition, researchers also designed intuitive glyphs by combining basic charts, such as pie charts and bar charts, to present data of different dimensions [48,66]. For example, the glyph designed by Wu and Qu [51] used a pie chart to present the distribution of rhetorical modes in TED talks, a bar chart to present the percentage of gestures, and two triangles to present the most frequent hand posture (Fig. 8D). All charts were organized based on the metaphor of the human upper body to summarize the characteristics of the speaker in a TED talk.

Image-based visualization is another frequently used to directly show the details in original images [70,90,91]. It can facilitate the exploration of a large image collection. Besides, Grid-based [31,53], Graph-based [16,49], and Line-based [45,51] visualizations are also preferred by researchers when analyzing image data.

5.2. Text

Text is one of the most straightforward data in multimodal data visualization. It is often visualized to provide contextual summaries [54,92], reflect individuals' sentiments [16,66], and analyze large-scale similarities [34,61].

Basic: In basic visualization techniques, Scatterplot [13,41], Bar Chart [46,67], and Area Chart [18,61] are frequently used for text (Fig. 6). To apply these techniques to text, researchers often need to extract high-level information from the raw text. For example, Zhang et al. [34] first used LDA to extract ten themes from tourist posts. Then they applied t-SNE to reduce the dimensionality of each post which was depicted by ten theme probability values. Finally, all posts were projected on a scatterplot (Fig. 9A). With the scatterplot, users could easily navigate to a set of tourist posts (i.e., tourism routes) of interest based on themes. Wu et al. [48] classified the sales pitches of live streaming into six categories by applying ERNIE 3.0 [93] to sentence data. With the labels of each sentence, they visualized the temporal distribution of sales pitches by an area chart (Fig. 9B). Furthermore, they employed SHAP values [94] as model explanation metrics to display the positive/negative contributions of different sales pitches to time-series modeling. The contributions were presented by a series of

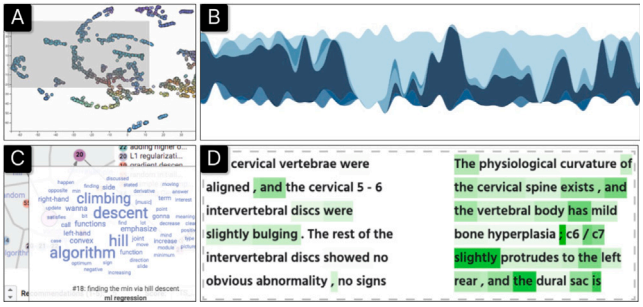

Fig. 9. Examples of text visualization. (A) Zhang et al. [34] projected tourist posts on a scatterplot. (B) Wu et al. [48] leveraged area charts to display sales pitches. (C) Cao et al. [92] used a word cloud to display the summary of users' communication records. (D) Ouyang et al. [41] visualized the weight of each word in multimodal prediction models through a heatmap.

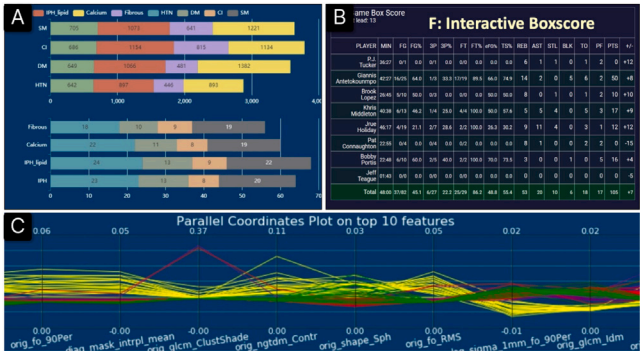

Fig. 10. Examples of tabular data visualization. (A) Xu et al. [45] visualized the relationship among carotid artery plaque and other patients' characteristic by a stacked bar chart. (B) Fu et al. [55] displayed the raw tabular data directly. (C) Xu et al. [45] leveraged PCP to visualize multi-dimensional clinical data.

bar charts in time order to help analysts identify the key sales pitches influencing streaming statistics.

Advanced: Text, itself can provide descriptive information distinctly and accurately. Therefore, the most frequently used visualization technique for text is Word-based. Word-based visualization consists of Text and Word Cloud. Text refers to displaying the raw text data directly for analysts. It is usually used as a context for analysis [18,54]. Word Cloud is often employed to present summative information [32,95]. For example, both Cao et al. [92] and Zhao et al. [77] used a word cloud to display the summary of users' communication records and video content, respectively (Fig. 9C). Zeng et al. [16] used Tone Analyzer API[1] to extract emotion tones of each sentence within a video. They further employed a word cloud to highlight important words in sentences with different emotion tones. The word cloud provides a rich context for presenting the emotion coherence among the face channel, text channel, and audio channel.

Other popular techniques include Glyph-based [49,66], Grid-based [13,35], Line-based [46,48], and Graph-based [32,92] visualizations. For example, Wu et al. [48] designed a glyph to encode the contributions of different text features to time-series models. In addition, Ouyang et al. [41] visualized the weight of each word in multimodal prediction models through a heatmap (Fig. 9D).

5.3. Tabular data

Tabular data is very common in multimodal data analysis. Considering its simple data format, it is easily visualized by various visualization techniques.

Basic: In basic visualization techniques, tabular data is most often visualized by the Bar Chart based on the statistical results of its particular dimensions [41,78]. For example, patients' clinical recording data which includes demographics, chronic disease information, etc., were visualized by a stacked bar chart to [45] reveal the relationship among carotid artery plaque and other patients' characteristics (Fig. 10A). Due to its high dimension, dimension reduction algorithms, such as t-SNE, were used to plot tabluar data in a scatterplot [30,45] to provide an overview of the whole dataset. Moreover, some researchers [19,55] chose to directly display the raw tabular data to provide context since tabular data itself is straightforward for analysts (Fig. 10B).

Advanced: Tabular data's intrinsic nature of multi-dimension facilitates the application of Line-based visualization, especially Parallel Coordinate Plot (PCP). For example, Xu et al. [45], Mörth et al. [30], and Jönsson et al. [31] all utilized PCP to present multi-dimensional clinical data (Fig. 10C). Chen et al. [72] used PCP to visualize online learners' profiles and clickstream data, revealing the correlation between learners and video segments with many clickstreams. Glyph-based visualization [52,78,96] is also effective for multi-dimensional tabular data. Moreover, tabular data could also be transformed into other data formats, such as graph and geolocation, and visualized by Grid-based [19,68], Graph-based [13,19], and Map-based [18,72] visualizations.

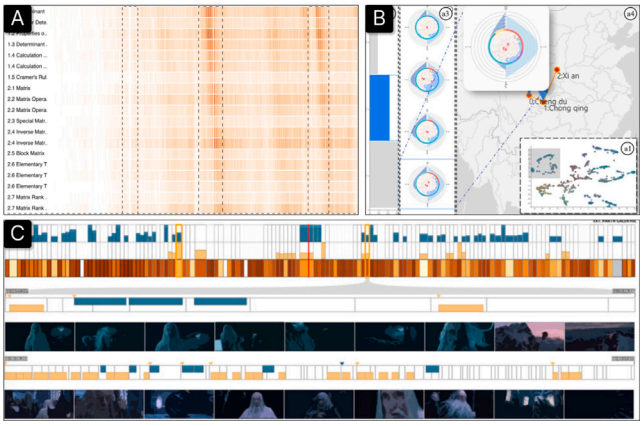

Fig. 11. Examples of event sequence visualization. (A) Chen et al. [74] used stripes to display the temporal distribution of user comments. (B) Zhang et al. [34] designed a glyph based on maps to show an overview of a travel route. (C) Kurzhals et al. [97] used a heatmap to present details in a movie.

5.4. Event sequences

Event sequences are usually visualized to facilitate users' behaviors in multimodal data analysis [42,71,73].

Basic: Important statistics of event sequences are often presented by Bar Chart [16,55] (Fig. 6). For example, Zhang et al. [34] counted the number of destinations in each travel route and presented the result by a bar chart to allow users to navigate the travel routes based on length. Similar cases also occur in Anchorage [42] where customers' satisfaction scores derived from machine logs were visualized by bar charts. Another popular technique for event sequences is Stripe (Fig. 6), which is often used to show the distribution of various events [67,78]. For example, DanmuVis [74] displays the temporal distribution of different danmu comments by stripes based on users' streaming comments (Fig. 11A). As for the sequence containing various kinds of events, the events can be visualized as unit visualizations in temporal order. These units are usually colored differently to distinguish them [98].

Advanced: Glyph-based technique is most preferred in advanced visualizations (Fig. 7) by visualizing multi-dimensional event features

[1] https://www.ibm.com/watson/services/tone-analyzer/

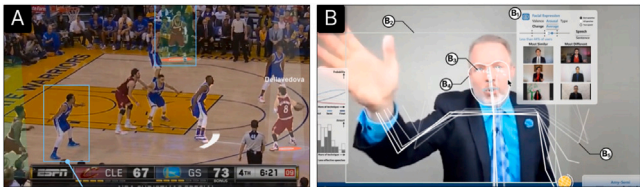

Fig. 12. Examples of video visualization. (A) Chen et al. [52] embedded visualizations into basketball videos. (B) Huang et al. [46] overlayed skeleton points on the video to enhance the understanding.

