Leveraging Foundation Models for Crafting Narrative Visualization: A Survey

Yi He*, Ke Xu*, Shixiong Cao, Yang Shi, Qing Chen, and Nan Cao

Abstract—Narrative visualization transforms data into engaging stories, making complex information accessible to a broad audience. Foundation models, with their advanced capabilities such as natural language processing, content generation, and multimodal integration, hold substantial potential for enriching narrative visualization. Recently, a collection of techniques have been introduced for crafting narrative visualizations based on foundation models from different aspects. We build our survey upon 66 papers to study how foundation models can progressively engage in this process and then propose a reference model categorizing the reviewed literature into four essential phases: Analysis, Narration, Visualization, and Interaction. Furthermore, we identify eight specific tasks (e.g. Insight Extraction and Authoring) where foundation models are applied across these stages to facilitate the creation of visual narratives. Detailed descriptions, related literature, and reflections are presented for each task. To make it a more impactful and informative experience for diverse readers, we discuss key research problems and provide the strengths and weaknesses in each task to guide people in identifying and seizing opportunities while navigating challenges in this field.

Index Terms—Narrative Visualization; Automatic Visualization; Foundation Models; Generative AI; Multi-modal; Survey

1 Introduction

Narrative visualization, a powerful approach to transforming data into compelling visual data stories [1], is designed to be accessible and engaging for a wide audience. It is an effective combination of data narrative and information visualization, enabling data-driven stories that convey context, causality, and data insights. Crafting a narrative visualization requires diverse creators' skills, including expertise in data analysis, design, storytelling, and data visualization [2], [3]. These skills surpassed the capability of most existing automatic generation techniques.

At the same time, foundation models such as Large Language Models (LLMs), image generation models, and multimodal foundation models [4] demonstrate remarkable capabilities [5] that can be harnessed for generating visual narratives [6]. Unlike traditional AI and machine learning approaches, foundation models acquire extensive knowledge through pretraining, enabling rapid adaptation to novel tasks. Moreover, foundation models can process natural language inputs, lowering the threshold for users to perform complex tasks like data processing, design, or programming. Recently, many techniques have been introduced for crafting narrative visualizations based on foundation models from different aspects. However, a comprehensive review paper that systematically summarizes these techniques and highlights potential research directions is still missing. Therefore, we believe a dedicated review is desired to track these developments and analyze their implications for the crafting of narrative visualization.

This paper summarizes recent advances in using foundation models to craft a narrative visualization. A standard crafting reference model is summarized based on the literature review (Fig. 1), which consists of four major steps : ***Analysis***, i.e., using foundation models to analyze the raw data to chose the relevant data content [7]–[10] or extract data insights [3], [11]–[15]; ***Narration***, i.e., using foundation models to construct the narrative logic and structure [16]–[20] or generate narrative content such as title, caption, annotation, and script [21]–[26]; ***Visualization***, i.e., using foundation models to generate charts with image embellishments and animations to facilitate information communication [6], [11], [20], [27]–[35]; ***Interaction***, i.e., using foundation models to support and implement complex user interactions such as nature language query and data exploration, which makes the narrative visualization more comprehensible, accessible, and controllable to its readers [36], [36]–[42]. Specifically, the contributions of this paper are summarized as follows:

- We categorize existing literature on using foundation models for narrative visualization and introduce a reference model consisting of four stages and eight tasks, exploring how foundation models are used in supporting the generation of narrative visualization.
- We made an in-depth discussion on the existing research limitations and future research opportunities to help inspire future research.

The rest of the survey is structured as follows: In Section 2, we introduce the related work and the methodology for literature collection. In Section 3, we describe how to construct the reference model and provide a comprehensive definition of it. In Sections 4, 5, 6, and 7, we present each stage along with the associated problems. Following that, in Section 8, we discuss the methods of evaluation. In Section 9, we delve into the discussion of research challenges and opportunities. Finally, we bring our survey to a conclusion in Section 10. In addition to the state-of-the-art survey, we developed an interactive browser to facilitate the exploration and presentation of the collected papers at **http://lm4vis.idvxlab.com/**.

- *Yi He, Shixiong Cao, Yang Shi, Qing Chen, and Nan Cao are with Intelligent Big Data Visualization Lab, Tongji University. Email: {heyi_11,caoshixiong,yangshi.idvx,qingchen,nan.cao}@tongji.edu.cn. Nan Cao is the corresponding author.*
- *Ke Xu is with the School of Intelligence Science and Technology at Nanjing University. E-mail: lukexuke@gmail.com.*
- ** These authors contributed equally to this work.*

2 Related Survey and Methodology

In this section, we first review the relevant surveys to highlight the value of our work. After that, we introduce the scope and methodology of our survey.

2.1 Related Surveys

For a long time, the process of creating visualization diagrams has been time-consuming and labor-intensive, requiring people to have skills in data analysis, design, and programming. With the rapid development of new techniques, artificial intelligence have been used to help with the generation of visualization diagrams. These techniques have been well summarized in several relevant survey papers. For example, Wang et al. [43] surveyed 88 ML4VIS papers, identifying seven main visualization processes where ML techniques can be applied, and aligning these with existing theoretical models in an ML4VIS pipeline. Wu et al. [44] provided an overview of AI techniques applied to generate information visualization diagrams (AI4VIS). Di et al. [45] explores the challenges and opportunities of integrating generative models into visualization workflows. Yang et al. [46] also broadly explored the integration of visualization and foundation models by respectively exploring visualization for foundation models (VIS4FM) and foundation models for visualization (FM4VIS). All these survey papers mainly focused on information visualization techniques, but none of them focused on visual narrative. Recently, Chen et al. [47] made a comprehensive survey on visualization tools to explore how automation engages in visualization design and narrative processes. They summarized six narrative visualization genres and categorized them by intelligence and automation levels. This work is the most relevant to our survey, however, the authors only focused on tools instead of foundation models and detailed techniques.

Compared to the above survey papers, we would like to explore how foundation models are used for generating narrative visualizations from different aspects, including analysis, narration, visualization, and interaction. Especially, a foundation model is any model trained on broad data, typically using self-supervision at scale, which can be adapted (e.g., fine-tuned) to a wide range of downstream tasks. Current examples include BERT, GPT, and CLIP [4], [48]. In this paper, we focus on models with over 100 million parameters.

2.2 Literature Selection

We collect papers from a set of relevant journals and conference proceedings using both the **search-driven** and **reference-driven** methods. In particular, for the search-driven selection, we conducted four rounds of paper selection:

TABLE 1: Relevant Venues

VIS	VIS(infoVIS VAST), PacificVIS TVCG, EuroVIS(Computer graphics forum)
HCI	CHI, TOCHI, UIST, IUI, CSCW, IJHCS
AI	AAAI, IJCAI, TPAMI, AI ICML, NeurIPS, ICLR, JMLR(Machine Learning) CVPR, IJCV, ICCV(Computer Vision) ACL, EMNLP(Natural Language Processing) TIST, TiiS

Identification: we select relevant papers from the top venues in the field of artificial intelligence, data visualization, and human-computer interaction as summarized in Table 1. We specifically select papers containing keywords related to foundation models (e.g., 'generative AI,' 'large language models,' 'pre-trained models,' 'foundation models,' 'image-generation models,' 'GPT,' 'AIGC') and narrative visualization (e.g., 'visualization,' 'infographic,' 'data videos,' 'narrative,' 'data comics,' 'annotated chart,' 'slideshow,' 'data story'). As a result, 2417 papers were selected.

Code Screening: we used a program for preliminary screening and compiled a keyword list that included terms related to narrative visualization and foundation models. We matched these keywords against the titles and abstracts of the initially selected papers, including only those that encompassed both aspects in our selection. As a result, only 361 papers are retained. **Manual Screening:** we meticulously reviewed the abstracts of the remaining articles to ensure their alignment with our research scope. This thorough assessment allowed us to refine our selection, resulting in 121 articles being included in our research corpus.

Eligibility Assessment: finally, we checked the eligibility of each paper based on the following exclusion criteria : (1) articles using foundation models with fewer than 100 million parameters were excluded; (2) articles that are irrelevant to the process of crafting narrative visualizations were excluded. As a result, 121 papers were filtered out and we conducted a detailed full-text review of these papers. This meticulous review process involved carefully examining each article to ensure it met all inclusion criteria. We paid particular attention to the methodologies, results, and discussions presented in each article to confirm their alignment with the scope of our study. Finally, only 55 of the most relevant papers were selected in our review corpus.

In the reference-driven selection, we went through all the papers cited in the above 55 papers. Some important papers (i.e., highly relevant and highly cited) from journals or conference proceedings outside our searching scope were also included in our review corpus. Finally, we have compiled a total of 66 most relevant papers. The papers in our corpus are meticulously cataloged in Fig. 2.

2.3 Literature Review

The paper analysis was conducted in three phases. In the first phase, we extracted a concise description for each paper, including the foundation model mentioned (e.g., GPT, BERT, etc.), the type of foundation model, the scale of training parameters used, the specific visualization problem addressed by the model, the context or scenario where the model was applied, and a summary of the paper's main theme or contribution. In the second phase, four authors summarized and merged similar items of the encoding above, enabling us to identify common patterns and determine categories. During weekly discussions, we iteratively adapted, split, and refined the codes multiple times until no further disagreements arose. We then classified the literature into four stages and eight main tasks., with eachtask further divided into detailed subtasks specifying the specific activities involved. In the third phase, each author was responsible for classifying papers within a specific stage until ultimately completing the paper selection.

3 Reference Model

To understand how foundation models support the generation of narrative visualizations, it is essential to analyze the creation process and identify the key techniques at each generation stage where foundation models can contribute effectively.

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2025.3542504

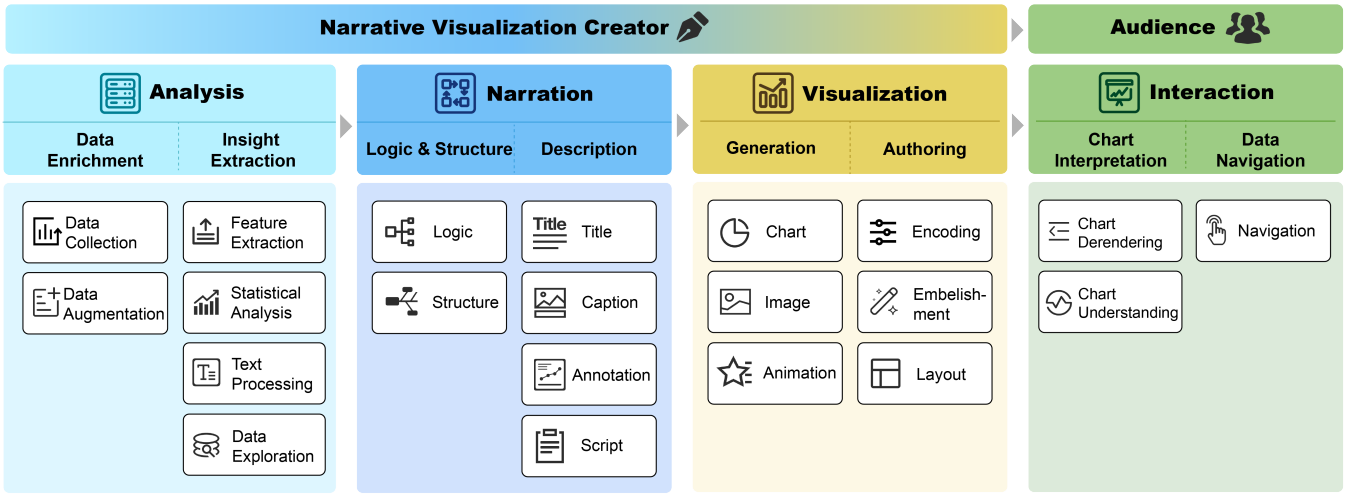


Fig. 1: A Reference Model for Creating Narrative Visualizations Based on Foundation Models

Several researchers have attempted to outline the process of creating narrative visualization. For instance, Lee et al. [49] identified three primary phases in crafting data stories: exploring data, making a story, and telling a story. Segel et al. summarized three key elements of narrative visualization: narrative genres, narrative structure, and visual narrative [50]. Li et al. [51] also investigated a framework encompassing analysis, planning, implementation, and communication. Amini et al. [52], through observations of how people create data videos, proposed a process involving data reading and interpretation, data selection, narrative crafting, and viewer engagement. However, these summaries still have some limitations. Specifically, they tend to generalize the process of story design and lack detailed guidance on specific design actions that designers can undertake during the creation process.

To provide a more comprehensive understanding of how foundation models contribute to the creation of narrative visualization, we have synthesized and refined previous research findings. As illustrated in Fig. 1, we introduce a reference model for creating narrative visualizations based on foundation models. The reference model consists of four major stages, including analysis, narration, visualization, and interactions. In particular, the first three stages focus on the tasks that an analyzer needs to perform when creating a narrative visualization. The last stage focuses on how an audience interacts with an existing visualization to gain more insights.

Analysis: the analysis stage deals with the input data and extracts meaningful data insight for narration regarding the story creators' requirements [43]. Here, foundation models are often used to help with data processing and uncover data insights. In particular, they are frequently used in two major tasks: (1) data enrichment and (2) insight extraction. ***Data enrichment*** is to augment the raw data to improve its quality and usability. Here, foundation models are used either to find relevant datasets [7] or fix data errors such as eliminating outliers or imputing missing data [10]. ***Insight extraction*** is to use foundation models to automatically analyze the raw data to extract useful knowledge regarding a narration topic [53].

Narration: the narration stage involves arranging a series of data insights logically and structurally to create a coherent and engaging story. Here, foundation models are often used to identify logical relationships between data insights and organize them. In particular, they are frequently used in two major tasks: (1) narrative logic and structure construction and (2) narrative content generation. ***Narrative logic and structure construction*** is to deliberately guide the audience's attention in a specific sequence, helping them maintain a sense of direction throughout the story. Here, foundation models can identify relationships between data facts and adjust the narrative order [54]. ***Narrative content generation*** is to use foundation models to generate summarizing or descriptive text [55].

Visualization: the visualization stage transforms the prepared data insights into visual representations for the audience. Here, foundation models are often used to help users automatically generate accurate and visually appealing narrative visualizations. In particular, foundation models are primarily used to accomplish two main tasks: (1) visualization generation and (2) visual content embellishment. ***Visualization generation*** is to interpret data insights and select suitable methods for visual representation. Here, foundation models are often used to recommend the most suitable chart types and generate the necessary code to create charts [6]. ***Visual content embellishment*** is to use foundation models to generate stylized visualization using the text-conditioned image-to-image generation capabilities [6], [56].

Interaction: the interaction stage deals with how users read and interact with the visualization to extract valuable information [43]. Here, foundation models can help the audience understand the narrative visualization more thoroughly and explore further. In particular, foundation models are primarily used to accomplish two main tasks: (1) facilitating natural language interactions and (2) enhancing interactive exploration. ***Facilitating natural language interactions*** is to use foundation models to communicate with the audience using natural language, improving the effectiveness of information communication [42]. ***Enhancing interactive exploration*** is to use foundation models to generate code to interact with visualizations by triggering corresponding events in the underlying JS code, such as switching data mappings, zooming in, and zooming out [7].

In conclusion, foundation models play different roles in various stages to help users complete the process of creating narrative visualizations. These models provide robust support, allowing users to focus on the narrative content and the knowledge they want to convey rather than technical details. In the following sections, 4 through 7, we will provide a detailed overview of the tasks

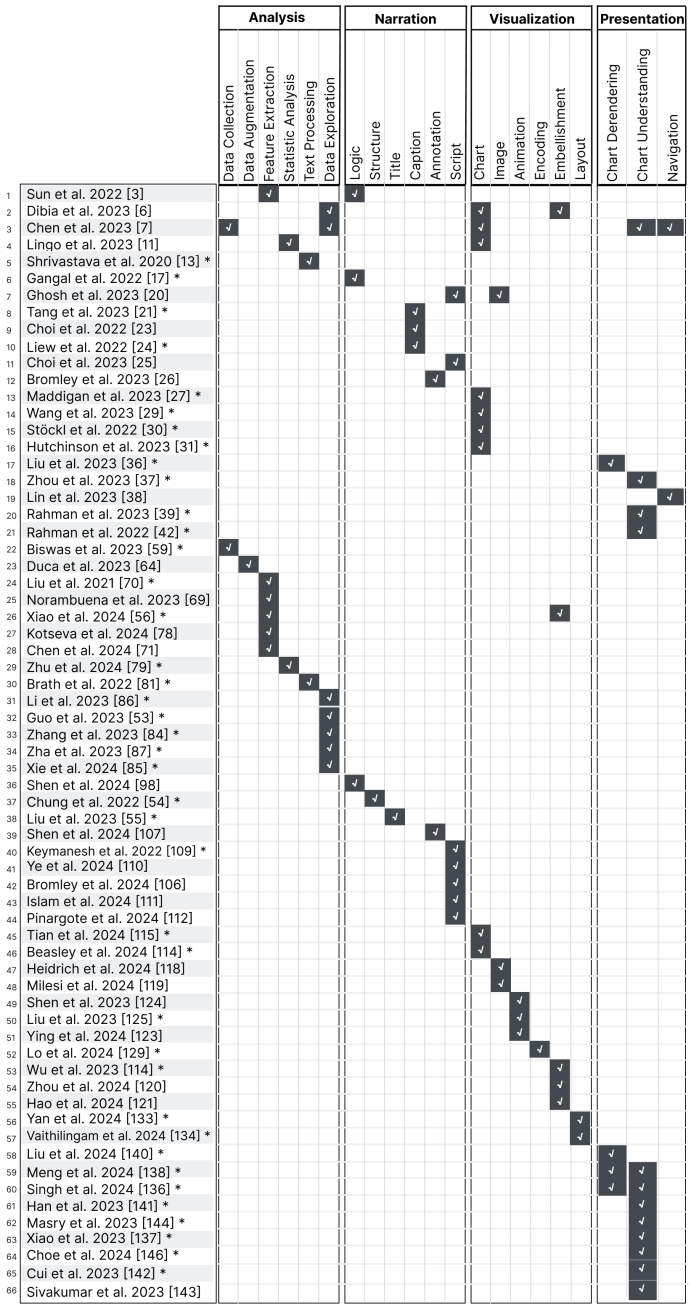


Fig. 2: An overview of the papers related to the reference model and their associated code. Papers marked with (*) represent tools that can be used in creating narrative visualizations, even if the primary focus of the research is not specifically on narrative visualization. A clearer and more detailed collection of the papers can be found at http://lm4vis.idvxlab.com/.

foundation models accomplish. At the end of each subsection, we explore the existing challenges and future research directions.

4 Analysis

To tell a data story or craft a narrative visualization, the first thing an author should handle is to collect and analyze the data. We reviewed 19 papers related to the data analysis stage, which can be broadly classified into two major categories based on the role of foundation models: (1) data enrichment and (2) insight extraction, which will be discussed in this section.

4.1 Data Enrichment

Data is not a natural resource [45]. When crafting a narrative visualization, users usually need to gather datasets from various sources, which is usually time-consuming and labor-intensive, especially for large-scale datasets. In addition, the collected raw data often contain missing values, inconsistencies, or even errors, which can greatly impact the accuracy of subsequent analyses. To address these challenges, recent research has focused on two main aspects: data collection and data augmentation based on foundation models.

Data Collection involves integrating multiple data sources [57]. Foundation models, with their information retrieval capabilities, can align users' narrative goals and requirements with various data sources and effectively identify the most relevant content [58]. For example, Chen et al. [7] demonstrated the utility of GPT-4 in suggesting relevant website links and providing access to downloadable datasets, thus facilitating the collection of diverse and pertinent data. Similarly, Biswas et al. [59] showed that ChatGPT could collect and organize data from social media platforms. Both studies highlight the ability of foundation models to access a wide range of data sources and aggregate relevant information.

Data Augmentation is the process of augmenting the existing datasets to make them more complete and comprehensive [60]–[62]. Here, Foundation models are frequently used to augment existing datasets using Retrieval-Augmented Generation (RAG) technology. RAG is a natural language processing technique that combines information retrieval and generation models. RAG's main idea is to enhance the content generation process by retrieving relevant information from external knowledge bases [63]. Using RAG, foundation models can integrate relevant information according to specific domain requirements for users, thereby providing richer context and background information for narrative visualization. For example, Duca et al. [64] proposed a method using RAG technology to generate context for data-driven stories. First, documents related to specific topics were input into a vector database for indexing to facilitate efficient retrieval. Then, in the retrieval step, relevant information was queried from external knowledge bases. In the generation step, this information was integrated into an LLM, which generated accurate and detailed textual content based on the retrieved knowledge. This approach supplemented data narratives with rich background knowledge and contextual information, achieving the goal of data augmentation.

Reflection. Current research shows that foundation models have substantial potential to automate the laborious process of data collection and augmentation [58]. Data collection involves using foundation models' information retrieval capabilities to identify and aggregate relevant datasets from diverse sources. Data augmentation is enhanced by foundation models through the use of RAG, which integrates external knowledge to enrich and supplement existing datasets.

However, the generated content may contain errors [65]. To prevent foundation models from generating incorrect information, enhancing the model's ability to recognize and acknowledge its knowledge gaps is essential [66]. Therefore, future research should focus either on developing algorithms to help foundation models identify their limitations and avoid generating inaccurate content [67], or integrating real-time human feedback to help refine the outputs generated by foundation models [56], [67]. Research should also explore methods for displaying the basis and context of AI-generated content, making it more trustworthy.

4.2 Insight Extraction

Insight extraction involves performing in-depth analysis to identify crucial insights and patterns in the data [68]. Foundation models, with their advanced analytical capabilities, enable users to go beyond superficial data analysis, allowing for deeper exploration and discovery of hidden value within the data [58]. In particular, recent research in this direction has focused on four main aspects: feature embedding, statistical analysis, text processing, and data exploration.

Feature Embedding converts unstructured data, like text [72], images [73], and structured data, like data facts [3], [71], into semantic feature vectors [46], [74]. Foundation models such as BERT [75] and ELMo [76] transform textual data into concise, low-dimensional vector spaces. In narrative visualization, these feature vectors can be used to identify patterns, trends, and connections within the data, which in turn enhances the storytelling aspect by providing deeper insights and more coherent narratives [56], [69], [70], [77], [78]. For example, Norambuena et al. [69] used text embedding to capture deep semantic information from text, supporting the construction of graph-based narrative visualizations. The paper took news articles as input data, where each article was considered an event. By using the all-MiniLM-L6-v2 model to generate text embeddings, and combining UMAP algorithm for dimensionality reduction and HDBSCAN algorithm for clustering, it calculated the coherence values between events to construct a high-level discrete structure narrative map. This framework supported analysts in incrementally refining the narrative model through semantic interactions, leading to a better understanding of the relationships between news events, as shown in Fig. 3 (1). These vector were used to generate the projection space and compute coherence. Another example is ADVISor [70](Fig. 3 (2)), which automatically generates annotated charts to answer natural language questions about tabular data based on BERT [75]. Here, BERT encodes natural language questions and table headers into latent vectors. These vectors are then used to extract relevant data regions from the table and identify potential data aggregation methods. Based on this information, charts are generated.

Despite the above techniques, recent research also focuses on how to fine-tune foundation models to design effective embedding methods that map data facts appropriately in a vector space that reflects the logical and semantic relationships of data facts. For example, Erato [3] fine-tuned a BERT to convert data facts into vectors with the distances between two vectors indicating logical relationships between data facts. In particular, two fully connected layers were added on top of BERT (Fig. 3 (3)). The input format converted each data fact into a tokenized string, capturing both structural and semantic information. The model was fine-tuned on a dataset of manually designed data stories, where each training sample was a trigram of data facts capturing a logical relationship in the story. Another notable work is Chart2Vec [71]. It proposed a method to create an embedding for narrative visualizations that incorporates context-aware information (Fig. 3 (4)). In this work, authors collected a dataset of multi-view visualizations, including data stories from Calliope [2] and dashboards from Tableau Public, and transformed these visualizations into data facts similar to Erato. The structural information was encoded using a one-hot vector representation of the parse tree derived from a context-free grammar, while the semantic information was encoded using pre-trained Word2Vec embedding for the textual content. These features were then fused, including positional encoding to preserve the connection between structural and semantic data, creating a final vector representation for each chart. The chart embedding results support various downstream tasks such as visualization recommendation and data story generation. Compared to general visualization, feature embedding in narrative visualization focuses more on the relationships between data facts and the contextual semantics. This emphasis helps in crafting a coherent and logical story.

Statistical Analysis provides an overview of a dataset's central tendencies, dispersion, and overall distribution [11]. These statistics offer a preliminary understanding of the data, highlighting key attributes such as the mean, median, range, and standard deviation, and laying the groundwork for more in-depth data insight.

Foundation models excel in generating code to meet users' computational needs. For example, Zhu et al. [79] used pre-designed templates to generate statistical questions, which were refined by GPT-3.5-Turbo. The model preprocessed the tabular data, such as cleaning and standardizing formats, and generated SQL queries or statistical code to perform the actual calculations. The results were then validated and interpreted to ensure accuracy and reliability. Similarly, Lingo et al. [11] demonstrated that by adjusting prompts, ChatGPT could perform summary statistics, data grouping, hypothesis testing, and other operations on tabular data by generating code to provide real-time computational assistance. Both studies validate the capability of foundation models to perform statistical computations on tabular data, lowering the learning curve and barriers compared to traditional complex data analysis methods. However, for narrative visualization, it is essential not only to compute the statistical data but also to highlight the exciting results, such as outliers and change points, which can drive the story. These statistical data can serve as anchor points, drawing the reader's attention and encouraging further exploration.

Text Summarization extracts key information and insights from large volumes of textual data, condensing it to provide a clear summary that helps users quickly grasp the story behind the data [80], [81]. Compared to traditional topic modeling methods in NLP like NewsViews [82], foundation models help users quickly understand the semantics of input text, offering preliminary exploration and visualization suggestions. Textual data, such as user feedback, comments, and subjective evaluations, are important sources for narrative visualization. However, they often present challenges for users in quickly organizing and extracting key insights [81]. However, emerging NLP techniques of foundation models like document summarization addressed these challenges. For example, iSeqL [13] used ELMo to generate context-aware word embedding, allowing users to annotate text data interactively and update model predictions in real-time. Through iterative optimization and visualization tools, users could continuously adjust annotations and model parameters, gaining deeper insights into textual data distribution and model performance, thereby effectively conducting exploratory data analysis.

Data Exploration is to extract knowledge from data. Query-based data exploration is a simple example. This approach involves extracting insights through Question-Answering (QA) interactions [83]. Foundation models have demonstrated remarkable proficiency in interacting with user-generated prompts and questions to facilitate data exploration. These models, equipped with advanced natural language processing and machine learning capabilities, can interpret and respond to a wide range of user inquiries, allowing users to delve into datasets and uncover insights in an intuitive and user-friendly manner [53], [84], [85], improving human-computer

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2025.3542504

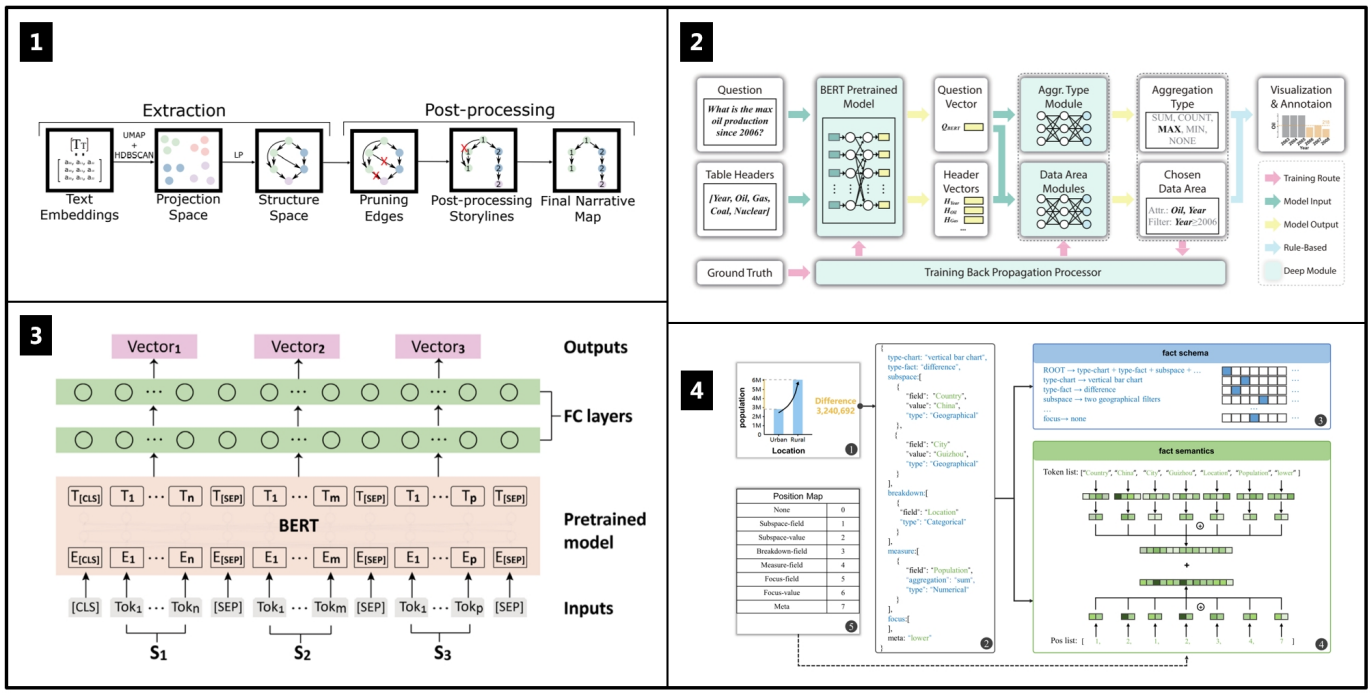


Fig. 3: Selected examples of feature embedding: (1) Use text embedding technology to capture the deep semantic information of text, supporting the construction of graph-based narrative visualizations [69]. (2) ADVISor: Use BERT to convert the natural language in table headers and questions to vectors presenting the semantic meaning [70]. (3) Erato: schematic diagrams of the fact embedding model [3]. (4) Chart2Vec: learn a universal embedding of visualizations with context-aware information [71].

interaction by moving away from complex commands and facilitating a more intuitive and user-friendly mode of interaction [58]. For example, Sheet Copilot [86] could autonomously manage, process, analyze, predict, and visualize data. When a request was received, they transformed raw data into informative results that best aligned with the user's intent. Furthermore, TableGPT [87] leveraged foundation models for comprehensive table understanding by encoding entire tables into global representation vectors, enabling complex data manipulation and analysis. It utilized a Chain-of-Command approach to break down complex tasks into simpler steps, executing them sequentially. Additionally, TableGPT enhanced its performance in specific domains through domain-specific fine-tuning and supports private deployment to ensure data privacy. TableGPT excels by specifically optimizing and fine-tuning LLMs for table-related tasks, enabling comprehensive table understanding and utilizing a chain-of-command approach to break down complex tasks into simple steps. Additionally, TableGPT supports domain-specific fine-tuning and private deployment, ensuring data privacy and security, making it superior to SheetCopilot in handling complex table tasks and sensitive data.