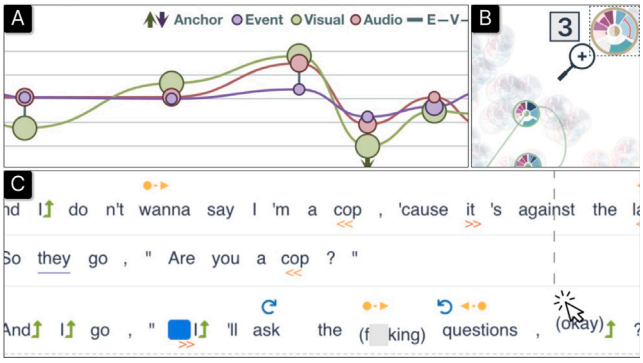

Fig. 13. Examples of audio visualization. (A) Wong et al. [42] used line charts to visualize different vocal features. (B) Wu et al. [48] designed a circular glyph to encode audio features including pitch and volume. (C) Wang et al. [67] leveraged customized icons to display audio patterns.

with customized glyphs [72,73]. For example, Zhang et al. [34] extracted the location of each destination within a travel rote and designed a map-based glyph to provide an overview of the geographic mode of the travel route (Fig. 11B). In addition, researchers also used Heatmap [97] (Fig. 11C) and Matrix [67,73] (Grid-based) to illustrate details of even sequences (Fig. 7).

5.5. Video

Videos are often segmented into single frames/images (Section 5.1), transformed into audio (Section 5.6), or transcribed into text (Section 5.2) in multimodal analysis. We have discussed these cases separately. When videos are treated as a whole, all of them are visualized using Image-based techniques, especially Video player and Augmented video (Fig. 7). Video player [42,47] enables analysts to browse the original videos for more details. Augmented video overlays visual hints on the original videos. For example, iBall [52] embeds visualizations presenting players' statistics into basketball videos to enhance fans' game comprehension and engagement (Fig. 12A). SpeechMirror [46] overlays the face mark key points, the direction of eye gaze, and the skeleton of the upper body of a speaker on the original video to enhance the understanding of his/her presentation techniques in a speech (Fig. 12B).

5.6. Audio

Audio is sometimes transcribed into text (e.g., subtitles) in multimodal analysis. You can refer to Section 5.2 for more details about visualization of text. Beyond these cases, audio is often visualized for the analysis of speech performance [46,48] and sentiment [18,42].

Basic: Audio features such as tone, pitch, and pauses are mostly visualized by Bar Chart to facilitate performance evaluation of speakers (Fig. 6). For example, Huang et al. [46], Wu [48], and Wang [67] all used bar charts to present the statistical results of diverse vocal features.

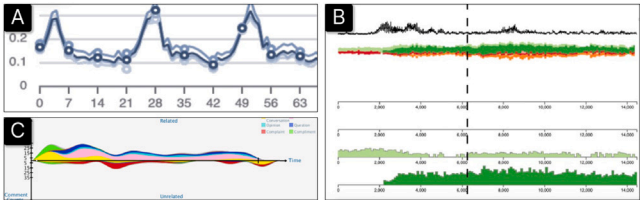

Fig. 14. Examples of time series visualization. (A) Jiang et al. [19] displayed the temporal distribution of motion features by line charts. (B) Chan et al. [29] used line charts and area charts to present motion features varying over time. (C) Sung et al. [61] used area charts to display the number of related and unrelated comments over time.

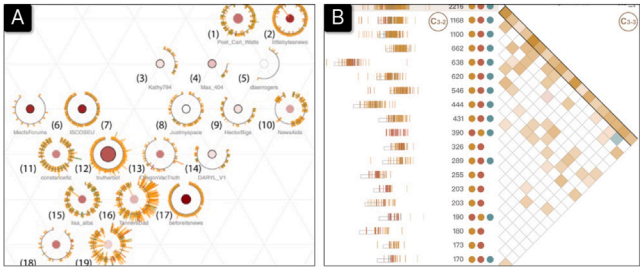

Fig. 15. Examples of graph visualization. (A) Cao et al. [77] presented important features on social media based on their designed glyph. (B) Zhang et al. [78] leveraged a matrix to present the quantitative distribution of events that happened between two figures.

In addition to Bar Chart, Line Chart [18,42] also serves as an effective way to reveal the variation of various vocal features (Fig. 13A).

Advanced: Similar to image (Section 5.1), the most frequent technique for audio is Glyph-based. For example, both Wang et al. [13] and Wu et al. [48] designed a circular glyph to encode audio features such as pitch and volume (Fig. 13B). Besides, Wang et al. [67] and Soure et al. [66] used customized icons as glyphs to encode various audio patterns, such as fast speech rate, high pitch, and low loudness (Fig. 13C). Other works also employed Line-based (i.e., Storyline [15] and Polar Chart [48]) and Grid-based techniques (i.e., Heatmap [67]) to visualize similar audio features.

5.7. Time series

Time series data is rare in multimodal data analysis (Fig. 5). In HealthPrism [19], the inertial sensor data is intuitively visualized by line charts to illustrate the distribution of motion features (Fig. 14A). In Motion Browser [29], similar data is visualized by line charts and area charts to facilitate analysis of upper limb movement (Fig. 14B). Additionally, Sung et al. [61] used area charts to present the temporal distribution of the number of related and unrelated comments (Fig. 14C). Moreover, Chen et al. [52] embedded customized icons into basketball videos to highlight the dynamic positions of basketball players.

5.8. Graph

Graphs record relationships between entities. The most intuitive visualization for this modality is Graph-based techniques [68] (Fig. 7). To encode more information about entities, Glyph-based techniques (i.e., Glyph) will be designed to present important features such as users' posting and responding patterns [77] (Fig. 15A). To be mentioned, another efficient technique for displaying relationships is Matrix in Grid-based techniques. Zhang et al. [78] integrated a matrix and a heatmap to present the quantitative distribution of events that happened between two figures (Fig. 15B).

Fig. 16. Examples of geolocation visualization. (A) Chen et al. [35] leveraged a semantic map on base maps to display different information. (B) Zhang et al. [34] leveraged a base map to present travel routes.

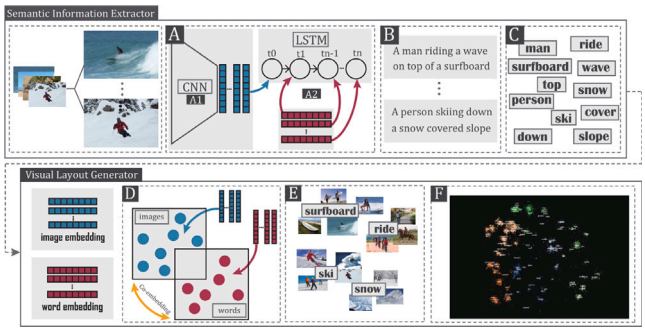

Fig. 17. An example pipeline of semantic fusion [43].

5.9. Geolocation

Geolocation is also a rare modality in multimodal data analysis (Fig. 5). Only two papers visualize this modality. Both of them employed a map to illustrate the geographical context during the analysis. Chen et al. [35] applied topic modeling to comments for user-generated POIs in urban space (Fig. 16A). They further overlaid a semantic map on base maps to display the quantitative distribution, impact on urban performance, and semantic information of POIs. Similarly, Zhang et al. [34] leveraged a base map to present travel routes derived from comments for user-generated POIs (Fig. 16B). Moreover, they designed a map-based glyph to present the geographical patterns of frequent routes.

6. Multimodal visualization

According to the fusion strategies of different data modalities, multimodal visualizations can be divided into four categories, namely, semantic fusion, visual integration, visual juxtaposition, and cross-view linking. The visual cohesion between different modalities progressively weakens.

6.1. Semantic fusion

Semantic fusion involves combining data from different modalities at the data or feature level before visualization. The aim is to create a unified representation of the data that integrates the semantic information from all sources [41,43]. For example, Xie et al.'s visualization pipeline [43] includes a co-embedding approach to fuse the semantics of images and their descriptions, as shown in Fig. 17.

The visualization of the fused semantics can be used to serve as analysis entries. For example, in Wong et al.'s study [42], the extracted facial and audio features are used to further classify emotions into positive, negative, and neutral, and design a line chart to provide an overview as an exploration entry. In Zhang et al.'s study [78], the features of different modalities are first weighted and summed for every history figure. These fused features were then used to group the figures into cohorts, which are the basis of the following analyses. In Li et al.'s study [73], the numbers of abnormalities in each modality in the online exam scenario were weighted and summed to determine the degree of abnormality in the online exam.