Reflection. Feature embedding with foundation models transforms unstructured and structured data into semantic vectors. Statistical analysis using foundation models generates and validates computations on tabular data. Text summarization through foundation models efficiently condenses and interprets textual data, facilitating exploration and enhancing the integration of qualitative insights into visualizations. Data exploration with foundation models is enhanced by their ability to interpret and respond to user-generated queries, enabling intuitive and interactive knowledge extraction from datasets.

However, most research focuses primarily on the extraction of textual features [69], [78]. As the demand for more complex narrative visualizations grows, particularly in data videos, there is a need for effective integration and alignment of multimodal information, including visuals, audio, and text [47]. Furthermore, studies have shown that integrating multimodal information can help users better understand and remember visualizations [88]. Therefore, exploring the potential of large models in processing multimodal data becomes crucial for enhancing the quality of narrative visualizations. Future research should focus on the integration of multimodal information, exploring the application of foundation models in multimodal data processing and how to achieve semantic alignment across different modalities. This will aid in processing multimodal data, providing more material for narrative visualizations, and designing semantically consistent multimodal narrative visualizations. Furturemore, future research should investigate Chain-of-Thought strategies and systematically designed prompt engineering [16], [89]. These approaches can enhance the interpretability and accuracy of foundation models, while also clearly defining model tasks to prevent aimless exploration and improve the efficiency of narrative visualization creation.

5 Narration

We reviewed 17 papers related to the narration stage in our reference model. These papers are broadly classified into two major categories based on the roles played by foundation models, i.e., (1) logic generation and (2) structure generation, which will be discussed in this section.

5.1 Logic & Structure

Logic and structure are the backbones of narrative visualization [90]. An effective structure with a coherent logic will directly help with the understanding and

memorization [91]. In a data story, editors usually interlink data facts that are logically relevant to craft a storyline following a narrative structure [92]. However, analysis to detect data logic and fit the content into a narrative structure usually introduces technique barriers that are difficult for users with little data analysis knowledge to achieve. Foundation models, especially the pre-trained language models such as GPT [93] and Bert [75], have a great ability to help with logic reasoning in nature languages, which has also been used to build data logic [3], [71] and organize structural data narrations [54], recently.

Logic refers to the relationship between data facts [94]. A disordered narrative sequence can confuse the audience due to the lack of logical dependence between the elements [95]. Based on large amounts of training data, foundation models can identify and capture logical relationships within data facts, which helps users automatically arrange the positioning of events when creating narrative visualizations. For example, Erato [3] and Chart2Vec [71], as introduced in Section 4, converted data facts into vectors whose similarities directly indicated not only the semantic relationships but also the underlying logical connections between the corresponding facts. These approaches allow foundation models to effectively identify and represent complex data relationships within their respective contexts. Additionally, foundation models can also function as intelligent agents [96]. By leveraging their inherent capabilities in reasoning and contextual understanding, these models can be guided to perceive and respond to user input, exhibiting goal-oriented behavior [97]. For example, Shen et al. [98] proposed Data Director to automatically create data videos. In the Data Director system, GPT-4 was utilized as an intelligent agent to generate narrative text that established logical connections between data facts. Specifically, the system placed the GPT-4 in the role of a data analyst. By following a chain-of-thought approach, carefully designed prompts guide the GPT-4 to generate narrative text that not only connects insights into a coherent story but also ensures that the logic is compelling. The prompts explicitly instruct GPT-4 to avoid merely enumerating insights, thereby encouraging the creation of a more integrated and logically sound narrative.

Structure, compared to logic, is a higher-level concept that determines how the entire narrative content is organized. There are many types of structures [99], [100]. One famous example is known as Freytag's Pyramid Structure [101], which organizes the narrative content into four different stages, including initializing, rising, peaking, and releasing.

Foundation models have been used to help users interactively create coherent narratives that conform to a specific structure. Some advanced techniques have been developed. In particular, instead of generating narrations based on a fixed structure, TaleBrush [54] allows users to interactively create a structure by simply sketching a horizontal curve on canvas. GPT is then used to generate the narrative content based on this curve with the changing slopes in the curve indicating the positive or negative change of the generated storyline. In this way, by using this tool, users can create data stories with multiple climaxes and turning points in the plot.

Despite the interactive structure adjustment, narrative structures can also be automatically adjusted by fine-tuning the foundation models. For example, NAREOR [17] fine-tuned foundation models using two innovative training methods to control narrative structure. The first method, NAR-denoise, mimicked the human rewriting process by introducing noise through deletion and swapping of tokens. It learned to generate high-quality text from noisy input and was further fine-tuned on supervised data. The second method, NAR-reorder, directly addressed the reordering task. It used an input encoding scheme that enabled the model to recognize and handle different sentence structures and coreference chains. Through these two stages of training, the models could generate text in a target narrative order while preserving the story's plot. These methods not only enhanced model performance in terms of fluency and logical coherence but also demonstrated the potential of fine-tuning foundation models to create more interesting and diverse narrative structures.

Reflection. Logic is enhanced by foundation models, which identify and organize relationships between data facts to create coherent and contextually accurate narratives. Structure is shaped by foundation models, which enable both interactive creation [54] and automatic adjustment of narrative forms [17], allowing for diverse and well-organized storytelling. In narrative visualization, maintaining logical coherence is crucial, as it often involves various forms of reasoning, such as elaboration, similarity, generalization, contrast, temporal sequencing, and cause-effect. These logical structures help create coherent and compelling narratives [2], [94]. However, the current tools and methodologies do not yet fully leverage foundation models to explicitly clarify and categorize the diverse types of narrative logic.

To address these challenges, future research should focus on developing methods to explicitly clarify and categorize logical relationships between data facts. For example, develop algorithms within foundation models to identify temporal sequences by recognizing time-based patterns in data, or they could detect cause-effect relationships by analyzing correlations and dependencies between variables. Furthermore, future research could create interactive visual tools that allow users to map the logic. These tools could work in tandem with foundation models to automatically generate narratives based on the user-defined logic.

5.2 Description

After establishing the logical framework of a narrative visualization, the crafting of detailed descriptions becomes an essential step in adding content to the narration. Description adds textual details to narrative visualization. It weaves the data insights into a coherent and engaging story to captivate and guide the audience. Foundation models, with their sophisticated ability to recognize and generate coherent and contextually relevant text [102], play a critical role in this phase. Recent research has focused on four main aspects: generating titles, captions, annotations, and scripts.

Title is a line of text or a phrase that typically appears at the top of content, summarizing the main content or theme. The primary purpose of a title is to capture the reader's attention and provide an initial impression of the subject matter. It is often a brief and concise overview of the content [55], which is an important component of effective narrative visualization [103]. Foundation models are particularly adept at extracting essential information from visualizations and condensing it into impactful and succinct titles. For example, AutoTitle [55] proposed an interactive system designed to generate titles for visualizations by extracting data, computing relevant facts, and using a deep learning model to create natural language titles. The system begins by reverse-engineering the given visualization to extract the underlying data and create atomic facts, which form the foundation for higher-level facts

through operations like aggregation, comparison, trend analysis, and merging. A diverse dataset of fact-title pairs is constructed and used to fine-tune the T5 model, generating fluent and diverse titles. Users can interact with an interface to upload visualizations, explore generated titles using Radar and RadViz views, and select the most appropriate ones. The interface also includes a representative title view that displays high-quality titles based on the calculated metrics. A user study confirmed the system's ability to produce high-quality titles and validated the usefulness of the metrics.

Caption is typically a brief piece of text placed under images, charts, or graphics to describe or explain these visual elements. The main purpose of a caption is to interpret the picture or chart, provide necessary background information, or explain the relationship between the image content and the main text [24]. The task of creating a caption for charts using foundation models is a complex one, involving the disciplines of computer vision and natural language processing [104]. For example, Kantharaj et al. [40] conducted a study collecting 44,096 pairs of chart titles to compare the performance of foundation models in chart captioning. Similarly, Vistext [21] proposed a dataset of 12,441 pairs of charts and captions, which describe the charts' construction, report key statistics, and identify perceptual and cognitive phenomena. This dataset was then used to fine-tune an LLM to generate coherent and semantically rich captions.

However, this end-to-end generation system limits author involvement in the title generation process, potentially overlooking the author's intended message for the titles. To address this, Intentable [23] introduced a mixed-initiative caption authoring system. Intentable allowed users to express their intents through an interactive interface. These intents were encoded as JSON-based 'caption recipes'. Utilizing Transformer architectures fine-tuned for this purpose, the system generated accurate and fluent natural language captions based on these recipes. Evaluation results demonstrated that Intentable produces more accurate and flexible captions compared to traditional fully automated methods. Similarly, Liew et al. [24] used GPT-3 to generate engaging captions for visualizations. The datasets were sourced from Kaggle to create scatter plots, followed by clustering analysis to identify data structures. Effective prompt engineering was employed, involving basic templates, instructional guidance, and an interactive question-and-answer format with the user. The study demonstrated that the incorporation of user interaction and detailed prompt design improved the quality and engagement of GPT-3-generated captions.

Both Intentable and Liew's research utilize LLMs to generate captions for visualizations but employ different methods. The Intentable system offers highly customized caption generation through user interaction and various intent types, providing flexibility but requiring complex design. In contrast, the system using GPT-3 focuses on rapid caption generation with a tiered approach to prompt engineering, making it quick and user-friendly. The former is ideal for scenarios requiring personalized expression, while the latter suits quick analysis and display needs.

Annotation is an additional explanation or commentary provided for text, images, videos, or other narrative content. It can offer background information, explanations, critiques, or a brief analysis of the content [105]. Foundation models can generate annotations, enhancing the information conveyed in narrative visualization by providing more details and context for the audience [106], [107]. For instance, Bromley et al. [26] introduced a novel approach that combines feature-word distributions with the visual features and data domain of charts to annotate visual chart features. This feature-word-topic model can identify words associated with vocabulary that are semantically similar but subtly different. Compared to classic automated annotation pipelines like Contextifier [108], foundation models can simplify text mining and image processing techniques. They can even generate annotations through simple prompt engineering. This contrast highlights the advancements foundation models bring to the field, streamlining the annotation process and offering more sophisticated and context-aware outputs with minimal manual intervention.

Script in the context of data storytelling typically refers to a written guide used to organize and narrate the story of data [50]. Foundation models can generate personalized narrative texts based on specific needs or goals [25], [35], [58], [109], [110], which means that stories that adapt to different styles, accents, or themes can be generated according to the requirements of the user. This is one of the strongest abilities of the foundation model. For example, Islam et al. [111] proposed DataNarrative, a system for automated data-driven storytelling that combines visualizations and text. They presented a multi-agent framework that employs two LLMs: one for understanding and describing the data, generating an outline, and narrating the story, and another for verifying the output at each step. Similarly, Pinargote et al. [112] proposed a method to automatically generate data narratives in learning analytics dashboards using GPT-3.5. It analyzed and automatically coded classroom meeting transcripts with GPT-3.5 to generate coherent narratives. Furthermore, Ghosh et al. [20] proposed a new method for analyzing trajectory data called spatio-temporal narratives, interpreting trajectory data as stories. Now, foundation models can autonomously generate scripts and stories, meeting a diverse range of user needs and preferences [54].

Reflection. Title and caption generation by foundation models involve extracting key information from visualizations to create concise and impactful summaries. Annotation is enhanced by foundation models, which generate additional context and explanations to enrich the content. Script generation by foundation models allows for the creation of personalized narrative texts that adapt to different styles and themes. Due to the potential for hallucinations, the use of foundation models can lead to ethical issues, particularly when the generated narratives are disseminated to a broad audience. These biases can skew the information, potentially reinforcing stereotypes or spreading misinformation, which undermines the integrity and fairness of the content [11].

To address these challenges, it is essential to adopt a more careful and conscientious approach to using foundation models for narrative generation. Future research could develop advanced algorithms that can detect and correct biases in real-time as the narrative is generated. These algorithms could analyze the content for signs of bias or cultural insensitivity and make adjustments before the narrative is finalized.

6 Visualization

In this section, we review 22 papers related to the narration stage, which can be broadly classified into two categories based on the role of foundation models: (1) visualization generation and (2) visualization authoring. Specifically, generation pertains to the automated creation of content using foundation models, whereas authoring highlights the user's capacity to create and edit narrative visualization content that has already been generated.

6.1 Generation

Generation refers to converting data into graphical representations guided by the narrative context and insights derived from the data. Foundation models suitable visualizations to accommodate user-generated visual types and personalization, which is key in narrative visualizations to fit specific user preferences [6]. Therefore, recent research on visualization generation has focused on three main aspects with foundation models, i.e., charts, images, and animations.

Chart refers to a graphical representation of data. Charts use visual elements such as points, lines, and bars to display relationships or trends in the data. Foundation models offer two primary advantages in narrative visualization. On the one hand, foundation models can recommend the most suitable chart type based on data insights and narrative purposes [29]. For example, LLM4Vis [29] introduced an interpretable visualization recommendation method. This approach guided ChatGPT to generate comprehensive text descriptions for each tabular dataset, convert these features into vectors, and then employ clustering algorithms for visual recommendations. Consequently, this method eliminated the inefficiencies associated with manual chart selection. On the other hand, foundation models demonstrate the capability to directly generate charts by producing code, a competency validated by multiple studies. For example, Pere-Pau Vázquez et al. [113] evaluated the performance of ChatGPT-3 and ChatGPT-4 in generating different types of charts using various visualization libraries such as matplotlib, Plotly, and Altair. Their findings revealed that ChatGPT-4 excelled in producing up to 80% of the expected charts.

Recent studies have also explored the integration of foundation models' generative capabilities into systems designed to assist users in creating narrative visualizations for practical applications [27], [28], [30], [31], [84], [114]. For example, Chat2VIS [27], [28] converted natural language into code for appropriate visualization. The user started by selecting a dataset and entering a natural language query to specify the desired visualization. The system then created prompts that included a data description and a Python code template to assist the LLM in understanding the context and generating precise code. In contrast to Chat2VIS, another notable work, LIDA [6], has introduced a system that automatically employs a multi-stage, modular approach to generate data visualizations and infographics using foundation models. LIDA used these models in various modules to perform tasks such as data summarization, goal generation, visualization code creation, and the generation of stylized visualizations. Compared to Chat2VIS, LIDA is a more comprehensive end-to-end visualization generation system. LIDA allows user intervention at every stage. It also features robust multilingual interfaces for optimizing, explaining, and evaluating charts, making it suitable for novice users and those with technical backgrounds. The previously mentioned research uses online models, which might have suffered from inherent hallucination issues associated with foundation models. In contrast, ChartGPT [115] developed a visualization-specific LLM aimed at improving chart recommendations. The approach entailed the development of a comprehensive dataset comprising abstract statements paired with their corresponding charts. Although foundation models can assist in generating charts, current research predominantly focuses on the creation of individual charts, whereas narrative visualizations often require multiple logically connected charts.

Image refers to the illustrations and graphics used to enhance, interpret, or supplement data narratives. These images may include characters, scenes, emotional expressions, or other visual elements that support the storytelling aspect of the narrative. Foundation models like DALL-E [116], [117] have revolutionized this domain by enabling the automatic creation of images from textual descriptions. Such models enrich narrative content by producing images that align closely with the underlying story, thereby elevating its appeal and comprehensibility [118], [119]. Schetinger et al. [45] highlighted that generative text-to-image models present substantial opportunities in visualization processes. This included the design of visualizations, the creation of mood boards, and the generation of explanatory images. For example, Ghosh et al. [20] leveraged DALL-E to provide visual explanations for generated semantic trajectory visualizations, offering a more intuitive way to understand and interpret mobility patterns. This method provides a richer and more comprehensive depiction of the scenarios under study. Foundation models can generate images through text-to-image or image-to-image methods, effectively enriching the visual aspect of narrative visualizations. Future research could explore how to generate images in different styles based on user preferences, further enhancing user engagement and appeal. **Animation** is a technique that brings static images and graphics to life visuals by creating the illusion of movement [122]. In the context of narrative visualization, data video is an important form of narrative visualization, integrates visualization with animation to convey data-driven stories [47]. Traditionally, creating a data video required professional skills. However, foundation models offer two primary advantages in narrative visualization.

The first advantage is that foundation models can connect narrative text with visual elements. For example, Ying et al. [123] used Graph Neural Networks to parse static SVG charts and extract their data and visual encodings. They then utilized GPT-3 to generate descriptive and insightful audio narrations based on the charts. By adding appropriate animations to the chart elements and synchronizing them with the narrations, static charts were transformed into dynamic charts with enhanced interactivity and information delivery. Similarly, Data Player [124] proposed a method for automatically generating data videos. This technical solution first extracted data by parsing static SVG charts and mapping visual elements to rows in a data table. Then, it used GPT-3.5-turbo to semantically match narrative text segments with data table rows, creating links between text and visual elements. Both studies involved extracting data from static charts and transforming it into a more manageable format. They utilized foundation models for natural language processing to generate descriptive narrations and establish links between text and visual elements, and employed programming or constraint-solving solutions to generate animation sequences synchronized with narrations.

The second advantage is that foundation models can create animations by generating interpolations between the start and end images. Generative Disco [125] used foundation models to integrate visual elements with music rhythm to generate animations. It utilized models such as GPT-4 to generate prompts that encompassed text and visual objectives, aligning these with the lyrics and rhythm of the music to ensure coherence with the musical content. For the animation interpolation process, Generative Disco employed Stable Diffusion to create transitions between the start and end images, analyzing music beats and dynamics to produce animation effects synchronized with the musical rhythm. In general, current animation generation techniques focus primarily on connecting individual images and charts to animations using interpolation or

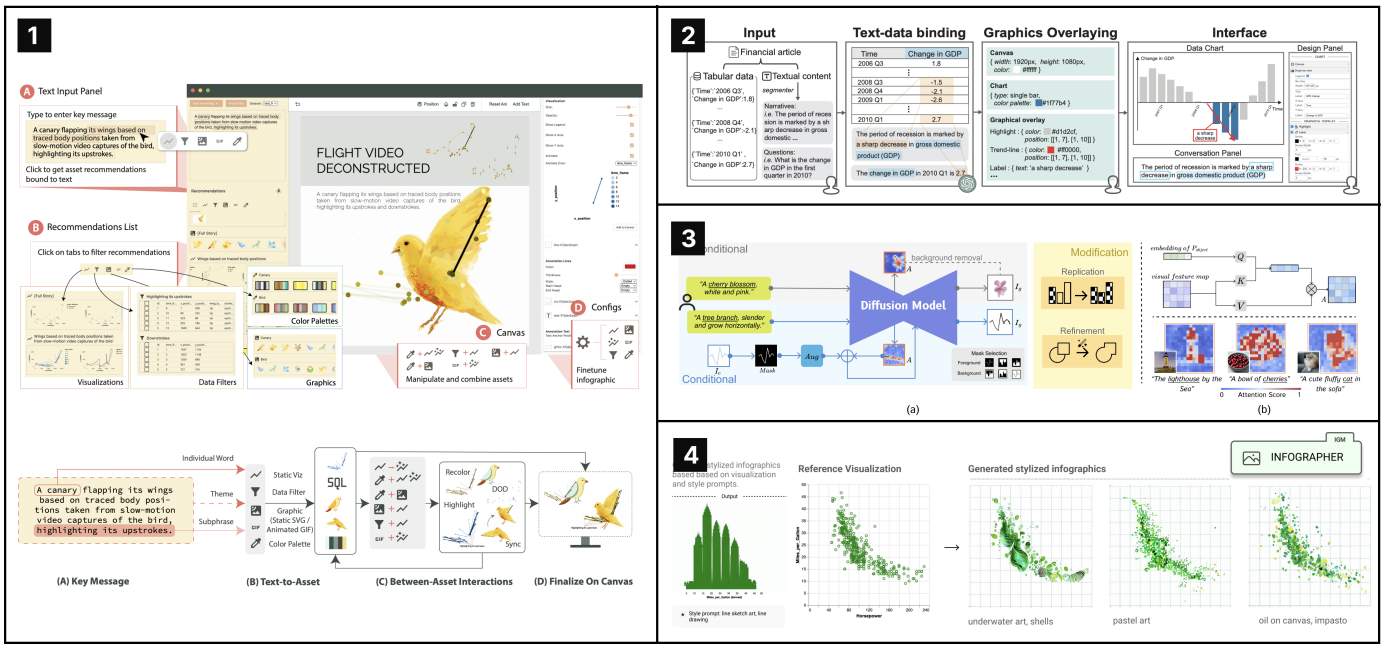


Fig. 4: Selected examples of editing. (1) Epigraphics: message-driven infographics authoring [120]. (2) FinFlier: layer visual elements onto charts of financial narrative visualizations [121].(3) ChartSpark: embedding semantic context into charts with text-to-image generative model [56]. (4) LIDA: generate stylized graphics based on visualizations for infographics [6].

rule-based syntax. However, there is a lack of examples in which foundation models are used directly to generate animations.

Reflection. Charts can be generated by foundation models through code production, creating visualizations that represent data. Images can be produced by leveraging the text-to-image capabilities of foundation models, enriching the content of data comics and other narrative visualizations. Animations are created when these charts and images are connected according to a narrative structure, with transitions such as interpolation applied, effectively conveying complex data stories. Although foundation models perform well in generating most standard charts, they still face notable difficulties when producing more complex visualizations, such as Sankey diagrams and composite charts. Additionally, another challenge arises when foundation models are required to configure multiple visual variables simultaneously [126]. For example, when multiple variables need to be adjusted, the generated code may not execute correctly, resulting in errors.

To address these challenges, future research should focus on improving foundation models in designing more complex and composite charts [124]. Overcoming limitations in handling comprehensive visualizations will enhance the utility of foundation models. In addition, most methods currently rely on linking charts to create animations. However, with the advancement of video models, video foundation model is expected in the future that is fine-tuned specifically for narrative visualizations.

6.2 Authoring

Authoring is the process of transforming individual materials—such as narratives, charts, and images—into a cohesive and compelling narrative visualization [120]. It involves editing and enhancing each element to align with user needs while establishing connections among them to create a unified flow. In this process, foundation models primarily assist users in establishing relationships between materials and correcting errors. Additionally, their generative capabilities enable the editing of materials to enhance clarity and aesthetics and the effectiveness of information communication. Therefore, recent research in this domain has focused primarily on three main aspects: encoding, embellishment, and layout.

Encoding refers to the specification used to map data attributes to visual elements. It enables data to be represented visually, allowing people to understand and analyze the data intuitively [127]. Previous research has aimed to help users detect and correct encoding errors in visualizations. For example, Chen et al. proposed VizLinter [128], a technical solution that incorporates a visualization linter to detect errors in visualizations and a visualization fixer that employs a linear optimization algorithm to automatically suggest and apply corrections. More recently, Lo et al. [129] evaluated the abilities of foundation models to identify improper encoding. In this study, multimodal foundation models detected chart errors, particularly encoding errors, by employing prompts of varying complexity, ranging from simple direct questions to complex chain-of-thought strategies. The researchers utilized multi-round conversation prompts to guide the models through a progressive analysis of chart issues. Experimental results demonstrated that foundation models performed significantly better in identifying encoding errors when guided by chain-of-thought prompts. Therefore, this study highlighted the potential of multimodal foundation models to enhance encoding accuracy and reliability through advanced prompt engineering techniques. By employing chain-of-thought reasoning, foundation models can deduce encoding errors. Future research could explore whether these models can be further guided through dialogue to generate code that corrects encoding errors.

Embelishment refers to the addition of decorative and ornamental elements in the process of data visualization to enhance visual impact and user comprehension [130]. Foundation models can be leveraged to analyze data and context, enabling the recommenda-

tion of suitable graphical elements such as icons and illustrations to enhance the visual impact and information delivery [121]. For example, Zhou et al. [120] proposed Epigraphics, an authoring tool that enhances the aesthetic cohesiveness of the materials, as shown in Fig. 4 (1). The system recommends visualizations, graphics, data filters, color palettes, and animations based on text inputs, supporting interactions like recoloring and highlighting. GPT-3.5 analyzes user input to generate chart specifications and SQL for data filtering. Sentence-BERT optimizes recommendations by calculating text-graphic similarity, while BLIP provides image captions. Adobe Firefly creates images and suggests colors, collectively improving infographic content and presentation. In another study, FinFilter utilized GPT-3.5 to automate the graphical overlays for financial narrative visualizations, as shown in Fig. 4 (2). The study summarized common graphical overlays and narrative structures from a survey of 1,752 layered charts. By applying prompt engineering techniques, financial knowledge was integrated into GPT-3.5, enabling it to efficiently identify complex structures and link terminology with data. Foundation models were used to automatically generate layered charts, optimizing overlay positions and colors to enhance visualization effectiveness.

Recently, research also focused on stylized and pictorial visualizations based on text-to-image or image-to-image models. These approaches replace the foreground and background of charts with semantically relevant graphics. For example, Viz2viz [131] employed a combination of generative diffusion models and image processing techniques to transform traditional visualizations such as bar charts, area charts, pie charts, and network graphs into highly stylized forms. ChartSpark [56] introduced a three-stage framework—feature extraction, generation, and evaluation—to embed semantic context into charts using text-to-image generative models, as shown in Fig. 4 (3). MPNet and Word2Vec facilitated keyword extraction and thematic context. The Frozen Latent Diffusion Model integrated data and semantic context. An evaluation stage ensured chart fidelity through pixel-wise comparison and edge detection, enhancing interpretability and credibility. In another study, LIDA [6] generated stylized graphics to refine the infographics by using a user-editable library of visual styles and the text-conditioned image-to-image transformation with diffusion models, as shown in Fig. 4 (4).

Layout is the underlying semantic structure that links graphical elements to convey the information and story to the user [132]. Recent advancements in foundation models have enabled the parsing of natural language instructions to facilitate chart layout adjustments. ChartReformer [133] achieves this adjustment by parsing user natural language instructions, using a pre-trained Visual Language Model(VLM) to extract the current visual properties and layout information of the chart, generating new drawing parameters, and invoking charting libraries to redraw the chart. The specific steps include: first, the system parses user instructions using NLP techniques to understand the desired chart layout adjustments. Next, the system uses the VLM to analyze the input chart image and extract its current attributes, such as axes, gridlines, and legend positions. Based on the parsed user instructions, the system is able to generate new drawing parameters. Finally, the system invokes charting libraries like Matplotlib or Vega-Lite to apply the new drawing parameters and redraw the chart such that the layout is adjusted according to the user's instructions. This process combines natural language processing, visual language models, and advanced charting libraries to achieve precise and user-driven chart layout editing. Similarly, the DynaVis system [134], powered by the GPT-3.5, enabled users to edit various aspects of visualizations using natural language commands, including chart layout, visual properties, data filtering and transformation, and axis range adjustments. Current layout editing tools primarily focus on basic adjustments such as modifying axes and legends. However, the overall distribution of visual elements, the sequence in which charts are presented, and the spatial organization within a visualization can greatly influence how users interpret and engage with the data. Future research could delve into more sophisticated layout strategies that optimize the user experience.

Reflection. Foundation models enhance the authoring process by improving encoding, embellishment, and layout in visualizations. For encoding, these models help detect and correct errors, ensuring accurate data-to-visual mappings. For embellishment, they recommend and generate graphical elements like icons and illustrations, enriching the visuals through advanced data analysis and contextual understanding. Finally, for layout, foundation models enable precise adjustments by parsing natural language instructions, extracting visual properties, and applying new parameters via charting libraries, allowing for user-driven editing.

However, current authoring tools exhibit several limitations. First, the range of supported chart types and narrative visualization formats is limited. Future authoring tools should expand to include a broader variety of chart types and narrative forms to better represent complex data and tell diverse stories. Second, the customization options in current tools are inadequate, restricting users' creative expression. To address this problem, incorporating sketching or hand-drawn authoring capabilities could allow users to adjust visual elements more freely [120]. Finally, layout design in current tools lacks effective control over the overall information flow and the ability to guide the viewer's reading sequence. Future research could adopt more innovative layout styles, such as spiral or star arrangements, to improve narrative readability and more effectively guide the viewer's attention [132].