Besides, there are some studies that focus on evaluating the semantic fusion of multimodal data. Most of the visualizations in these papers are designed to interpret the prediction based on the multimodal fusion results. Often, the interpretation requires reverting to unimodal visualizations with other general visualizations, such as trees [13] and parallel coordinate plots [19]. For example, Wang et al. [13] designed a tree-like visualization to present the conflict, dominance, and complement situation of modalities in sentiment analysis. Furthermore, Ye et al. [99] combined the projection with a well-designed "concept axis" to steer the alignment of multimodal fusion.

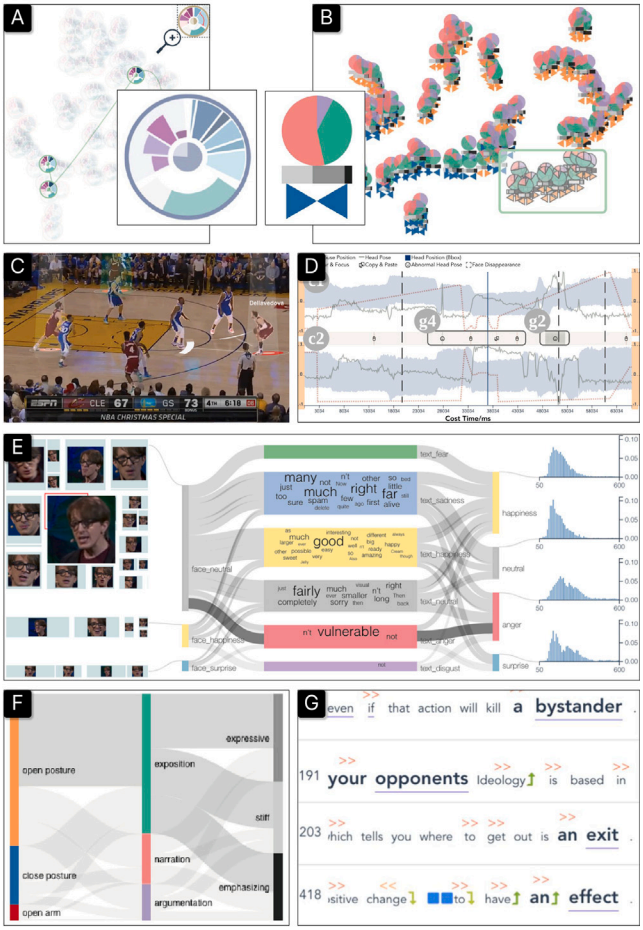

Fig. 18. Examples of visual integration. (A and B) We et al.'s [48] and Wu and Qu's [51] glyph designs encoding multimodal data, respectively. (C) Chen et al.'s method [52] overlaid basketball player statistics onto the game video. (D) Li et al.'s method [73] overlaid mouse position, head pose, and head position in one chart. (E) Zeng et al. [16] extended (F) Wu and Qu's design [51] by integrating visualizations of each modality into a Sankey diagram. (G) Wang et al.'s design [33] integrated the verbal feature and textual content.

6.2. Visual integration

One strategy for multimodal visualization is visual integration, which tightly integrates visualizations from various modalities through a well-designed visual layout. This strategy primarily includes three methods: glyphs, overlaying, and embedding.

Glyph. Multimodal data can be considered a form of multi-dimensional data. Glyphs are one of the multi-dimensional visualizations that effectively integrate information from multiple dimensions [100,101]. For example, in Wu et al.'s study [48], the emotions and selling

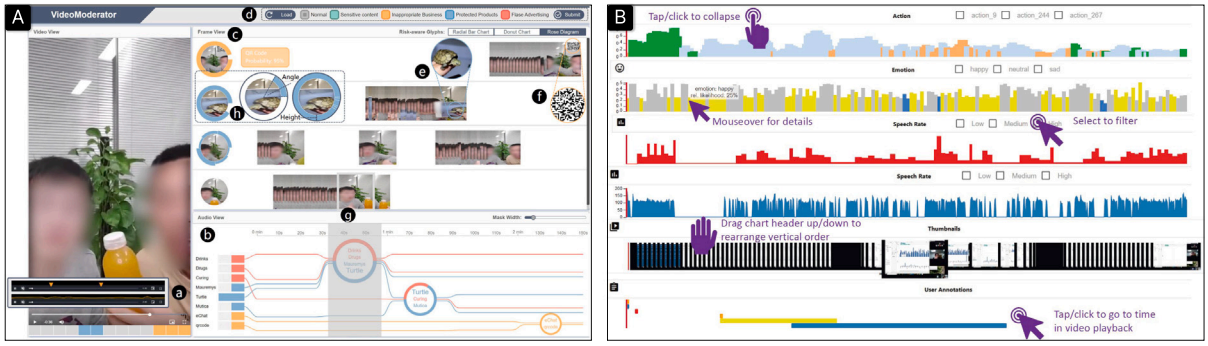

Fig. 19. Examples of visual juxtaposition. (A) Tang et al. [15] juxtaposed the audio and image (video frame) visualizations along the timeline. (B) Batch et al. [47] juxtaposed the visualizations of six modalities collected from a user study video along the timeline.

techniques for every live commerce video clip were first detected from text, audio and frame modalities, respectively, and encoded with a pie-like glyph. The glyphs of video clips are plotted as a scatterplot (Fig. 18A), where users can easily locate semantically similar video clips and identify the distribution of each modality. Similar designs can be found in Zeng et al.'s study [16] and Wu and Qu's study [51] (Fig. 18B).

Overlaying. Overlaying means that the visualizations of different modalities are directly superimposed onto one view according to the same coordinate system [35,42]. For example, Chen et al. [52] overlaid the statistics of basketball players onto the basketball game video (Fig. 18C). Li et al. [73] overlaid the mouse position (red line), head pose (blue line), and head position (blue area) extracted from the online exam video onto the same time-oriented chart (Fig. 18D). Similarly, to track the cell's growth, Lange et al. [80] displayed the cell's division trajectories with the reference of the cell image.

Embedding. Different from overlaying, the embedding method aims to embed multimodal data into some kind of visualization. For example, the presentation techniques extracted from the posture, gesture, and transcripts can be encoded with a Sankey-like design [51] (Fig. 18F). Zeng et al. [16] extended such a design by embedding word cloud, facial images, and histogram into the nodes of the Sankey (Fig. 18E), showing the co-occurrence of emotions across modalities. Besides, audio features (e.g., speed, pauses, and loudness) can be represented symbolically and then embedded into text visualizations (Fig. 18E), revealing both the audio and text modalities [33,67]. To track the cell's growth, Lange et al. [80] nested the cell images with the cell's division tree.

Guo et al. [81] considered text-to-image applications. To reveal the impact of the prompt changes on generated images, the words or phrases denoting prompt changes were embedded into the graph-based image layouts. Such a visual integration supports inspecting which words or phrases in the prompts lead to the difference between images.

6.3. Visual juxtaposition

Visual juxtaposition means the visualizations of different modalities are visually aligned with their common reference. This strategy is most commonly seen in the scenarios with a timeline, such as speech [67] and video [14] analyses. A speech involves the modalities of textual transcripts [54] and speech audio [66], and if the speech is recorded via video, it also involves the image modality [48] (i.e., video frame) from which gesture [50], face [16], and posture [73] can be further extracted. All of the visualizations of these modalities can be juxtaposed along the timeline, empowering users with a comprehensive progression analysis.

Fig. 19 shows two examples. Fig. 19A is Tang et al.'s visualization system [15] for moderating live commerce videos. The upper right part shows the consecutive frames (images) of key video clips, and the lower right part is the Storyline of the keywords mentioned in the video audio. These two parts are aligned based on the video's timestamps. Fig. 19B is Batch et al.'s visualization system [47] for evaluating the subject's experiences on using a certain prototype. The system consists of six views from top to bottom, aligned in time from left to right according to the user study. Each view corresponds to a modality or information extracted from a modality, such as the user action records in the first view and the user emotions in the second view. Visual juxtaposition facilitates an intuitive comparison of conflicts and consistencies between different modalities [102], while aligning them along the timeline helps to reveal temporal progressive patterns.

6.4. Cross-view linking

The above two strategies achieve multimodal data analysis through the well-designed layout of visual elements. Additionally, some works achieve multimodal visual analysis through interaction-based visual linking across views, a strategy we call cross-view linking. This strategy is particularly useful when data from different modalities lack a common reference or when the visualization of a single modality is already dense with potential visual clutter.

The former situation is commonly seen in medical domains [30,31,41,45,84]. For example, Xu et al. [45] coordinated the medical visualizations of medical indicators and medical images (Fig. 20A), which may be because these three modalities cannot be visually aligned due to the lack of a common reference. An example of the latter situation is Chen et al. study [53]. They linked labels and images to guide the image captioning (Fig. 20B), which is scalable for large numbers of labels and images. Besides, in Zeng et al.'s study [50], gesture glyphs and speech content for every speech clip are plotted on two scatterplots, respectively. Users can brush one of the scatterplots, and the corresponding clips in another scatterplot will be highlighted. In another view of their system, gestures and speech content are juxtaposed along the speech progress.