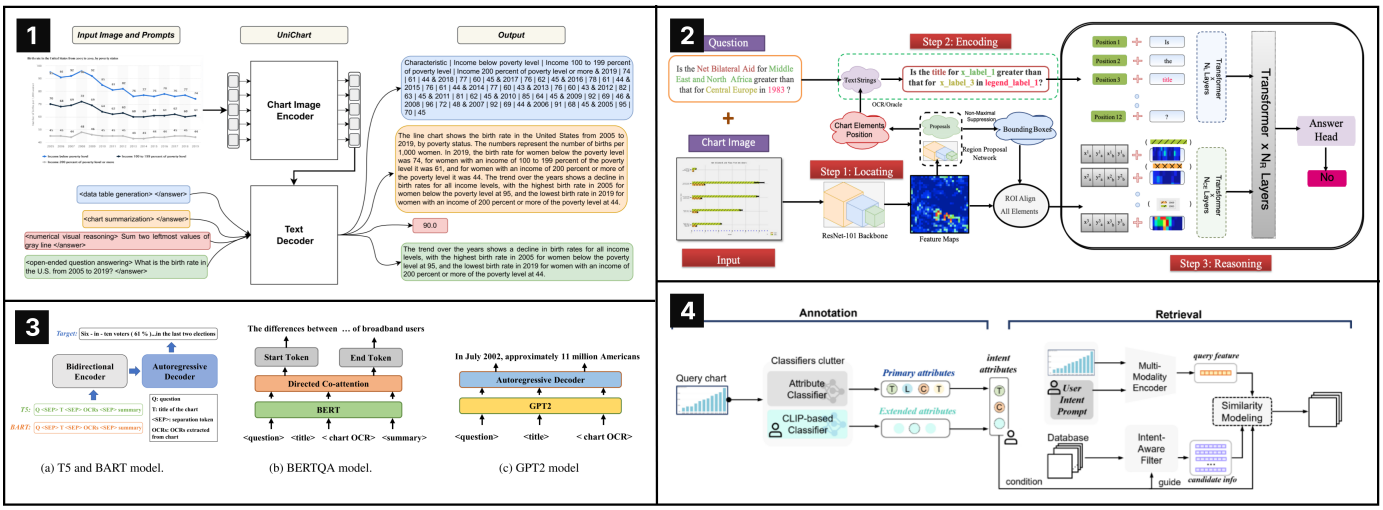
7 Interaction

We reviewed 15 papers related to the interaction stage, which can be broadly classified into two major categories based on the role of foundation models: (1) chart interpretation and (2) data navigation. This section will discuss both categories in detail.

7.1 Chart Interpretation

Chart interpretation involves the automated "reading" of narrative visualizations in a way that simulates human perception [43]. Understanding important patterns and trends from charts and answering complex questions related to charts can be cognitively demanding. Furthermore, accessibility challenges exist for people with visual impairments in understanding charts. To address these challenges, recent research has focused on two main aspects: chart derendering and chart understanding [135] based on foundation models.

Chart Derendering is the process of converting charts into tables [41], [138]. This process presents significant challenges due to the variety of chart types and the complexity of their components. Existing approaches are heavily based on heuristic rules for detecting chart components, which require substantial domain knowledge to develop [139]. Foundation models, however, have shown remarkable proficiency in plot-to-table tasks without explicitly recognizing chart structures and components [41], which



(a) T5 and BART model. (b) BERTQA model. (c) GPT2 model

Fig. 5: Selected examples of Chart Understanding. Chart Understanding is divided into three tasks: **Chart Summarization**: (1) Unicharts: a model that processes visual elements through a chart encoder and text decoder to generate natural language summaries [135]; **Chart Question Answering**: (2) STL-CQA: a structured Transformer model to localize chart elements and perform cross-modal reasoning [136], (3) OpenCQA: Multiple foundation models were fine-tuned to answer open-ended questions with charts [42]; **Chart Retrieval**: (4) WYTIWYR: a user intent-aware framework with multimodal inputs for visualization retrieval [137].

led to their increase in automatic content recognition and extraction. A notable example was MATCHA [140], which was designed based on the encoder-decoder structure of the Pix2Struct model and achieved chart-to-table conversion through end-to-end training. The model employed a Transformer architecture with 300 million parameters and was fine-tuned on real-world chart-table pairs to generate linearized tables. On the basis, Deplot [36] further fine-tuned MATCHA by focusing exclusively on the task of chart-to-table conversion. To better accommodate complex charts with numerous data, Deplot utilized a longer sequence length, enhancing its performance in chart derendering tasks. By decomposing the charts into tables, users could input the resulting tables along with the questions as text prompts, which can then be further processed by the foundation models for enhanced chart understanding.

Chart Understanding aims to attain a comprehensive interpretation of visualizations and tackle complex analytical tasks. [44]. Numerous studies have advanced this field by collecting extensive datasets, fine-tuning foundation models, and integrating multiple chart understanding tasks to improve the performance of narrative visualizations [37], [141]–[143]. Foundation models can solve three downstream tasks: (1) Chart Summarization [39], [40], (2) Chart Question Answering [144] and (3) Visualization Retrieval [137].

Chart Summarization aims to articulate the important data points and trends presented in a chart using natural language [39]. This task could be beneficial for readers seeking to quickly grasp the main points of a chart [40]. Leveraging their reasoning and generative abilities, foundation models can produce concise and accurate textual summaries that capture the essential information and trends in charts. For example, Unichart [135] comprised a chart encoder to process the relevant charts, text, data, and visual elements of the chart, and a text decoder to generate the corresponding natural language output, as illustrated in Fig. 5 (1). To enhance chart comprehension and reasoning capabilities, Unichart was trained on a large corpus of 611K charts with various pre-training tasks, including data table generation, data value estimation, numerical and visual reasoning, open-ended QA and chart summarization. Although Unichart excelled in generating concise and accurate textual summaries, it did not perform as well as specialized QA systems in handling open-ended questions or complex logical reasoning.

Chart Question Answering is a task to take a chart and a question in natural language as input to generate the desired answer as output [145]–[147]. Foundation models are capable of understanding and processing user queries to provide relevant and accurate answers. These models go beyond surface-level information analysis, employing logical reasoning and contextual analysis to provide deeper insights in response to user inquiries. For example, STL-CQA [136] employed a structured Transformer model to localize and encode chart elements for answering natural language questions about charts. STL-CQA's technical approach involved three key steps: first, detecting and localizing chart elements using a Mask R-CNN model; second, dynamically encoding questions by replacing specific words with standardized tokens; and finally, using a structured Transformer module for chart structure understanding, question understanding, and cross-modal reasoning, as shown in Fig. 5 (2). However, STL-CQA faced limitations when dealing with open-ended questions or tasks requiring larger parameter models. In contrast, OpenCQA [42] leveraged larger parameter models to handle open-ended questions about charts, enabling deeper logical reasoning and contextual analysis. OpenCQA's technical approach encompassed three task settings: using charts and their accompanying full articles, providing only relevant paragraphs, and generating answers based solely on the charts. As shown in Fig. 5 (3), OpenCQA's dataset included 7,724 open-ended questions and descriptive answers, and multiple foundation models were fine-tuned. Through various model architectures and rigorous evaluation methods, OpenCQA showed the potential to generate detailed descriptive responses. However, it still faced challenges in handling complex logic and mathematical reasoning. Although this method required high computational resources, it offered deeper insights when addressing complex problems.

Chart Retrieval refers to the process of finding charts from a set of charts that meet specific query conditions or requirements.

Chart retrieval can efficiently locate relevant visualizations from vast datasets, enabling users to quickly access and utilize the most pertinent charts for their needs. In this task, foundation models can be used to map textual and visual features into a shared embedding space, which enables zero-shot classification of the query charts. For example, WYTIWYR [137] introduced a user intent-aware framework for chart retrieval through multi-modal inputs. This framework involved two main stages: annotation and retrieval. In the annotation stage, a deep neural network classified query charts and decoupled their visual attributes using the CLIP model for zero-shot classification. In the retrieval stage, an intent-aware filter sifted through the chart database based on user-selected intent attributes, while multi-modal encoders generated joint feature representations from the inputs, as shown in Fig. 5 (4). WYTIWYR integrated user intent into the chart retrieval process, providing highly customized results. However, this approach required substantial computational resources, and designing effective text prompts remained a challenge.

Reflection. Chart Derendering leverages foundation models to convert charts into tables through advanced neural networks, effectively bypassing the need for explicit recognition of chart structures and components. Chart Understanding utilizes foundation models to process visual elements and textual data, enabling tasks such as summarization, question answering, and retrieval, which allow for comprehensive interpretation of narrative visualizations. However, current techniques are primarily focused on derendering and understanding individual elements, such as single charts, which are often applied in general visualization. In contrast, narrative visualization requires a cohesive flow of context and relationships between multiple charts, as these connections are essential for building a comprehensive and meaningful story.

To address these limitations, a promising area for future research lies in exploring how to develop a holistic approach that cohesively integrates multiple charts. This involves not just interpreting individual charts, but also understanding how they interact with each other to form a unified story. Foundation models can be further developed to analyze and interpret these interrelationships. Users are able to obtain a more comprehensive and logically coherent understanding of the entire narrative visualization [145]. For example, future research can explore the design of contextual links between charts, where hovering over or clicking on an element in one chart provides insights or highlights relevant data in other charts. This would enhance the user's ability to see connections across different visual elements and deepen their overall understanding of the narrative.

7.2 Data Navigation

In narrative visualization, data navigation is the process where users browse and explore data via a visual interface. This includes activities like selecting different data views, zooming, and scrolling to access various parts of the data, and altering the data representation using interactive elements like sliders, buttons, or drop-down menus [7].

Navigation refers to how individuals engage with visualizations to explore data [43]. Foundation models have enhanced user interaction with narrative visualizations by enabling more intuitive navigation and exploration through natural language queries. These models allow users to interact with visualizations simply by adjusting the prompts, which reduces the barriers to accessing and interpreting complex data. Furthermore, foundation models provide the ability to customize data views based on individual user interests and requirements. They adapt data displays according to user-specified parameters and can even generate code to offer a more personalized experience. For example, Chen et al. [7] demonstrated how GPT-4 interacted with charts by sending JavaScript events, generating code to update axes, resetting zoom levels, and mimicking hover events for tooltips. This extended the adaptability to tailoring data insights for various audiences, such as GPT-4, which can generate audience-specific reports for different entities such as the governments of African countries, the United Nations, and primary school students.

In addition to natural language interaction, graphical methods offered new ways to express analytical intent. InkSight [38] explored a novel interaction method that allowed users to directly draw paths on data visualization charts to express their analytical focus. This intuitive method allowed users to select subsets of data or patterns of interest with more precision, which were then automatically annotated by the system. InkSight's approach revolved around capturing user intent and automatically generating documentation that reflected this focus. By sketching directly on the charts, users expressed their analytical intent, which the system interpreted and converted into natural language documentation using GPT-3.5. This documentation could be further refined through editing, making it precise, and easy to understand, and enhancing the efficiency of recording and sharing data analysis results.

Reflection. Navigation is enhanced by the foundation models through the use of NLP and adaptive algorithms, allowing users to intuitively interact with and explore data by adjusting prompts and customizing data views. Although foundation models have shown the potential to lower barriers to audience engagement and interaction, previous research indicates that it remains challenging to describe interactions. Non-experts, in particular, often find it difficult to naturally express complex interactive functions using natural language [113]. Firstly, non-expert users are generally unfamiliar with the technical language required to effectively guide the model. Foundation models, while powerful, still require precise and clear prompts to generate the desired visualizations. However, users without a background in narrative visualization may struggle to articulate their intentions in a way that the model can accurately interpret. This gap between the user's natural language and the model's interpretive capabilities often leads to misunderstandings, resulting in interactions that do not fully meet the user's needs. Secondly, when presented with a narrative visualization, non-expert users often find themselves unsure of how to proceed. The complexity of interacting with and manipulating the visualization can be overwhelming, especially when the visualization includes multiple layers of data or advanced interactive elements.

To address these limitations, future research should focus on providing users with more explicit guidance on operational methods. Research should explore ways for users to learn these methods. Potential learning pathways include [50]: ***Explicit Instruction:*** Offering clear, step-by-step operational explanations to help users quickly grasp basic operations. ***Tacit Tutorial:*** Implementing implicit tutorials and progressive guidance to allow users to naturally master complex features during usage. ***Initial Configuration:*** Providing a well-designed initial setup that is easy to understand and operate, reducing the difficulty of getting started. By combining the generative capabilities of foundation models with these interaction design strategies, future narrative visualization tools can serve user needs better and facilitate the widespread application of narrative visualization technology.

8 EVALUATION

In this section, we discuss existing methods for evaluating the reliability and quality of narrative visualizations created by foundation models. Furthermore, we propose future research directions to offer valuable insights for researchers and practitioners in the field of narrative visualization.

8.1 Methods of Evaluation

Although foundation models provide convenience in creating narrative visualizations, their widespread application faces some potential challenges. One major issue is the variability in how these models can be instructed, as different prompts can lead to divergent and unpredictable results of the process. Moreover, the presence of hallucinations, where models generate inaccurate or non-sensical content even in tasks as structured as visualization generation, further complicates their use [148].

Recently, several studies have emerged to evaluate visualizations created by foundation models [7], [79], [129] through various human evaluations from various perspectives to identify their strengths and weaknesses. However, in practical applications, automated evaluation methods and a comprehensive set of metrics are necessary to evaluate the visualizations generated by these models effectively. For example, Podo et al. [148] proposed a conceptual framework called EvaLLM, which systematically dissects and models the evaluation of data visualizations generated by LLMs. This framework consists of five layers. From bottom to top, they are the **Code layer** (evaluating the consistency of the visualization within the code environment), the **Representation layer** (assessing the semantic aspects of the visualization representation), the **Presentation layer** (evaluating the data presentation quality of the visualization from the perceptual perspective), the **Informativeness layer** (measuring the intrinsic quality of the visualization), and the **LLM layer** (evaluating the generation strategy and the significance of the visualization). Each layer includes different levels that support the evaluation process, detailing aspects in terms of evaluation scope, importance, implementation approach, evaluation affordability, semantics, and human-machine ratio.

To enhance the generalizability of the evaluation, Chen et al. [149] proposed a dataset that includes 2,524 representative queries, covers 146 datasets, and is paired with accurately labeled ground truth. They also introduced a comprehensive automated evaluation method encompassing multiple dimensions: The **validity checker** ensures the code generates visualizations by first running it in a sandbox to avoid crashes, then checking for essential code snippets to render the visuals. The **legality checker** verifies if the visualization matches the query by extracting key details (e.g., chart type, data) and checking their accuracy and order using meta-information. The **readability evaluator** evaluates visualization clarity through layout and scale/tick checks. It assigns a readability score (1–5) based on factors like titles, labels, and colors.

Reflection. These evaluation frameworks effectively decompose the tasks involved in assessing data visualizations created by foundation models. However, beyond traditional evaluation criteria, narrative visualizations demand a more comprehensive framework that incorporates new metrics specifically related to narratives, such as narrative coherence, narrative relevance, and audience engagement. For instance, while it is important to assess the technical accuracy of the visualizations, assessing the narrative logic is equally important. This includes ensuring that the story flows coherently, the insights are effectively communicated, and the data meaningfully supports the narrative. Furthermore, the engagement factor should be considered, as a good narrative visualization must not only present data but also captivate the audience's attention. Evaluation should consider elements like the clarity of storytelling, as well as the emotional or cognitive impact the visualization has on the viewer.

8.2 Research Opportunities of Evaluation

Although effectively evaluating and enhancing the narrative quality of these models remains a critical challenge, relatively comprehensive visualization evaluation methods for data visualization are available. However, evaluation methods specifically tailored for narrative visualization require further refinement. Additionally, improving evaluation efficiency and enhancing user trust through evaluation are also future research directions.

Using the Foundation Model as an Evaluation Method. Future research can explore how to use the generative model itself to evaluate its narrative output [148]. Leveraging the self-evaluation capabilities of the model can lead to more precise and efficient assessments of narrative quality. This involves training the model to understand and evaluate the coherence, logic, and linguistic quality of its own generated content, thereby enhancing the overall standard of the narratives produced. This approach can reduce the subjectivity of human evaluation and improve the consistency and reliability of assessments.

Visualizing Evaluation and Error Correction. Visualizing the evaluation and error correction processes of generated narratives can help users more intuitively understand the quality and issues in the model output [56]. Future research could develop more tools that allow users to see evaluation results and perform real-time error correction. This not only aids users in better understanding and utilization of the model but also promotes the improvement and optimization of the model itself. Through visualization, users can more easily identify and correct errors in the generated content, enhancing the quality and acceptability of the narrative output.

9 DISCUSION

In the previous sections, we analyzed the application of foundational models to automate narrative visualization using the defined reference model. Having identified future research directions for each task, we now present a comprehensive analysis of prospective research avenues for the overall process.

Crafting Narrative Visualizations with Agent-based Workflow. Recent advancements in foundation models have facilitated the development of autonomous agent-based systems that simplify the data storytelling workflow. In this framework, foundation models serve as the cognitive core of agents, defining distinct agent roles and allowing them to perceive, decide, collaborate and act. This automation of tasks enhances the efficiency of narrative visualization creation [98] and positions multi-agent systems for narrative visualization as a promising avenue for future studies.

Agent Role Definition and Task Allocation. Foundation model serves as the brain of autonomous agent-based systems, facilitating the assignment of specific roles to manage various aspects of the narrative visualization workflow. Based on specific tasks drawing from our reference model, foundation models could design efficient systems that decompose the narrative process into manageable sub-tasks, and assign appropriate roles to agents specialized in

handling particular aspects of these tasks [98]. For example, in an automated data insight and narrative generation system, foundation models such as LLMs can assist in mimicking the human creative process for storytelling. They can establish a data analyst role to handle data exploration, alongside roles such as scripter and reviewer for story production, They can establish a data analyst role to handle data exploration, along with roles such as scripter and reviewer for story production, a concept similar to the discussion of the DataNarrative [111] system in the **Script**. Finally, an editor role can be designated for the presentation of the story. For a visualization system, we can assign a designer agent that utilizes image-based models or multimodal models to create visual representations, ensuring that the narrative remains compelling and that the visualizations are consistent with the story.

Information Flow and Collaboration between Agents. Effective communication and collaboration between agents are vital for maintaining a seamless crafting workflow. Foundation models can ensure that each agent converts input data into a format that can be easily processed by the models themselves, thereby strengthening the contextual foundation for subsequent tasks. Additionally, as we discussed in the **Logic**, prompt engineering for foundation models is particularly important, as it is essential to define the input-output formats and the chains of thought for each agent to ensure a comprehensive workflow.

Affective Narrative Visualization Design. Substantial research demonstrates that emotionally rich designs can significantly influence narrative visualization [150]. Recent studies indicate that foundation models are now capable of analyzing and generating emotional content [151]. Therefore, future research should focus further on enhancing the affective computing capabilities within foundation models and exploring novel applications that fully leverage the role of emotions in storytelling. Furthermore, by integrating foundation models with existing design spaces, researchers can create more engaging and emotionally compelling narrative visualization. Based on these insights, future research directions should address the following areas:

Personalized Emotional Experience. Designing systems that can tailor the emotional tone of narratives based on individual user preferences and profiles is a promising direction for enhancing personalized emotional experiences [152]. To craft a personalized visualization, user modeling and adaptive algorithms are necessary. As discussed in the **Navigation**, foundation models have the ability to customize data views according to the unique interests and needs of users. Through prompt engineering, foundation models adjust the narrative visualization based on user-defined parameters. For example, if a user prefers uplifting and motivational stories, the system can prioritize content with positive emotional tones. In contrast, for users who enjoy dramatic and intense narratives, the system can introduce elements that evoke stronger emotional responses. This personalized approach enhances the relevance and emotional impact of the narrative, making it more engaging for each user.

Emotion-Driven Storytelling. Understanding emotions in narratives to make models more human-like is a crucial direction for future research on foundation models. By improving the model's ability to comprehend and convey emotions, the generated narratives can become more vivid and impactful. Particularly, research can focus on collecting emotionally rich datasets to fine-tune foundation models [153], [154], or adjusting existing model architectures or prompts to help models accept and analyze emotional input, thus enhancing the naturalness of emotional expression in storytelling [124].

Interaction in Narrative Visualization. Interaction plays a crucial role in narrative visualization by bridging the gap between data representations and user experiences. Although tools like Tableau and APIs such as d3.js can create interactive visualizations, these efforts often fail to effectively facilitate narrative communication. Moreover, constructing interactive elements, even for simple charts, can be complex and time-consuming [155]. Therefore, leveraging foundation models to simplify and enhance the crafting of interaction in narrative visualization is a key focus for future research. The following directions merit exploration:

Enhancing User Experience through the Natural Language-based Interaction. In the previous discussion on various tasks involved in creating narrative visualizations, such as **Data Exploration, Generation**, and **Chart Understanding**, we emphasize that natural language-based interaction eliminates the need for users to master specialized technical skills. Users can simply pose questions or commands in natural language, and the models can accurately interpret and respond. This simplified interaction reduces the learning curve, making the creation of narrative visualizations more accessible to a broader audience. The natural language processing capabilities of foundation models enable them to comprehend and generate a wide range of semantic information, allowing them to handle complex user commands and natural language queries. Foundation models can not only understand the user's intent but also generate precise responses and recommendations tailored to specific needs [7].

Enhancing User Interaction through Exploratory Interaction. Some tools discussed in the **Generation** usually pre-determine a template, regardless of user input, which results in a limited variety of outcomes. While this end-to-end approach offers simplicity and clarity in certain applications, it restricts user interaction and limits the diversity of the narrative. As the application of foundation models in visualization deepens, there is growing interest in how user interaction with these models can lead to diverse narrative paths and multiple visualization forms.

Future research could prioritize the design and implementation of adaptive narrative structures. These structures would lead to different perspectives and outcomes with different user interactions, thereby improving the complexity and richness of narrative visualizations. This approach presents not only a technical challenge, but also a deep understanding of user behavior and cognitive processes to ensure that the generated narratives meet user expectations and effectively convey critical information.

10 CONCLUSION

In this study, we conducted a systematic review of 77 papers to study how foundation models progressively engage in narrative visualization. We propose a reference model that leverages foundation models for crafting narrative visualizations. In particular, we have identified four stages of narrative visualization based on previous research, and nine tasks for characterizing the state-of-the-art techniques. With the survey, we connect prior studies by integrating them into our proposed reference model. Furthermore, this paper explores the relationship between foundation models and narrative visualization, highlighting existing challenges and promising directions for future research. We hope our work could guide researchers in making informed decisions about effectively incorporating these advanced models into their work, thereby enhancing the quality and effectiveness of narrative visualization.

ACKNOWLEDGMENTS

Nan Cao is the corresponding author. This work was supported in part by NSFC 62072338, 62372327, NSF Shanghai 23ZR1464700, and China Postdoctoral Science Foundation (2023M732674). We would like to thank all the reviewers for their valuable feedback.

REFERENCES

[1] Jessica Hullman and Nick Diakopoulos. Visualization rhetoric: Framing effects in narrative visualization. IEEE Transactions on Visualization and Computer Graphics, 17(12):2231–2240, 2011.

[2] Danqing Shi, Xinyue Xu, Fuling Sun, Yang Shi, and Nan Cao. Calliope: Automatic visual data story generation from a spreadsheet. IEEE Transactions on Visualization and Computer Graphics, 27(2):453–463, 2020.

[3] Mengdi Sun, Ligan Cai, Weiwei Cui, Yanqiu Wu, Yang Shi, and Nan Cao. Erato: Cooperative data story editing via fact interpolation. IEEE Transactions on Visualization and Computer Graphics, 29(1):983–993, 2022.

[4] Sha Yuan, Hanyu Zhao, Shuai Zhao, Jiahong Leng, Yangxiao Liang, Xiaozhi Wang, Jifan Yu, Xin Lv, Zhou Shao, Jiaao He, et al. A roadmap for big model. arXiv preprint arXiv:2203.14101, 2022.

[5] Wahab Khan, Ali Daud, Khairullah Khan, Shakoor Muhammad, and Rafiul Haq. Exploring the frontiers of deep learning and natural language processing: A comprehensive overview of key challenges and emerging trends. Natural Language Processing Journal, 4:100026, 2023.

[6] Victor Dibia. LIDA: A tool for automatic generation of grammar-agnostic visualizations and infographics using large language models. In Danushka Bollegala, Ruihong Huang, and Alan Ritter, editors, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 113–126, Toronto, Canada, July 2023. Association for Computational Linguistics.

[7] Zhutian Chen, Chenyang Zhang, Qianwen Wang, Jakob Troidl, Simon Warchol, Johanna Beyer, Nils Gehlenborg, and Hanspeter Pfister. Beyond generating code: Evaluating gpt on a data visualization course. In 2023 IEEE VIS Workshop on Visualization Education, Literacy, and Activities (EduVis), pages 16–21, 2023.

[8] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. Evaluating large language models in generating synthetic hci research data: a case study. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pages 1–19, 2023.

[9] Mikke Tavast, Anton Kunnari, and Perttu Hämäläinen. Language models can generate human-like self-reports of emotion. In 27th International Conference on Intelligent User Interfaces, pages 69–72, 2022.

[10] Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators. arXiv preprint arXiv:2210.06280, 2022.

[11] Ryan Lingo. The role of chatgpt in democratizing data science: An exploration of ai-facilitated data analysis in telematics. arXiv preprint arXiv:2308.02045, 2023.

[12] Xiaoyu Zhang, Jianping Li, Po-Wei Chi, Senthil Chandrasegaran, and Kwan-Liu Ma. Concepteva: Concept-based interactive exploration and customization of document summaries. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pages 1–16, 2023.

[13] Akshat Shrivastava and Jeffrey Heer. Iseql: Interactive sequence learning. In Proceedings of the 25th International Conference on Intelligent User Interfaces, pages 43–54, 2020.

[14] Yamei Tu, Rui Qiu, Yu-Shuen Wang, Po-Yin Yen, and Han-Wei Shen. Phrasemap: Attention-based keyphrases recommendation for information seeking. IEEE Transactions on Visualization and Computer Graphics, pages 1–15, 2022.

[15] Yamei Tu, Olga Li, Junpeng Wang, Han-Wei Shen, Przemek Powałko, Irina Tomescu-Dubrow, Kazimierz M. Slomczynski, Spyros Blanas, and J. Craig Jenkins. Sdrquerier: A visual querying framework for cross-national survey data recycling. IEEE Transactions on Visualization and Computer Graphics, 29(6):2862–2874, 2023.

[16] Jinglong Gao, Xiao Ding, Bing Qin, and Ting Liu. Is ChatGPT a good causal reasoner? a comprehensive evaluation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, Findings of the Association for Computational Linguistics: EMNLP 2023, pages 11111–11126, Singapore, December 2023. Association for Computational Linguistics.

[17] Varun Gangal, Steven Y. Feng, Malihe Alikhani, Teruko Mitamura, and Eduard Hovy. Nareor: The narrative reordering problem. Proceedings of the AAAI Conference on Artificial Intelligence, 36(10):10645–10653, Jun. 2022.

[18] Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. arXiv preprint arXiv:2305.00050, 2023.

[19] Fengjie Wang, Xuye Liu, Oujing Liu, Ali Neshati, Tengfei Ma, Min Zhu, and Jian Zhao. Slide4n: Creating presentation slides from computational notebooks with human-ai collaboration. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[20] Shreya Ghosh, Saptarshi Sengupta, and Prasenjit Mitra. Spatio-temporal storytelling? leveraging generative models for semantic trajectory analysis. arXiv preprint arXiv:2306.13905, 2023.

[21] Benny Tang, Angie Boggust, and Arvind Satyanarayan. VisText: A benchmark for semantically rich chart captioning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7268–7298, Toronto, Canada, July 2023. Association for Computational Linguistics.

[22] Nina Shvetsova, Anna Kukleva, Xudong Hong, Christian Rupprecht, Bernt Schiele, and Hilde Kuehne. Howtocaption: Prompting llms to transform video annotations at scale. arXiv preprint arXiv:2310.04900, 2023.

[23] Jiwon Choi and Jaemin Jo. Intentable: A mixed-initiative system for intent-based chart captioning. In 2022 IEEE Visualization and Visual Analytics (VIS), pages 40–44, 2022.

[24] Ashley Liew and Klaus Mueller. Using large language models to generate engaging captions for data visualizations. arXiv preprint arXiv:2212.14047, 2022.

[25] Nicole Sultanum and Arjun Srinivasan. Datatales: Investigating the use of large language models for authoring data-driven articles. In 2023 IEEE Visualization and Visual Analytics (VIS), pages 231–235, 2023.

[26] Dennis Bromley and Vidya Setlur. What is the difference between a mountain and a molehill? quantifying semantic labeling of visual features in line charts. In 2023 IEEE Visualization and Visual Analytics (VIS), pages 161–165, 2023.

[27] Paula Maddigan and Teo Susnjak. Chat2vis: Generating data visualisations via natural language using chatgpt, codex and gpt-3 large language models. IEEE Access, 2023.

[28] Paula Maddigan and Teo Susnjak. Chat2vis: Fine-tuning data visualisations using multilingual natural language text and pre-trained large language models. arXiv preprint arXiv:2303.14292, 2023.

[29] Lei Wang, Songheng Zhang, Yun Wang, Ee-Peng Lim, and Yong Wang. LLM4Vis: Explainable visualization recommendation using ChatGPT. In Mingxuan Wang and Imed Zitouni, editors, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 675–692, Singapore, December 2023. Association for Computational Linguistics.

[30] Andreas Stöckl. Natural language interface for data visualization with deep learning based language models. In 2022 26th International Conference Information Visualisation (IV), pages 142–148, 2022.

[31] M. Hutchinson, A. Slingsby, R. Jianu, and P. Madhyastha. Towards visualisation specifications from multilingual natural language queries using large language models. In EuroVis 2023 - Posters, pages 77–79, Eindhoven, The Netherlands, June 2023. Eurographics.

[32] Ecem Kavaz, Anna Puig, and Inmaculada Rodríguez. Chatbot-based natural language interfaces for data visualisation: A scoping review. Applied Sciences, 13(12):7025, 2023.