Note that cross-view linking can be added to the visual integration and juxtaposition strategies. In our survey, we only consider and discuss the multimodal visualizations enhanced with cross-view linking only.

7. Research opportunity

In this section, we introduced research opportunities to use visualization techniques to analyze multimodal data. We focused on three aspects, including multimodal AI, AI for multimodal visualization, and application in domains.

7.1. Consideration of more modalities

In this survey, data mostly come from videos, images, and text according to the source modality in Fig. 3. Other modalities such as inertial sensors and proprioception [36] are rarely or even never mentioned in existing works. Although current multimodal AI has

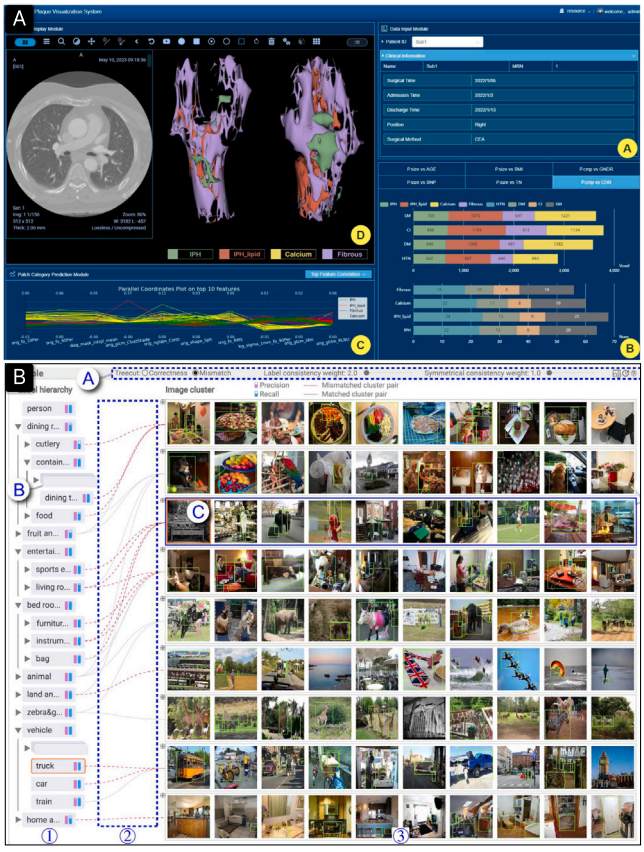

Fig. 20. Examples of cross-view linking. (A) Xu et al.'s system [45] visualized medical images and numeric indicators with linked views. (B) Chen et al.'s method [53] visually linked the textual labels and images.

experienced rapid development, most of them excel at extracting information from primary modalities (e.g., videos, images, and text) [103–109]. However, they still face limitations when processing data from other less common modalities [28]. Further studies on multimodal AI are needed to improve the capability to extract meaningful information from diverse data modalities, posing multiple AI research challenges throughout the whole lifecycle of AI models [110].

For example, during model development, one critical challenge is designing high-quality benchmarks. Existing benchmarks mainly target language and vision, leaving other modalities understudied [36]. Comprehensive datasets that include diverse modalities such as inertial sensors, physiological data, and optical flow are urgently needed. By establishing these benchmarks, researchers can more effectively measure progress and identify areas needing improvement in training multimodal AI. With powerful multimodal AI models, more valuable information can be extracted from diverse data modalities and visualized for analysts to discover valuable insights.

7.2. Better multimodal artificial intelligence (AI)

Vis4AI (Visualization for AI models) aims to diagnose and improve AI models via visualization techniques, a popular topic in the past decade. The previous Vis4AI techniques tend to visualize the parameters within the model to open the black box for better diagnosis and improvement. However, the recent development of foundation models makes these techniques inapplicable due to the prohibited number of parameters. Thus, the visualization techniques have been shifted to the input and output side and evaluate input–output pairs [111–113]. The multimodal large language model (MLLM) is a kind of foundation model, which allows inputs and outputs in various modalities. Many studies have proven its power, such as in detecting misleading visualization [114] and question answering [115]. However, many weaknesses still exist, for example, the notorious hallucination [116]. For MLLMs, multimodal visualizations become a potential technique that supports users in navigating multimodal inputs and outputs, comparing input–output pairs, and finally evaluating large multimodal models [44,81,117].

Visualization for large multimodal models is an emerging area that still requires substantial efforts from the community. We believe this survey provides design guidelines for practitioners to design visualizations and visual analytics approaches for diagnosing and improving large multimodal models.

7.3. Artificial intelligence (AI) for multimodal visualization

AI models have been increasingly applied to the visualization community [118–121]. Such a research topic is called AI4VIS [122]. According to Wang et al.'s work [123], AI4VIS methods can be roughly categorized into two kinds, visualization understanding [124] and visualization generation or recommendation [125]. Many works have tried to leverage advanced AI techniques, such as reinforcement learning [126], and RuleFit Binary Classifier [127] to recommend/generate visualizations based on the data and tasks provided by users [128]. However, users can only input tabular data in these works, which limits their generalizability to multimodal data visualizations. Generating visualizations from multimodal data requires AI models to discover the most effective way to combine different data modalities into a cohesive and informative visual representation. This survey provides a comprehensive summary of visualization methods for single modality and multiple modalities, which can provide rich expertise in multimodal visualizations for generation models to learn. Challenges may lie in two aspects, data processing and visualization fusion.

First, it is difficult for models to decide the features that should be extracted from multimodal data. According to our survey, tasks given by users always play an important role in determining the features for visualization. However, these tasks are often described in a high-level format. The model needs to understand users' intents underlying the tasks and decide what features should be extracted from each data modality.

Second, visualization generation is also challenging due to the need for modality fusion. The generated visualizations need to seamlessly merge different modalities while maintaining their intrinsic relevance (e.g., alignment of different data modalities in time and space). This requires models to identify the relevance among different data modalities before generating visualizations, which is quite challenging.

Addressing these challenges requires interdisciplinary efforts combining AI advances, human–computer interaction, data visualization, and application domains. Developing robust models and systems that can handle multimodal data will significantly enhance the ability to generate effective and efficient visualizations, ultimately leading to better multimodal data analysis.

7.4. Application in domains

According to Fig. 3, most of the existing works are applied to e-learning and health & medical. The analysis problems in these domains have been studied comprehensively. More research focus should be placed in other domains, especially urban, history, and sports. Specifically, in the urban domain, there are considerable multimodal data [129,130], such as transportation data, environmental data, and social media data. Potential directions can be how to combine multimodal data to facilitate urban management including transportation improvement, pollution control, and travel planning. In the history domain, data such as text from historical documents, images of artifacts, audio recordings of oral histories, and geospatial data from archaeological sites contain rich interpretations of historical events.

Multimodal visualizations can help historians and researchers uncover new patterns and connections that are not apparent when analyzing a single data modality. In the sports domain, integrating data from video footage, player statistics, and social media can offer deeper insights into player performance, fan engagement, and game strategies. Besides, combining video data, biometric data, and inertial sensor data can help develop personalized training programs and injury prevention strategies.

8. Conclusion

In this survey, we explored the methodologies for applying visualization to multimodal analysis from the perspectives of data, processing, modality visualization, and multimodal visualization, emphasizing the roles of computational and visualization techniques in human–machine collaborative multimodal analysis. We particularly discussed how visualization aids in multimodal analysis by enabling the intuitive presentation, integration, and comparison of different modalities. Finally, we identified future research trends and challenges, with a particular focus on AI4VIS's potential to enhance intelligent multimodal visualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The work was supported by the Guangdong Provincial Fund for Basic and Applied Basic Research—Regional Joint Fund Project (Key Project) (2023B1515120078) and the National Natural Science Foundation of China (62402184, 62402428).

Data availability

Data will be made available on request.

References

[1] D. Lahat, T. Adali, C. Jutten, Multimodal data fusion: An overview of methods, challenges, and prospects, Proc. IEEE 103 (9) (2015) 1449–1477.

[2] J. Feng, G. Wang, C. Zheng, Y. Cai, Z. Fu, Y. Wang, X. Wei, Q. Li, Towards bridged vision and language: Learning cross-modal knowledge representation for relation extraction, IEEE Trans. Circuits Syst. Video Technol. 34 (1) (2024) 561–575, http://dx.doi.org/10.1109/TCSVT.2023.3284474.

[3] M. Suzuki, Y. Matsuo, A survey of multimodal deep generative models, Adv. Robot. 36 (5–6) (2022) 261–278, http://dx.doi.org/10.1080/01691864.2022.2035253.

[4] W. Fang, J. Xie, H. Liu, J. Chen, Y. Cai, Diverse visual question generation based on multiple objects selection, ACM Trans. Multim. Comput. Commun. Appl. 20 (6) (2024) 161:1–161:22, http://dx.doi.org/10.1145/3640014.

[5] T. Baltrusaitis, C. Ahuja, L. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2019) 423–443, http://dx.doi.org/10.1109/TPAMI.2018.2798607.

[6] C. Zhang, Z. Yang, X. He, L. Deng, Multimodal intelligence: Representation learning, information fusion, and applications, IEEE J. Sel. Top. Signal Process. 14 (3) (2020) 478–493, http://dx.doi.org/10.1109/JSTSP.2020.2987728.

[7] Z. Wu, C. Zheng, Y. Cai, J. Chen, H. Leung, Q. Li, Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts, in: Proceedings of ACM International Conference on Multimedia, 2020, pp. 1038–1046, http://dx.doi.org/10.1145/3394171.3413650.

[8] J. Xie, W. Fang, Y. Cai, Q. Huang, Q. Li, Knowledge-based visual question generation, IEEE Trans. Circuits Syst. Video Technol. 32 (11) (2022) 7547–7558, http://dx.doi.org/10.1109/TCSVT.2022.3189242.

[9] J. Xie, Y. Cai, J. Chen, R. Xu, J. Wang, Q. Li, Knowledge-augmented visual question answering with natural language explanation, IEEE Trans. Image Process. 33 (2024) 2652–2664, http://dx.doi.org/10.1109/TIP.2024.3379900.

[10] Q. Huang, Y. Liang, J. Wei, Y. Cai, H. Liang, H. Leung, Q. Li, Image difference captioning with instance-level fine-grained feature representation, IEEE Trans. Multim. 24 (2022) 2004–2017, http://dx.doi.org/10.1109/TMM.2021.3074803.

[11] Q. Huang, P. Li, Y. Huang, F. Shuang, Y. Cai, Region-focused network for dense captioning, ACM Trans. Multim. Comput. Commun. Appl. 20 (6) (2024) 183:1–183:20, http://dx.doi.org/10.1145/3648370.

[12] S. Poria, E. Cambria, N. Howard, G.-B. Huang, A. Hussain, Fusing audio, visual and textual clues for sentiment analysis from multimodal content, Neurocomputing 174 (2016) 50–59.

[13] X. Wang, J. He, Z. Jin, M. Yang, Y. Wang, H. Qu, M2lens: Visualizing and explaining multimodal models for sentiment analysis, IEEE Trans. Vis. Comput. Graphics 28 (1) (2021) 802–812.

[14] A. Wu, H. Qu, Multimodal analysis of video collections: Visual exploration of presentation techniques in ted talks, IEEE Trans. Vis. Comput. Graphics 26 (7) (2018) 2429–2442.

[15] T. Tang, Y. Wu, Y. Wu, L. Yu, Y. Li, Videomoderator: A risk-aware framework for multimodal video moderation in e-commerce, IEEE Trans. Vis. Comput. Graphics 28 (1) (2021) 846–856.

[16] H. Zeng, X. Wang, A. Wu, Y. Wang, Q. Li, A. Endert, H. Qu, EmoCo: Visual analysis of emotion coherence in presentation videos, IEEE Trans. Vis. Comput. Graph. 26 (1) (2020) 927–937, http://dx.doi.org/10.1109/TVCG.2019.2934656.

[17] H. Zeng, X. Shu, Y. Wang, Y. Wang, L. Zhang, T. Pong, H. Qu, EmotionCues: Emotion-oriented visual summarization of classroom videos, IEEE Trans. Vis. Comput. Graph. 27 (7) (2021) 3168–3181, http://dx.doi.org/10.1109/TVCG.2019.2963659.

[18] K.T. Maher, Z. Huang, J. Song, X. Deng, Y. Lai, C. Ma, H. Wang, Y. Liu, H. Wang, E-ffective: A visual analytic system for exploring the emotion and effectiveness of inspirational speeches, IEEE Trans. Vis. Comput. Graph. 28 (1) (2022) 508–517, http://dx.doi.org/10.1109/TVCG.2021.3114789.

[19] Z. Jiang, H. Chen, R. Zhou, J. Deng, X. Zhang, R. Zhao, C. Xie, Y. Wang, E.H. Ngai, HealthPrism: A visual analytics system for exploring children's physical and mental health profiles with multimodal data, IEEE Trans. Vis. Comput. Graphics 30 (01) (2024) 1205–1215, http://dx.doi.org/10.1109/TVCG.2023.3326943.

[20] Q. Wang, Z. Chen, Y. Wang, H. Qu, A survey on ML4VIS: Applying machine learning advances to data visualization, IEEE Trans. Vis. Comput. Graph. 28 (12) (2022) 5134–5153, http://dx.doi.org/10.1109/TVCG.2021.3106142.

[21] A. Wu, Y. Wang, X. Shu, D. Moritz, W. Cui, H. Zhang, D. Zhang, H. Qu, AI4VIS: survey on artificial intelligence approaches for data visualization, IEEE Trans. Vis. Comput. Graph. 28 (12) (2022) 5049–5070, http://dx.doi.org/10.1109/TVCG.2021.3099002.

[22] Z. Zhou, Z. Yang, S. Jiang, T. Zhu, S. Ma, Y. Li, J. Zhuo, Design and validation of a navigation system of multimodal medical images for neurosurgery based on mixed reality, Vis. Inform. 7 (2) (2023) 64–71, http://dx.doi.org/10.1016/J.VISINF.2023.05.003.

[23] Z. Yuan, S. He, Y. Liu, L. Yu, MEinVR: Multimodal interaction techniques in immersive exploration, Vis. Inform. 7 (3) (2023) 37–48, http://dx.doi.org/10.1016/J.VISINF.2023.06.001.

[24] J. Kehrer, H. Hauser, Visualization and visual analysis of multifaceted scientific data: A survey, IEEE Trans. Vis. Comput. Graphics 19 (3) (2013) 495–513, http://dx.doi.org/10.1109/TVCG.2012.110.

[25] X. Huang, H. Miao, A. Townsend, K. Champley, J. Tringe, V. Pascucci, P.-T. Bremer, Bimodal visualization of industrial X-Ray and neutron computed tomography data, IEEE Trans. Vis. Comput. Graphics (2024).

[26] K. Lawonn, N. Smit, K. Bühler, B. Preim, A survey on multimodal medical data visualization, Comput. Graph. Forum 37 (1) (2018) 413–438, http://dx.doi.org/10.1111/cgf.13306.

[27] Y. Wu, X. Xie, J. Wang, D. Deng, H. Liang, H. Zhang, S. Cheng, W. Chen, ForVizor: Visualizing spatio-temporal team formations in soccer, IEEE Trans. Vis. Comput. Graphics 25 (1) (2019) 65–75, http://dx.doi.org/10.1109/TVCG.2018.2865041.

[28] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2018) 423–443.

[29] G.Y.-Y. Chan, L.G. Nonato, A. Chu, P. Raghavan, V. Aluru, C.T. Silva, Motion browser: Visualizing and understanding complex upper limb movement under obstetrical brachial plexus injuries, IEEE Trans. Vis. Comput. Graphics 26 (1) (2020) 981–990, http://dx.doi.org/10.1109/TVCG.2019.2934280.

[30] E. Mörth, K. Wagner-Larsen, E. Hodneland, C. Krakstad, I. Haldorsen, S. Bruckner, N.N. Smit, RadEx: Integrated visual exploration of multiparametric studies for radiomic tumor profiling, Comput. Graph. Forum 39 (7) (2020) 611–622, http://dx.doi.org/10.1111/CGF.14172.

[31] D. Jönsson, A. Bergström, C. Forsell, R. Simon, M. Engström, S. Walter, A. Ynnerman, I. Hotz, VisualNeuro: A hypothesis formation and reasoning application for multi-variate brain cohort study data, Comput. Graph. Forum 39 (6) (2020) 392–407, http://dx.doi.org/10.1111/CGF.14045.

[32] A. Karduni, I. Cho, R. Wesslen, S. Santhanam, S. Volkova, D.L. Arendt, S. Shaikh, W. Dou, Vulnerable to misinformation?: Verifi!, in: W. Fu, S. Pan, O. Brdiczka, P. Chau, G. Calvary (Eds.), Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI 2019, Marina Del Ray, CA, USA, March 17-20, 2019, ACM, 2019, pp. 312–323, http://dx.doi.org/10.1145/3301275.3302320.

[33] X. Wang, H. Zeng, Y. Wang, A. Wu, Z. Sun, X. Ma, H. Qu, VoiceCoach: Interactive evidence-based training for voice modulation skills in public speaking, in: Proceedings of CHI Conference on Human Factors in Computing Systems, ACM, 2020, pp. 1–12, http://dx.doi.org/10.1145/3313831.3376726.

[34] X. Zhang, X. Pang, X. Wen, F. Wang, C. Li, M. Zhu, TriPlan: an interactive visual analytics approach for better tourism route planning, J. Vis. 26 (1) (2023) 231–248, http://dx.doi.org/10.1007/S12650-022-00861-8.

[35] J. Chen, Q. Huang, C. Wang, C. Li, SenseMap: Urban performance visualization and analytics via semantic textual similarity, IEEE Trans. Vis. Comput. Graphics (2023).