[33] Sangho Suh and Pengcheng An. Leveraging generative conversational ai to develop a creative learning environment for computational thinking. In 27th International Conference on Intelligent User Interfaces, IUI '22 Companion, page 73–76, New York, NY, USA, 2022. Association for Computing Machinery.

[34] Victor Schetinger, Sara Di Bartolomeo, Edirlei Soares de Lima, Christofer Meinecke, and Rudolf Rosa. *n* walks in the fictional woods. arXiv preprint arXiv:2308.06266, 2023.

[35] Yuqian Sun, Ying Xu, Chenhang Cheng, Yihua Li, Chang Hee Lee, and Ali Asadipour. Explore the future earth with wander 2.0: AI Chatbot driven by knowledge-base story generation and text-to-image model. In Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems, CHI EA '23. Association for Computing Machinery, 2023.

[36] Fangyu Liu, Julian Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. Deplot: One-shot visual language reasoning by plot-

to-table translation. In Findings of the Association for Computational Linguistics: ACL 2023, pages 10381–10399, 2023.

[37] Mingyang Zhou, Yi Fung, Long Chen, Christopher Thomas, Heng Ji, and Shih-Fu Chang. Enhanced chart understanding via visual language pre-training on plot table pairs. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, Findings of the Association for Computational Linguistics: ACL 2023, pages 1314–1326, Toronto, Canada, July 2023. Association for Computational Linguistics.

[38] Yanna Lin, Haotian Li, Leni Yang, Aoyu Wu, and Huamin Qu. Inksight: Leveraging sketch interaction for documenting chart findings in computational notebooks. IEEE Transactions on Visualization and Computer Graphics, pages 1–11, 2023.

[39] Raian Rahman, Rizvi Hasan, Abdullah Al Farhad, Md Tahmid Rahman Laskar, Md Hamjajul Ashmafee, and Abu Raihan Mostofa Kamal. Chartsumm: A comprehensive benchmark for automatic chart summarization of long and short summaries. arXiv preprint arXiv:2304.13620, 2023.

[40] Shankar Kantharaj, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. Chart-to-text: A large-scale benchmark for chart summarization. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4005–4023, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[41] Zhi-Qi Cheng, Qi Dai, and Alexander G. Hauptmann. Chartreader: A unified framework for chart derendering and comprehension without heuristic rules. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 22202–22213, October 2023.

[42] Shankar Kantharaj, Xuan Long Do, Rixie Tiffany Leong, Jia Qing Tan, Enamul Hoque, and Shafiq Joty. OpenCQA: Open-ended question answering with charts. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 11817–11837, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[43] Qianwen Wang, Zhutian Chen, Yong Wang, and Huamin Qu. A survey on ML4VIS: Applying machine learning advances to data visualization. IEEE Transactions on Visualization and Computer Graphics, 28(12):5134–5153, 2021.

[44] Aoyu Wu, Yun Wang, Xinhuan Shu, Dominik Moritz, Weiwei Cui, Haidong Zhang, Dongmei Zhang, and Huamin Qu. AI4VIS: Survey on artificial intelligence approaches for data visualization. IEEE Transactions on Visualization and Computer Graphics, 28(12):5049–5070, 2022.

[45] V. Schetinger, S. Di Bartolomeo, M. El-Assady, A. McNutt, M. Miller, J. P. A. Passos, and J. L. Adams. Doom or deliciousness: Challenges and opportunities for visualization in the age of generative models. Computer Graphics Forum, 42(3):423–435, 2023.

[46] Weikai Yang, Mengchen Liu, Zheng Wang, and Shixia Liu. Foundation models meet visualizations: Challenges and opportunities. Computational Visual Media, pages 1–26, 2024.

[47] Qing Chen, Shixiong Cao, Jiazhe Wang, and Nan Cao. How does automation shape the process of narrative visualization: A survey of tools. IEEE Transactions on Visualization and Computer Graphics, 2023.

[48] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258, 2021.

[49] Bongshin Lee, Nathalie Henry Riche, Petra Isenberg, and Sheelagh Carpendale. More than telling a story: Transforming data into visually shared stories. IEEE Computer Graphics and Applications, 35(5):84–90, 2015.

[50] Edward Segel and Jeffrey Heer. Narrative visualization: Telling stories with data. IEEE Transactions on Visualization and Computer Graphics, 16(6):1139–1148, 2010.

[51] Haotian Li, Yun Wang, and Huamin Qu. Where are we so far? understanding data storytelling tools from the perspective of human-ai collaboration. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[52] Fereshteh Amini, Nathalie Henry Riche, Bongshin Lee, Christophe Hurter, and Pourang Irani. Understanding data videos: Looking at narrative visualization through the cinematography lens. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI '15, page 1459–1468, New York, NY, USA, 2015. Association for Computing Machinery.

[53] Yi Guo, Nan Cao, Xiaoyu Qi, Haoyang Li, Danqing Shi, Jing Zhang, Qing Chen, and Daniel Weiskopf. Urania: Visualizing data analysis pipelines for natural language-based data exploration. arXiv preprint arXiv:2306.07760, 2023.

[54] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. Talebrush: Sketching stories with generative pretrained language models. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, pages 1–19, 2022.

[55] Can Liu, Yuhan Guo, and Xiaoru Yuan. Autotitle: An interactive title generator for visualizations. IEEE Transactions on Visualization and Computer Graphics, pages 1–12, 2023.

[56] Shishi Xiao, Suizi Huang, Yue Lin, Yilin Ye, and Wei Zeng. Let the chart spark: Embedding semantic context into chart with text-to-image generative model. IEEE Transactions on Visualization and Computer Graphics, 30(1):284–294, 2024.

[57] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009.

[58] Katikapalli Subramanyam Kalyan. A survey of gpt-3 family large language models including chatgpt and gpt-4. Natural Language Processing Journal, page 100048, 2023.

[59] Som Biswas. The function of chat GPT in social media: According to chat GPT. Available at SSRN 4405389, 2023.

[60] David E Losada, David Elsweiler, Morgan Harvey, and Christoph Trattner. A day at the races: using best arm identification algorithms to reduce the cost of information retrieval user studies. Applied Intelligence, 52(5):5617–5632, 2022.

[61] Selina Meyer, David Elsweiler, Bernd Ludwig, Marcos Fernandez-Pichel, and David E. Losada. Do we still need human assessors? prompt-based gpt-3 user simulation in conversational ai. In Proceedings of the 4th Conference on Conversational User Interfaces, CUI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[62] Ashish Mishra, Gyanaranjan Nayak, Suparna Bhattacharya, Tarun Kumar, Arpit Shah, and Martin Foltin. Llm-guided counterfactual data generation for fairer ai. In Companion Proceedings of the ACM on Web Conference 2024, WWW '24, page 1538–1545, New York, NY, USA, 2024. Association for Computing Machinery.

[63] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 9459–9474. Curran Associates, Inc., 2020.

[64] Angelica Lo Duca. Using retrieval augmented generation to build the context for data-driven stories. In VISIGRAPP (1): GRAPP, HUCAPP, IVAPP, pages 690–696, 2024.

[65] Vipula Rawte, Amit Sheth, and Amitava Das. A survey of hallucination in large foundation models. arXiv preprint arXiv:2309.05922, 2023.

[66] Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. arXiv preprint arXiv:2402.00367, 2024.

[67] Brett A. Halperin and Stephanie M Lukin. Artificial dreams: Surreal visual storytelling as inquiry into ai 'hallucination'. In Proceedings of the 2024 ACM Designing Interactive Systems Conference, DIS '24, page 619–637, New York, NY, USA, 2024. Association for Computing Machinery.

[68] Hua Guo, Steven R. Gomez, Caroline Ziemkiewicz, and David H. Laidlaw. A case study using visualization interaction logs and insight metrics to understand how analysts arrive at insights. IEEE Transactions on Visualization and Computer Graphics, 22(1):51–60, 2016.

[69] Brian Felipe Keith Norambuena, Tanushree Mitra, and Chris North. Mixed multi-model semantic interaction for graph-based narrative visualizations. In Proceedings of the 28th International Conference on Intelligent User Interfaces, IUI '23, page 866–888, New York, NY, USA, 2023. Association for Computing Machinery.

[70] Can Liu, Yun Han, Ruike Jiang, and Xiaoru Yuan. Advisor: Automatic visualization answer for natural-language question on tabular data. In 2021 IEEE 14th Pacific Visualization Symposium (PacificVis), pages 11–20. IEEE, 2021.

[71] Qing Chen, Ying Chen, Ruishi Zou, Wei Shuai, Yi Guo, Jiazhe Wang, and Nan Cao. Chart2vec: A universal embedding of context-aware visualizations. IEEE Transactions on Visualization and Computer Graphics, pages 1–12, 2024.

[72] Nan Cao, Weiwei Cui, Nan Cao, and Weiwei Cui. Overview of text visualization techniques. Introduction to Text Visualization, pages 11–40, 2016.

[73] Fariba Lotfi, Amin Beheshti, Helia Farhood, Matineh Pooshideh, Mansour Jamzad, and Hamid Beigy. Storytelling with image data: A systematic review and comparative analysis of methods and tools. Algorithms, 16(3):135, 2023.

[74] Qingsong Wen, Liang Sun, Fan Yang, Xiaomin Song, Jingkun Gao, Xue Wang, and Huan Xu. Time series data augmentation for deep learning: A survey. In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, 2021.

[75] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[76] Justyna Sarzynska-Wawer, Aleksander Wawer, Aleksandra Pawlak, Julia Szymanowska, Izabela Stefaniak, Michal Jarkiewicz, and Lukasz Okruszek. Detecting formal thought disorder by deep contextualized word representations. Psychiatry Research, 304:114135, 2021.

[77] Adyasha Maharana, Darryl Hannan, and Mohit Bansal. Storydall-e: Adapting pretrained text-to-image transformers for story continuation. In European Conference on Computer Vision, pages 70–87. Springer, 2022.

[78] Bonka Kotseva, Irene Vianini, Nikolaos Nikolaidis, Nicolò Faggiani, Kristina Potapova, Caroline Gasparro, Yaniv Steiner, Jessica Scornavacche, Guillaume Jacquet, Vlad Dragu, et al. Trend analysis of covid-19 mis/disinformation narratives–a 3-year study. Plos one, 18(11):e0291423, 2023.

[79] Yizhang Zhu, Shiyin Du, Boyan Li, Yuyu Luo, and Nan Tang. Are large language models good statisticians? arXiv preprint arXiv:2406.07815, 2024.

[80] Brenda Santana, Ricardo Campos, Evelin Amorim, Alípio Jorge, Purificação Silvano, and Sérgio Nunes. A survey on narrative extraction from textual data. Artificial Intelligence Review, pages 1–43, 2023.

[81] Richard Brath. Summarizing text to embed qualitative data into visualizations. arXiv preprint arXiv:2209.15041, 2022.

[82] Tong Gao, Jessica R. Hullman, Eytan Adar, Brent Hecht, and Nicholas Diakopoulos. Newsviews: an automated pipeline for creating custom geovisualizations for news. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '14, page 3005–3014, New York, NY, USA, 2014. Association for Computing Machinery.

[83] Chuan Lei, Fatma Özcan, Abdul Quamar, Ashish R Mittal, Jaydeep Sen, Diptikalyan Saha, and Karthik Sankaranarayanan. Ontology-based natural language query interfaces for data exploration. IEEE Data Eng. Bull., 41(3):52–63, 2018.

[84] Wenqi Zhang, Yongliang Shen, Weiming Lu, and Yueting Zhuang. Data-copilot: Bridging billions of data and humans with autonomous workflow. arXiv preprint arXiv:2306.07209, 2023.

[85] Liwenhan Xie, Chengbo Zheng, Haijun Xia, Huamin Qu, and Chen Zhu-Tian. Waitgpt: Monitoring and steering conversational llm agent in data analysis with on-the-fly code visualization. arXiv preprint arXiv:2408.01703, 2024.

[86] Hongxin Li, Jingran Su, Yuntao Chen, Qing Li, and ZHAO-XIANG ZHANG. Sheetcopilot: Bringing software productivity to the next level through large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, Advances in Neural Information Processing Systems, volume 36, pages 4952–4984. Curran Associates, Inc., 2023.

[87] Liangyu Zha, Junlin Zhou, Liyao Li, Rui Wang, Qingyi Huang, Saisai Yang, Jing Yuan, Changbao Su, Xiang Li, Aofeng Su, et al. Tablegpt: Towards unifying tables, nature language and commands into one gpt. arXiv preprint arXiv:2307.08674, 2023.

[88] N. Errey, J. Liang, T. W. Leong, et al. Evaluating narrative visualization: a survey of practitioners. International Journal of Data Science and Analytics, pages 1–16, 2023.

[89] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery.

[90] S. McKenna, N. Henry Riche, B. Lee, J. Boy, and M. Meyer. Visual narrative flow: Exploring factors shaping data visualization story reading experiences. Computer Graphics Forum, 36(3):377–387, 2017.

[91] Humphrey O. Obie, Caslon Chua, Iman Avazpour, Mohamed Abdelrazek, John Grundy, and Tomasz Bednarz. A study of the effects of narration on comprehension and memorability of visualisations. Journal of Computer Languages, 52:113–124, 2019.

[92] Daniel Pietschmann, Sabine Völkel, and Peter Ohler. Transmedia critical— limitations of transmedia storytelling for children: A cognitive developmental analysis. International Journal of Communication, 8:22, 2014.

[93] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. OpenAI, 2018.

[94] Florian Wolf and Edward Gibson. Representing Discourse Coherence: A Corpus-Based Study. Computational Linguistics, 31(2):249–287, 06 2005.

[95] Qianwen Wang, Zhen Li, Siwei Fu, Weiwei Cui, and Huamin Qu. Narvis: Authoring narrative slideshows for introducing data visualization designs. IEEE Transactions on Visualization and Computer Graphics, 25(1):779–788, 2018.

[96] Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, Findings of the Association for Computational Linguistics: ACL 2023, pages 1049–1065, Toronto, Canada, July 2023. Association for Computational Linguistics.

[97] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. arXiv preprint arXiv:2309.07864, 2023.

[98] Leixian Shen, Haotian Li, Yun Wang, and Huamin Qu. From data to story: Towards automatic animated data video creation with llm-based multi-agent systems. In 2024 IEEE VIS Workshop on Data Storytelling in an Era of Generative AI (GEN4DS), pages 20–27, 2024.

[99] Xingyu Lan, Xinyue Xu, and Nan Cao. Understanding narrative linearity for telling expressive time-oriented stories. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[100] Nam Wook Kim, Benjamin Bach, Hyejin Im, Sasha Schriber, Markus Gross, and Hanspeter Pfister. Visualizing nonlinear narratives with story curves. IEEE Transactions on Visualization and Computer Graphics, 24(1):595–604, 2018.

[101] Leni Yang, Xian Xu, XingYu Lan, Ziyan Liu, Shunan Guo, Yang Shi, Huamin Qu, and Nan Cao. A design space for applying the freytag's pyramid structure to data stories. IEEE Transactions on Visualization and Computer Graphics, 28(1):922–932, 2021.