[36] P.P. Liang, Y. Lyu, X. Fan, Z. Wu, Y. Cheng, J. Wu, L.Y. Chen, P. Wu, M.A. Lee, Y. Zhu, et al., MultiBench: Multiscale benchmarks for multimodal representation learning, in: Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1), 2021.

[37] R. Amar, J. Eagan, J. Stasko, Low-level components of analytic activity in information visualization, in: IEEE Symposium on Information Visualization, 2005, INFOVIS 2005., IEEE, 2005, pp. 111–117.

[38] H.-J. Schulz, T. Nocke, M. Heitzler, H. Schumann, A design space of visualization tasks, IEEE Trans. Vis. Comput. Graphics 19 (12) (2013) 2366–2375.

[39] D. Deng, Y. Wu, X. Shu, J. Wu, S. Fu, W. Cui, Y. Wu, VisImages: A fine-grained expert-annotated visualization dataset, IEEE Trans. Vis. Comput. Graphics 29 (7) (2023) 3298–3311, http://dx.doi.org/10.1109/TVCG.2022.3155440.

[40] J. Knittel, S. Koch, T. Tang, W. Chen, Y. Wu, S. Liu, T. Ertl, Real-time visual analysis of high-volume social media posts, IEEE Trans. Vis. Comput. Graph. 28 (1) (2022) 879–889, http://dx.doi.org/10.1109/TVCG.2021.3114800.

[41] Y. Ouyang, Y. Wu, H. Wang, C. Zhang, F. Cheng, C. Jiang, L. Jin, Y. Cao, Q. Li, Leveraging historical medical records as a proxy via multimodal modeling and visualization to enrich medical diagnostic learning, IEEE Trans. Vis. Comput. Graph. 30 (1) (2024) 1238–1248, http://dx.doi.org/10.1109/TVCG.2023.3326929.

[42] K.K. Wong, X. Wang, Y. Wang, J. He, R. Zhang, H. Qu, Anchorage: Visual analysis of satisfaction in customer service videos via anchor events, IEEE Trans. Vis. Comput. Graphics (2023) 1–13, http://dx.doi.org/10.1109/TVCG.2023.3245609.

[43] X. Xie, X. Cai, J. Zhou, N. Cao, Y. Wu, A semantic-based method for visualizing large image collections, IEEE Trans. Vis. Comput. Graph. 25 (7) (2019) 2362–2377, http://dx.doi.org/10.1109/TVCG.2018.2835485.

[44] Y. Feng, X. Wang, K. Wong, S. Wang, Y. Lu, M. Zhu, B. Wang, W. Chen, PromptMagician: Interactive prompt engineering for text-to-image creation, IEEE Trans. Vis. Comput. Graph. 30 (1) (2024) 295–305, http://dx.doi.org/10.1109/TVCG.2023.3327168.

[45] C. Xu, Z. Zheng, Y. Fu, B. Chang, L. Chen, M. Wu, M. Song, J. Jiang, An integrated visual analytics system for studying clinical carotid artery plaques, J. Vis. (2024) 1–13.

[46] Z. Huang, Q. He, K.T. Maher, X. Deng, Y. Lai, C. Ma, S.F. Qin, Y. Liu, H. Wang, SpeechMirror: A multimodal visual analytics system for personalized reflection of online public speaking effectiveness, IEEE Trans. Vis. Comput. Graph. 30 (1) (2024) 606–616, http://dx.doi.org/10.1109/TVCG.2023.3326932.

[47] A. Batch, Y. Ji, M. Fan, J. Zhao, N. Elmqvist, uxSense: Supporting user experience analysis with visualization and computer vision, IEEE Trans. Vis. Comput. Graphics (2023) 1–15, http://dx.doi.org/10.1109/TVCG.2023.3241581.

[48] Y. Wu, Y. Xu, S. Gao, X. Wang, W. Song, Z. Nie, X. Fan, Q. Li, LiveRetro: Visual analytics for strategic retrospect in livestream E-commerce, IEEE Trans. Vis. Comput. Graph. 30 (1) (2024) 1117–1127, http://dx.doi.org/10.1109/TVCG.2023.3326911.

[49] Y. Feng, J. Chen, K. Huang, J.K. Wong, H. Ye, W. Zhang, R. Zhu, X. Luo, W. Chen, iPoet: interactive painting poetry creation with visual multimodal analysis, J. Vis. 25 (3) (2022) 671–685, http://dx.doi.org/10.1007/S12650-021-00780-0.

[50] H. Zeng, X. Wang, Y. Wang, A. Wu, T. Pong, H. Qu, GestureLens: Visual analysis of gestures in presentation videos, IEEE Trans. Vis. Comput. Graph. 29 (8) (2023) 3685–3697, http://dx.doi.org/10.1109/TVCG.2022.3169175.

[51] A. Wu, H. Qu, Multimodal analysis of video collections: Visual exploration of presentation techniques in TED talks, IEEE Trans. Vis. Comput. Graph. 26 (7) (2020) 2429–2442, http://dx.doi.org/10.1109/TVCG.2018.2889081.

[52] Z. Chen, Q. Yang, J. Shan, T. Lin, J. Beyer, H. Xia, H. Pfister, iBall: Augmenting basketball videos with gaze-moderated embedded visualizations, in: Proceedings of the CHI Conference on Human Factors in Computing Systems, 2023, pp. 841:1–841:18, http://dx.doi.org/10.1145/3544548.3581266.

[53] C. Chen, J. Wu, X. Wang, S. Xiang, S. Zhang, Q. Tang, S. Liu, Towards better caption supervision for object detection, IEEE Trans. Vis. Comput. Graph. 28 (4) (2022) 1941–1954, http://dx.doi.org/10.1109/TVCG.2021.3138933.

[54] T. Blascheck, F. Beck, S. Baltes, T. Ertl, D. Weiskopf, Visual analysis and coding of data-rich user behavior, in: Proceedings of IEEE Conference on Visual Analytics Science and Technology, 2016, pp. 141–150, http://dx.doi.org/10.1109/VAST.2016.7883520.

[55] Y. Fu, J.T. Stasko, Supporting data-driven basketball journalism through interactive visualization, in: S.D.J. Barbosa, C. Lampe, C. Appert, D.A. Shamma, S.M. Drucker, J.R. Williamson, K. Yatani (Eds.), CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans, la, USA, 29 April 2022 - 5 May 2022, ACM, 2022, pp. 598:1–598:17, http://dx.doi.org/10.1145/3491102.3502078.

[56] K. Yang, Z. Chen, Y. Cai, D. Huang, H. Leung, Improved automatic keyword extraction given more semantic knowledge, in: Proceedigns of International Workshops of Database Systems for Advanced Applications, Vol. 9645, 2016, pp. 112–125, http://dx.doi.org/10.1007/978-3-319-32055-7_10.

[57] C. Zhang, Automatic keyword extraction from documents using conditional random fields, J. Comput. Inf. Syst. 4 (3) (2008) 1169–1180.

[58] B. Zhu, Y. Cai, H. Ren, Graph neural topic model with commonsense knowledge, Inf. Process. Manag. 60 (2) (2023) 103215, http://dx.doi.org/10.1016/J.IPM.2022.103215.

[59] H. Zhang, Y. Cai, H. Ren, Q. Li, Multimodal topic modeling by exploring characteristics of short text social media, IEEE Trans. Multim. 25 (2023) 2430–2445, http://dx.doi.org/10.1109/TMM.2022.3147064.

[60] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, L. Zhao, Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey, Multimedia Tools Appl. 78 (2019) 15169–15211.

[61] C. Sung, X. Huang, Y. Shen, F. Cherng, W. Lin, H. Wang, Exploring online learners' interactive dynamics by visually analyzing their time-anchored comments, Comput. Graph. Forum 36 (7) (2017) 145–155, http://dx.doi.org/10.1111/CGF.13280.

[62] K.S. Jones, A statistical interpretation of term specificity and its application in retrieval, J. Doc. 60 (5) (2004) 493–502, http://dx.doi.org/10.1108/00220410410560573.

[63] Y. Zeng, G. Wang, H. Ren, Y. Cai, H. Leung, Q. Li, Q. Huang, A knowledge-enhanced and topic-guided domain adaptation model for aspect-based sentiment analysis, IEEE Trans. Affect. Comput. 15 (2) (2024) 709–721, http://dx.doi.org/10.1109/TAFFC.2023.3292213.

[64] Y. Cai, Q. Huang, Z. Lin, J. Xu, Z. Chen, Q. Li, Recurrent neural network with pooling operation and attention mechanism for sentiment analysis: A multi-task learning approach, Knowl.-Based Syst. 203 (2020) 105856, http://dx.doi.org/10.1016/J.KNOSYS.2020.105856.

[65] X. Zhang, J. Xu, Y. Cai, X. Tan, C. Zhu, Detecting dependency-related sentiment features for aspect-level sentiment classification, IEEE Trans. Affect. Comput. 14 (1) (2023) 196–210, http://dx.doi.org/10.1109/TAFFC.2021.3063259.

[66] E.J. Soure, E. Kuang, M. Fan, J. Zhao, CoUX: Collaborative visual analysis of think-aloud usability test videos for digital interfaces, IEEE Trans. Vis. Comput. Graph. 28 (1) (2022) 643–653, http://dx.doi.org/10.1109/TVCG.2021.3114822.

[67] X. Wang, Y. Ming, T. Wu, H. Zeng, Y. Wang, H. Qu, DeHumor: Visual analytics for decomposing humor, IEEE Trans. Vis. Comput. Graph. 28 (12) (2022) 4609–4623, http://dx.doi.org/10.1109/TVCG.2021.3097709.

[68] L. Shi, J. Hu, Z. Tan, J. Tao, J. Ding, Y. Jin, Y. Wu, P.M. Thompson, MVNet: Multi-variate multi-view brain network comparison over uncertain data, IEEE Trans. Vis. Comput. Graph. 28 (12) (2022) 4640–4657, http://dx.doi.org/10.1109/TVCG.2021.3098123.

[69] E.E. Firat, B. Swallow, R.S. Laramee, PCP-Ed: Parallel coordinate plots for ensemble data, Vis. Inform. 7 (1) (2023) 56–65, http://dx.doi.org/10.1016/J.VISINF.2022.10.003.

[70] P. Klemm, S. Oeltze-Jafra, K. Lawonn, K. Hegenscheid, H. Völzke, B. Preim, Interactive visual analysis of image-centric cohort study data, IEEE Trans. Vis. Comput. Graph. 20 (12) (2014) 1673–1682, http://dx.doi.org/10.1109/TVCG.2014.2346591.

[71] C. Shi, S. Fu, Q. Chen, H. Qu, VisMOOC: Visualizing video clickstream data from massive open online courses, in: S. Liu, G. Scheuermann, S. Takahashi (Eds.), 2015 IEEE Pacific Visualization Symposium, PacificVis 2015, Hangzhou, China, April 14-17, 2015, IEEE Computer Society, 2015, pp. 159–166, http://dx.doi.org/10.1109/PACIFICVIS.2015.7156373.

[72] Q. Chen, Y. Chen, D. Liu, C. Shi, Y. Wu, H. Qu, PeakVizor: Visual analytics of peaks in video clickstreams from massive open online courses, IEEE Trans. Vis. Comput. Graph. 22 (10) (2016) 2315–2330, http://dx.doi.org/10.1109/TVCG.2015.2505305.

[73] H. Li, M. Xu, Y. Wang, H. Wei, H. Qu, A visual analytics approach to facilitate the proctoring of online exams, in: Proceedings of the CHI Conference on Human Factors in Computing Systems, 2021, pp. 682:1–682:17, http://dx.doi.org/10.1145/3411764.3445294.

[74] S. Chen, S. Li, Y. Li, J. Zhu, J. Long, S. Chen, J. Zhang, X. Yuan, DanmuVis: Visualizing danmu content dynamics and associated viewer behaviors in online videos, Comput. Graph. Forum 41 (3) (2022) 429–440, http://dx.doi.org/10.1111/CGF.14552.

[75] A. Preston, K. Ma, Communicating uncertainty and risk in air quality maps, IEEE Trans. Vis. Comput. Graph. 29 (9) (2023) 3746–3757, http://dx.doi.org/10.1109/TVCG.2022.3171443.

[76] H. Han, C. Lian, Z. Zeng, B. Xu, J. Zang, C. Xue, Multimodal multi-instance learning for long-term ECG classification, Knowl.-Based Syst. 270 (2023) 110555, http://dx.doi.org/10.1016/J.KNOSYS.2023.110555.

[77] N. Cao, C. Shi, W.S. Lin, J. Lu, Y. Lin, C. Lin, TargetVue: Visual analysis of anomalous user behaviors in online communication systems, IEEE Trans. Vis. Comput. Graph. 22 (1) (2016) 280–289, http://dx.doi.org/10.1109/TVCG.2015.2467196.

[78] W. Zhang, J.K. Wong, X. Wang, Y. Gong, R. Zhu, K. Liu, Z. Yan, S. Tan, H. Qu, S. Chen, W. Chen, CohortVA: A visual analytic system for interactive exploration of cohorts based on historical data, IEEE Trans. Vis. Comput. Graph. 29 (1) (2023) 756–766, http://dx.doi.org/10.1109/TVCG.2022.3209483.

[79] B. Shneiderman, The eyes have it: a task by data type taxonomy for information visualizations, in: Proceedings 1996 IEEE Symposium on Visual Languages, 1996, pp. 336–343, http://dx.doi.org/10.1109/VL.1996.545307.

[80] D. Lange, R. Judson-Torres, T.A. Zangle, A. Lex, Aardvark: Composite visualizations of trees, time-series, and images, IEEE Trans. Vis. Comput. Graphics 31 (1) (2025) 1290–1300, http://dx.doi.org/10.1109/TVCG.2024.3456193.

[81] Y. Guo, H. Shao, C. Liu, K. Xu, X. Yuan, PrompTHis: Visualizing the process and influence of prompt editing during text-to-image creation, IEEE Trans. Vis. Comput. Graphics (2024).

[82] K. Kucher, A. Kerren, Text visualization techniques: Taxonomy, visual survey, and community insights, in: IEEE Pacific Visualization Symposium, 2015, pp. 117–121.

[83] A. Šilić, B.D. Bašić, Visualization of text streams: A survey, in: International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, Springer, 2010, pp. 31–43.

[84] S. Afzal, S. Ghani, M.M. Hittawe, S.F. Rashid, O.M. Knio, M. Hadwiger, I. Hoteit, Visualization and visual analytics approaches for image and video datasets: A survey, ACM Trans. Interact. Intell. Syst. 13 (1) (2023) 1–41.

[85] Y. Fang, H. Xu, J. Jiang, A survey of time series data visualization research, in: IOP Conference Series: Materials Science and Engineering, Vol. 782, No. 2, 2020, 022013.

[86] Y. Guo, S. Guo, Z. Jin, S. Kaul, D. Gotz, N. Cao, Survey on visual analysis of event sequence data, IEEE Trans. Vis. Comput. Graphics 28 (12) (2021) 5091–5112.

[87] S. Lee, S.-H. Kim, B.C. Kwon, VLAT: Development of a visualization literacy assessment test, IEEE Trans. Vis. Comput. Graphics 23 (1) (2017) 551–560, http://dx.doi.org/10.1109/TVCG.2016.2598920.

[88] R. Borgo, J. Kehrer, D.H. Chung, E. Maguire, R.S. Laramee, H. Hauser, M. Ward, M. Chen, Glyph-based visualization: Foundations, design guidelines, techniques and applications, in: Eurographics State of the Art Reports, 2013, pp. 39–63.

[89] H. Chernoff, The use of faces to represent points in k-dimensional space graphically, J. Amer. Statist. Assoc. 68 (342) (1973) 361–368.

[90] X. Xie, X. Cai, J. Zhou, N. Cao, Y. Wu, A semantic-based method for visualizing large image collections, IEEE Trans. Vis. Comput. Graphics 25 (7) (2018) 2362–2377.

[91] J. Zahálka, M. Worring, J.J. Van Wijk, II-20: Intelligent and pragmatic analytic categorization of image collections, IEEE Trans. Vis. Comput. Graphics 27 (2) (2021) 422–431, http://dx.doi.org/10.1109/TVCG.2020.3030383.

[92] J. Zhao, C.A. Bhatt, M. Cooper, D.A. Shamma, Flexible learning with semantic visual exploration and sequence-based recommendation of MOOC videos, in: R.L. Mandryk, M. Hancock, M. Perry, A.L. Cox (Eds.), Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018, ACM, 2018, p. 329, http://dx.doi.org/10.1145/3173574.3173903.

[93] Y. Sun, S. Wang, S. Feng, S. Ding, C. Pang, J. Shang, J. Liu, X. Chen, Y. Zhao, Y. Lu, et al., Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation, 2021, arXiv preprint arXiv:2107.02137.

[94] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017).

[95] J. Feng, K. Wu, S. Chen, TopicBubbler: An interactive visual analytics system for cross-level fine-grained exploration of social media data, Vis. Inform.s 7 (4) (2023) 41–56, http://dx.doi.org/10.1016/J.VISINF.2023.08.002.

[96] A. Dong, W. Zeng, X. Chen, Z. Cheng, VIStory: Interactive storyboard for exploring visual information in scientific publications, in: Proceedings of the 12th International Symposium on Visual Information Communication and Interaction, 2019, http://dx.doi.org/10.1145/3356422.3356430.

[97] K. Kurzhals, M. John, F. Heimerl, P. Kuznecov, D. Weiskopf, Visual movie analytics, IEEE Trans. Multimed. 18 (11) (2016) 2149–2160.