[102] Dinesh D. Patil, Dhanraj R. Dhotre, Gopal S. Gawande, Dipali S. Mate, Mayura V. Shelke, and Tejaswini S. Bhoye. Transformative trends in generative ai: Harnessing large language models for natural language understanding and generation. International Journal of Intelligent Systems and Applications in Engineering, 12(4s):309–319, 2023.

[103] Ha-Kyung Kong, Zhicheng Liu, and Karrie Karahalios. Trust and recall of information across varying degrees of title-visualization misalignment. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19, page 1–13, New York, NY, USA, 2019. Association for Computing Machinery.

[104] Shuang Bai and Shan An. A survey on automatic image caption generation. Neurocomputing, 311:291–304, 2018.

[105] Donghao Ren, Matthew Brehmer, Bongshin Lee, Tobias Höllerer, and Eun Kyoung Choe. Chartaccent: Annotation for data-driven storytelling. In 2017 IEEE Pacific Visualization Symposium (PacificVis), pages 230–239, 2017.

[106] Dennis Bromley and Vidya Setlur. Dash: A bimodal data exploration tool for interactive text and visualizations. In 2024 IEEE Visualization and Visual Analytics (VIS), pages 256–260, 2024.

[107] Leixian Shen, Haotian Li, Yun Wang, Tianqi Luo, Yuyu Luo, and Huamin Qu. Data playwright: Authoring data videos with annotated narration. IEEE Transactions on Visualization and Computer Graphics, pages 1–14, 2024.

[108] Jessica Hullman, Nicholas Diakopoulos, and Eytan Adar. Contextifier: automatic generation of annotated stock visualizations. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '13, page 2707–2716, New York, NY, USA, 2013. Association for Computing Machinery.

[109] Moniba Keymanesh, Adrian Benton, and Mark Dredze. What makes data-to-text generation hard for pretrained language models? In Antoine Bosselut, Khyathi Chandu, Kaustubh Dhole, Varun Gangal, Sebastian Gehrmann, Yacine Jernite, Jekaterina Novikova, and Laura Perez-Beltrachini, editors, Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pages 539–554, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[110] Li Ye, Lei Wang, Shaolun Ruan, Yuwei Meng, Yigang Wang, Wei Chen, and Zhiguang Zhou. Storyexplorer: A visualization framework for story-

line generation of textual narratives. arXiv preprint arXiv:2411.05435, 2024.

[111] Mohammed Saidul Islam, Enamul Hoque, Shafiq Joty, Md Tahmid Rahman Laskar, and Md Rizwan Parvez. Datanarrative: Automated data-driven storytelling with visualizations and texts. arXiv preprint arXiv:2408.05346, 2024.

[112] Adriano Pinargote, Eddy Calderón, Kevin Cevallos, Gladys Carrillo, Katherine Chiluiza, and Vanessa Echeverría. Automating data narratives in learning analytics dashboards using genai. In LAK Workshops, pages 150–161, 2024.

[113] Pere-Pau Vázquez. Are llms ready for visualization? In 2024 IEEE 17th Pacific Visualization Conference (PacificVis), pages 343–352. IEEE, 2024.

[114] Cole Beasley and Azza Abouzied. Pipe (line) dreams: Fully automated end-to-end analysis and visualization. In Proceedings of the 2024 Workshop on Human-In-the-Loop Data Analytics, pages 1–7, 2024.

[115] Yuan Tian, Weiwei Cui, Dazhen Deng, Xinjing Yi, Yurun Yang, Haidong Zhang, and Yingcai Wu. Chartgpt: Leveraging llms to generate charts from abstract natural language. IEEE Transactions on Visualization and Computer Graphics, pages 1–15, 2024.

[116] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 1(2):3, 2022.

[117] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In International Conference on Machine Learning, pages 8821–8831. PMLR, 2021.

[118] David Heidrich, Andreas Schreiber, and Sabine Theis. Generative artificial intelligence for the visualization of source code as comics. In International Conference on Human-Computer Interaction, pages 35–49. Springer, 2024.

[119] Mikaela E Milesi, Riordan Alfredo, Vanessa Echeverria, Lixiang Yan, Linxuan Zhao, Yi-Shan Tsai, and Roberto Martinez-Maldonado. "it's really enjoyable to see me solve the problem like a hero": Genai-enhanced data comics as a learning analytics tool. In Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems, CHI EA '24, New York, NY, USA, 2024. Association for Computing Machinery.

[120] Tongyu Zhou, Jeff Huang, and Gromit Yeuk-Yin Chan. Epigraphics: Message-driven infographics authoring. In Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[121] Jianing Hao, Manling Yang, Qing Shi, Yuzhe Jiang, Guang Zhang, and Wei Zeng. Finflier: Automating graphical overlays for financial visualizations with knowledge-grounding large language model. IEEE Transactions on Visualization and Computer Graphics, pages 1–17, 2024.

[122] Howard Beckerman. Animation: the whole story. Allworth, 2003.

[123] Lu Ying, Yun Wang, Haotian Li, Shuguang Dou, Haidong Zhang, Xinyang Jiang, Huamin Qu, and Yingcai Wu. Reviving static charts into live charts. IEEE Transactions on Visualization and Computer Graphics, pages 1–16, 2024.

[124] Leixian Shen, Yizhi Zhang, Haidong Zhang, and Yun Wang. Data player: Automatic generation of data videos with narration-animation interplay. IEEE Transactions on Visualization and Computer Graphics, 2023.

[125] Vivian Liu, Tao Long, Nathan Raw, and Lydia Chilton. Generative disco: Text-to-video generation for music visualization. arXiv preprint arXiv:2304.08551, 2023.

[126] Melissa Lin, Heer Patel, Medina Lamkin, Tukey Tu, Hannah Bako, Soham Raut, and Leilani Battle. How do observable users decompose d3 code? an exploratory study. arXiv preprint arXiv:2405.14341, 2024.

[127] Diansheng Guo. Flow mapping and multivariate visualization of large spatial interaction data. IEEE Transactions on Visualization and Computer Graphics, 15(6):1041–1048, 2009.

[128] Qing Chen, Fuling Sun, Xinyue Xu, Zui Chen, Jiazhe Wang, and Nan Cao. Vizlinter: A linter and fixer framework for data visualization. IEEE Transactions on Visualization and Computer Graphics, 28(1):206–216, 2022.

[129] Leo Yu-Ho Lo and Huamin Qu. How good (or bad) are llms at detecting misleading visualizations? IEEE Transactions on Visualization and Computer Graphics, 31(1):1116–1125, 2025.

[130] Yang Shi, Pei Liu, Siji Chen, Mengdi Sun, and Nan Cao. Supporting expressive and faithful pictorial visualization design with visual style transfer. IEEE Transactions on Visualization and Computer Graphics, 29(1):236–246, 2023.

[131] Jiaqi Wu, John Joon Young Chung, and Eytan Adar. viz2viz: Prompt-driven stylized visualization generation using a diffusion model. arXiv preprint arXiv:2304.01919, 2023.

[132] Min Lu, Chufeng Wang, Joel Lanir, Nanxuan Zhao, Hanspeter Pfister, Daniel Cohen-Or, and Hui Huang. Exploring visual information flows in infographics. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20, page 1–12, New York, NY, USA, 2020. Association for Computing Machinery.

[133] Pengyu Yan, Mahesh Bhosale, Jay Lal, Bikhyat Adhikari, and David Doermann. Chartreformer: Natural language-driven chart image editing. arXiv preprint arXiv:2403.00209, 2024.

[134] Priyan Vaithilingam, Elena L. Glassman, Jeevana Priya Inala, and Chenglong Wang. Dynavis: Dynamically synthesized ui widgets for visualization editing. In Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[135] Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. UniChart: A universal vision-language pretrained model for chart comprehension and reasoning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 14662–14684, Singapore, December 2023. Association for Computational Linguistics.

[136] Hrituraj Singh and Sumit Shekhar. STL-CQA: Structure-based transformers with localization and encoding for chart question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 3275–3284, Online, November 2020. Association for Computational Linguistics.

[137] Shishi Xiao, Yihan Hou, Cheng Jin, and Wei Zeng. Wytiwyr: A user intent-aware framework with multi-modal inputs for visualization retrieval. Computer Graphics Forum, 42(3):311–322, 2023.

[138] Fanqing Meng, Wenqi Shao, Quanfeng Lu, Peng Gao, Kaipeng Zhang, Yu Qiao, and Ping Luo. Chartassisstant: A universal chart multimodal language model via chart-to-table pre-training and multitask instruction tuning. arXiv preprint arXiv:2401.02384, 2024.

[139] Junyu Luo, Zekun Li, Jinpeng Wang, and Chin-Yew Lin. Chartocr: Data extraction from charts images via a deep hybrid framework. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 1917–1925, January 2021.

[140] Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Eisenschlos. Matcha: Enhancing visual language pretraining with math reasoning and chart derendering. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12756–12770, 2023.

[141] Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. Chartllama: A multimodal llm for chart understanding and generation. arXiv preprint arXiv:2311.16483, 2023.

[142] Yuan Cui, Lily W. Ge, Yiren Ding, Lane Harrison, Fumeng Yang, and Matthew Kay. Promises and pitfalls: Using large language models to generate visualization items. IEEE Transactions on Visualization and Computer Graphics, 31(1):1094–1104, 2025.

[143] Naren Sivakumar, Lujie Karen Chen, Pravalika Papasani, Vigna Majmundar, Jinjuan Heidi Feng, Louise Yarnall, and Jiaqi Gong. Show and tell: Exploring large language model's potential in formative educational assessment of data stories. In 2024 IEEE VIS Workshop on Data Storytelling in an Era of Generative AI (GEN4DS), pages 13–19, 2024.

[144] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In Findings of the Association for Computational Linguistics: ACL 2022, pages 2263–2279, 2022.

[145] E. Hoque, P. Kavehzadeh, and A. Masry. Chart question answering: State of the art and future directions. Computer Graphics Forum, 41(3):555–572, 2022.

[146] Kiroong Choe, Chaerin Lee, Soohyun Lee, Jiwon Song, Aeri Cho, Nam Wook Kim, and Jinwook Seo. Enhancing data literacy on-demand: Llms as guides for novices in chart interpretation. IEEE Transactions on Visualization and Computer Graphics, pages 1–17, 2024.

[147] Lin Gao, Jing Lu, Zekai Shao, Ziyue Lin, Shengbin Yue, Chiokit Ieong, Yi Sun, Rory James Zauner, Zhongyu Wei, and Siming Chen. Fine-tuned large language model for visualization system: A study on self-regulated learning in education. arXiv preprint arXiv:2407.20570, 2024.

[148] Luca Podo, Muhammad Ishmal, and Marco Angelini. Vi (e) va llm! a conceptual stack for evaluating and interpreting generative ai-based visualizations. arXiv preprint arXiv:2402.02167, 2024.

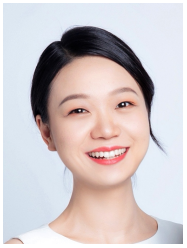
[149] Nan Chen, Yuge Zhang, Jiahang Xu, Kan Ren, and Yuqing Yang. Viseval: A benchmark for data visualization in the era of large language models. arXiv preprint arXiv:2407.00981, 2024.
[150] Xingyu Lan, Yanqiu Wu, and Nan Cao. Affective visualization design: Leveraging the emotional impact of data. IEEE Transactions on Visualization and Computer Graphics, 2023.
[151] Zixing Zhang, Liyizhe Peng, Tao Pang, Jing Han, Huan Zhao, and Björn W Schuller. Refashioning emotion recognition modelling: The advent of generalised large models. IEEE Transactions on Computational Social Systems, 2024.
[152] Oswald Barral, Sébastien Lallé, Grigorii Guz, Alireza Iranpour, and Cristina Conati. Eye-tracking to predict user cognitive abilities and performance for user-adaptive narrative visualizations. In Proceedings of the 2020 International Conference on Multimodal Interaction, ICMI '20, page 163–173, New York, NY, USA, 2020. Association for Computing Machinery.
[153] Jingyuan Yang, Jiawei Feng, and Hui Huang. Emogen: Emotional image content generation with text-to-image diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 6358–6368, June 2024.
[154] Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Danny Cohen-Or, and Hui Huang. Emoset: A large-scale visual emotion dataset with rich attributes. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 20383–20394, October 2023.
[155] Bongshin Lee, Rubaiat Habib Kazi, and Greg Smith. Sketchstory: Telling more engaging stories with data through freeform sketching. IEEE Transactions on Visualization and Computer Graphics, 19(12):2416–2425, 2013.

Yi He received her B.Eng degree from School of Digital Media & Design Arts, Beijing University of Posts and Telecommunications. She is currently working toward his Ph.D. degree as part of the Intelligent Big Data Visualization (iDV^x) Lab, Tongji University. Her research interests include data visualization and artificial intelligence.

Ke Xu is currently a tenure-track associate professor at Nanjing University. Prior to this, he was an scientist in the Data Intelligence Lab at HUAWEI Cloud and a postdoctoral research associate in the Visualization and Data Analytics Research Center (VIDA) at New York University (NYU). He obtained his Ph.D. in the Department of Electronic and Computer Engineering at the Hong Kong University of Science and Technology in 2019, and B.S. in Electronic Science and Engineering from Nanjing University, China in 2015. His research interests include data visualization, human-computer interaction, with focus on visual anomaly detection and explainable AI.

Shixiong Cao received his Master's degree in Design from Sangmyung University in South Korea in 2019, and subsequently obtained a Ph.D. degree from Sungkyunkwan University in South Korea in 2023. Currently, he works as a postdoctoral researcher at Tongji University, and his research interests include information design, narrative visualization design, and user experience design.

Yang Shi is an associate professor and doctoral supervisor in the College of Design and Innovation, Tongji University, and a member of the Visualization and Visual Analysis Committee of CSIG. She focuses on the intersection of computer and design, and her research interests are theoretical models for visualization design and intelligent design for visualization. She was awarded the second prize of the Natural Science Award of the Chinese Society of Graphics 2022, and was listed in the 2019 Forbes China 30 Under 30 (Science). Meanwhile, she served as the IEEE VIS2022-2023 Meetup Chair and IEEE PacificVis2020 Poster Chair.



Qing Chen received her B.Eng degree from the Department of Computer Science, Zhejiang University and her Ph.D. degree from the Department of Computer Science and Engineering, Hong Kong University of Science and Technology (HKUST). After receiving her PhD degree, she worked as a postdoc at Inria and Ecole Polytechnique. She is currently an associate professor at Tongji University. Her research interests include information visualization, visual analytics, human-computer interaction, generative AI and their applications in education, healthcare, design, and business intelligence.

Nan Cao received his Ph.D. degree in Computer Science and Engineering from the Hong Kong University of Science and Technology (HKUST), Hong Kong, China in 2012. He is currently a professor at Tongji University and the Assistant Dean of the Tongji College of Design and Innovation. He also directs the Tongji Intelligent Big Data Visualization Lab (iDV^x Lab) and conducts interdisciplinary research across multiple fields, including data visualization, human computer interaction, machine learning, and data mining. He was a research staff member at the IBM T.J. Watson Research Center, New York, NY, USA before joining the Tongji faculty in 2016.