[98] J. Kim, S. Lee, H. Jeon, K.-J. Lee, H.-J. Bae, B. Kim, J. Seo, PhenoFlow: A human-LLM driven visual analytics system for exploring large and complex stroke datasets, IEEE Trans. Vis. Comput. Graphics (2024).

[99] Y. Ye, S. Xiao, X. Zeng, W. Zeng, ModalChorus: Visual probing and alignment of multi-modal embeddings via modal fusion map, IEEE Trans. Vis. Comput. Graphics (2024).

[100] M. Nylin, J. Lundberg, M. Bång, K. Kucher, Glyph design for communication initiation in real-time human-automation collaboration, Vis. Inform. 8 (1) (2024) 23–35, http://dx.doi.org/10.1016/J.VISINF.2024.09.006.

[101] L. Ying, X. Shu, D. Deng, Y. Yang, T. Tang, L. Yu, Y. Wu, MetaGlyph: Automatic generation of metaphoric glyph-based visualization, IEEE Trans. Vis. Comput. Graph. 29 (1) (2023) 331–341, http://dx.doi.org/10.1109/TVCG.2022.3209447.

[102] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C.D. Hansen, J.C. Roberts, Visual comparison for information visualization, Inf. Vis. 10 (4) (2011) 289–309, http://dx.doi.org/10.1177/1473871611416549.

[103] B. He, H. Li, Y.K. Jang, M. Jia, X. Cao, A. Shah, A. Shrivastava, S.-N. Lim, MA-LMM: Memory-augmented large multimodal model for long-term video understanding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2024, pp. 13504–13514.

[104] S. Ren, L. Yao, S. Li, X. Sun, L. Hou, TimeChat: A time-sensitive multimodal large language model for long video understanding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2024, pp. 14313–14323.

[105] K. Ma, X. Zang, Z. Feng, H. Fang, C. Ban, Y. Wei, Z. He, Y. Li, H. Sun, LLaViLo: Boosting video moment retrieval via adapter-based multimodal modeling, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, 2023, pp. 2798–2803.

[106] T.-J. Fu, L. Yu, N. Zhang, C.-Y. Fu, J.-C. Su, W.Y. Wang, S. Bell, Tell me what happened: Unifying text-guided video completion via multimodal masked video generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 10681–10692.

[107] Q. Yu, Q. Sun, X. Zhang, Y. Cui, F. Zhang, Y. Cao, X. Wang, J. Liu, CapsFusion: Rethinking image-text data at scale, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2024, pp. 14022–14032.

[108] P.K.A. Vasu, H. Pouransari, F. Faghri, R. Vemulapalli, O. Tuzel, MobileCLIP: Fast image-text models through multi-modal reinforced training, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2024, pp. 15963–15974.

[109] D. Feng, X. He, Y. Peng, MKVSE: Multimodal knowledge enhanced visual-semantic embedding for image-text retrieval, ACM Trans. Multimed. Comput. Commun. Appl. 19 (5) (2023) http://dx.doi.org/10.1145/3580501.

[110] R. Ashmore, R. Calinescu, C. Paterson, Assuring the machine learning lifecycle: Desiderata, methods, and challenges, ACM Comput. Surv. 54 (5) (2021) http://dx.doi.org/10.1145/3453444.

[111] S.Y.-T. Lee, A. Bahukhandi, D. Liu, K.-L. Ma, Towards dataset-scale and feature-oriented evaluation of text summarization in large language model prompts, IEEE Trans. Vis. Comput. Graphics 31 (1) (2025) 481–491, http://dx.doi.org/10.1109/TVCG.2024.3456398.

[112] A. Coscia, A. Endert, KnowledgeVIS: Interpreting language models by comparing fill-in-the-blank prompts, IEEE Trans. Vis. Comput. Graph. 30 (9) (2024) 6520–6532, http://dx.doi.org/10.1109/TVCG.2023.3346713.

[113] H. Strobelt, A. Webson, V. Sanh, B. Hoover, J. Beyer, H. Pfister, A.M. Rush, Interactive and visual prompt engineering for Ad-hoc task adaptation with large language models, IEEE Trans. Vis. Comput. Graph. 29 (1) (2023) 1146–1156, http://dx.doi.org/10.1109/TVCG.2022.3209479.

[114] L.Y.-H. Lo, H. Qu, How good (or bad) are LLMs at detecting misleading visualizations? IEEE Trans. Vis. Comput. Graphics 31 (1) (2025) 1116–1125, http://dx.doi.org/10.1109/TVCG.2024.3456333.

[115] X. Zeng, H. Lin, Y. Ye, W. Zeng, Advancing multimodal large language models in chart question answering with visualization-referenced instruction tuning, IEEE Trans. Vis. Comput. Graphics 31 (1) (2025) 525–535, http://dx.doi.org/10.1109/TVCG.2024.3456159.

[116] A. Bendeck, J. Stasko, An empirical evaluation of the GPT-4 multimodal language model on visualization literacy tasks, IEEE Trans. Vis. Comput. Graphics 31 (1) (2025) 1105–1115, http://dx.doi.org/10.1109/TVCG.2024.3456155.

[117] J. He, X. Wang, S. Liu, G. Wu, C. Silva, H. Qu, POEM: Interactive prompt optimization for enhancing multimodal reasoning of large language models, 2024, arXiv preprint arXiv:2406.03843.

[118] P. Soni, C. de Runz, F. Bouali, G. Venturini, A survey on automatic dashboard recommendation systems, Vis. Inform. 8 (1) (2024) 67–79, http://dx.doi.org/10.1016/J.VISINF.2024.01.002.

[119] X. Wang, Z. Wu, W. Huang, Y. Wei, Z. Huang, M. Xu, W. Chen, VIS+ AI: integrating visualization with artificial intelligence for efficient data analysis, Front. Comput. Sci. 17 (6) (2023) 176709.

[120] Y. Ye, J. Hao, Y. Hou, Z. Wang, S. Xiao, Y. Luo, W. Zeng, Generative AI for visualization: State of the art and future directions, Vis. Inform. 8 (1) (2024) 43–66, http://dx.doi.org/10.1016/J.VISINF.2024.04.003.

[121] L. Weng, X. Wang, J. Lu, Y. Feng, Y. Liu, W. Chen, InsightLens: Discovering and exploring insights from conversational contexts in large-language-model-powered data analysis, 2024, arXiv preprint arXiv:2404.01644.

[122] A. Wu, Y. Wang, X. Shu, D. Moritz, W. Cui, H. Zhang, D. Zhang, H. Qu, AI4VIS: Survey on artificial intelligence approaches for data visualization, IEEE Trans. Vis. Comput. Graphics 28 (12) (2021) 5049–5070.

[123] J. Wang, X. Li, C. Li, D. Peng, A.Z. Wang, Y. Gu, X. Lai, H. Zhang, X. Xu, X. Dong, Z. Lin, J. Zhou, X. Liu, W. Chen, AVA: An automated and AI-driven intelligent visual analytics framework, Vis. Inform. 8 (1) (2024) 106–114, http://dx.doi.org/10.1016/J.VISINF.2024.06.002.

[124] C. Stoiber, D. Moitzi, H. Stitz, F. Grassinger, A.S.G. Prakash, D. Girardi, M. Streit, W. Aigner, VisAhoi: Towards a library to generate and integrate visualization onboarding using high-level visualization grammars, Vis. Inform. 8 (1) (2024) 1–17, http://dx.doi.org/10.1016/J.VISINF.2024.06.001.

[125] V. Dibia, Ç. Demiralp, Data2vis: Automatic generation of data visualizations using sequence-to-sequence recurrent neural networks, IEEE Comput. Graph. Appl. 39 (5) (2019) 33–46.

[126] D. Deng, A. Wu, H. Qu, Y. Wu, DashBot: Insight-driven dashboard generation based on deep reinforcement learning, IEEE Trans. Vis. Comput. Graphics 29 (1) (2023) 690–700, http://dx.doi.org/10.1109/TVCG.2022.3209468.

[127] Y. Lin, H. Li, A. Wu, Y. Wang, H. Qu, DMiner: Dashboard design mining and recommendation, IEEE Trans. Vis. Comput. Graphics (2023) 1–15, http://dx.doi.org/10.1109/TVCG.2023.3251344.

[128] A. Wu, L. Xie, B. Lee, Y. Wang, W. Cui, H. Qu, Learning to automate chart layout configurations using crowdsourced paired comparison, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, http://dx.doi.org/10.1145/3411764.3445179.

[129] D. Weng, H. Zhu, J. Bao, Y. Zheng, Y. Wu, HomeFinder revisited: Finding ideal homes with reachability-centric multi-criteria decision making, 2018, pp. 1–12, http://dx.doi.org/10.1145/3173574.3173821.

[130] Z. Deng, D. Weng, J. Chen, R. Liu, Z. Wang, J. Bao, Y. Zheng, Y. Wu, AirVis: Visual analytics of air pollution propagation, IEEE Trans. Vis. Comput. Graphics 26 (1) (2020) 800–810, http://dx.doi.org/10.1109/TVCG.2019.2934670.