Computational Visual Media
https://doi.org/10.1007/s41095-023-0393-x Vol. 10, No. 3, June 2024, 399–424

Review Article

Foundation models meet visualizations: Challenges and opportunities

Weikai Yang[1], Mengchen Liu[2], Zheng Wang[1], and Shixia Liu[1] (✉)

Abstract Recent studies have indicated that foundation models, such as BERT and GPT, excel at adapting to various downstream tasks. This adaptability has made them a dominant force in building artificial intelligence (AI) systems. Moreover, a new research paradigm has emerged as visualization techniques are incorporated into these models. This study divides these intersections into two research areas: visualization for foundation model (VIS4FM) and foundation model for visualization (FM4VIS). In terms of VIS4FM, we explore the primary role of visualizations in understanding, refining, and evaluating these intricate foundation models. VIS4FM addresses the pressing need for transparency, explainability, fairness, and robustness. Conversely, in terms of FM4VIS, we highlight how foundation models can be used to advance the visualization field itself. The intersection of foundation models with visualizations is promising but also introduces a set of challenges. By highlighting these challenges and promising opportunities, this study aims to provide a starting point for the continued exploration of this research avenue.

Keywords visualization; artificial intelligence (AI); machine learning; foundation models; visualization for foundation model (VIS4FM); foundation model for visualization (FM4VIS)

1 Introduction

A foundation model is a large-scale machine learning model trained on a large amount of data across different domains, generally using self-supervision [1]. Notable examples include bidirectional encoder representations from transformers (BERT) [2] for natural language processing, VisionTransformer [3] and InternImage [4] for computer vision, Contrastive Language-Image Pretraining (CLIP) [5] for cross-modal learning, and the generative pre-trained transformer (GPT) series models [6–8] for text generation. Unlike traditional machine learning models, foundation models typically possess parameters ranging from hundreds of millions to billions and require extensive training on vast datasets over several weeks or months. These immense scales of parameters and training data enable foundation models to capture general knowledge regarding the world and serve as a "foundation" to effectively adapt to various downstream tasks such as information extraction, object recognition, image captioning, and instruction following [1]. To illustrate this, consider a BERT model. After pretraining on a substantial text corpus to predict randomly masked words, the BERT model acquires a foundational understanding of natural language. This enables the model to rapidly adapt to various natural language processing tasks, such as text classification, sentiment analysis, and question answering. Such tasks often require minimal task-specific finetuning. Owing to the adaptability of foundation models, they have become a leading force in shaping the creation of versatile, high-performance artificial intelligence (AI) systems across multiple applications. A recent OpenAI report indicated that approximately 19% of jobs have undergone considerable changes, and at least 50% of the tasks were affected by these models [9].

In this era of big data and AI, the need to visualize large-scale datasets and machine learning models has been increasingly observed for

1 School of Software, Tsinghua University, Beijing 100084, China. E-mail: W. Yang, yangwk21@mails.tsinghua.edu.cn; Z. Wang, zheng-wa19@mails.tsinghua.edu.cn; S. Liu, shixia@tsinghua.edu.cn (✉).

2 Microsoft, Redmond 98052, USA. E-mail: mengcliu@microsoft.com.

Manuscript received: 2023-10-04; accepted: 2023-11-15

efficient analyses. Recent studies have indicated that incorporating humans into the analysis process can make visualization techniques a critical bridge for the human comprehension of complex models [10–17]. This enhanced human–AI collaboration facilitates effective insight communication, informed decision-making, and improved AI trustworthiness. A new research paradigm has emerged from incorporating both visualization techniques and foundation models. Figure 1 shows the two promising research areas that arise from this paradigm: visualization for foundation model (VIS4FM) and foundation model for visualization (FM4VIS). In VIS4FM, visualization is an indispensable mechanism for facilitating the understanding, analysis, and refinement of foundation models. Conversely, FM4VIS focuses on how foundation models can be employed to improve visualization techniques by adapting them to different visualization-related tasks, such as automatically generate visualizations and communicate richer insights to users. Embracing these intersections between foundation models and visualizations will advance both fields and improve collaboration between humans and AI.

While the integration of foundation models and visualizations is promising, it also introduces some challenges and new opportunities. On the one hand, the increasing scale and complexity of foundation models make the models difficult to analyze and interpret using traditional manners. This highlights the need for novel visualization techniques tailored to large-scale models. On the other hand, while foundation models have demonstrated a capability to unlock new dimensions of visualization, methods for maximizing their capability and the seamless integration of humans and AI in developing visualizations are topics that remain largely underexplored. Despite the promising potential of combining foundation models and visualizations, to the best of our knowledge, no comprehensive review has been made available on this topic. Previous studies have primarily summarized the intersections between traditional machine learning models (e.g., boosting trees and convolutional neural networks) and visualizations, such as machine learning for visualization [15, 16, 18] and visualization for machine learning [12–14, 19]. In this survey, we took initial steps to highlight both the challenges and opportunities of this emerging research topic to invite further research.

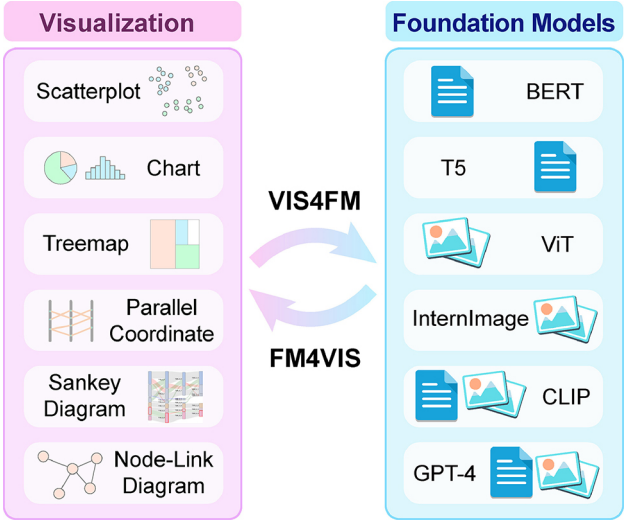

Fig. 1 Intersections between visualizations and foundation models divided into two categories: VIS4FM and FM4VIS.

2 Overview

The intersections between visualizations and foundation models concern two perspectives: VIS4FM and FM4VIS.

2.1 VIS4FM

VIS4FM focuses on leveraging the power of visualization tools to understand, refine, and evaluate intricate foundation models. Figure 2 shows that foundation models conduct two primary phases: training and adaptation [1].

Data are the basis for building foundation models and are critical in determining the performance, reliability, and ethical standing of the resulting models. Therefore, ensuring that the data are of high quality is crucial, such as broad coverage and precise annotations [46–49]. Given that foundation models often have billions or even trillions of parameters, they can learn from vast datasets and absorb both the beneficial and problematic aspects of the data. Consequently, the data must be ensured to be not only extensive but also of high quality. Visualizations facilitate the **data curation** process based on four aspects. First, visualizations guide the data generation process using real-time feedback regarding the data coverage and correctness. This allows for immediate adjustments to be made such that the generated data adequately represent the intended

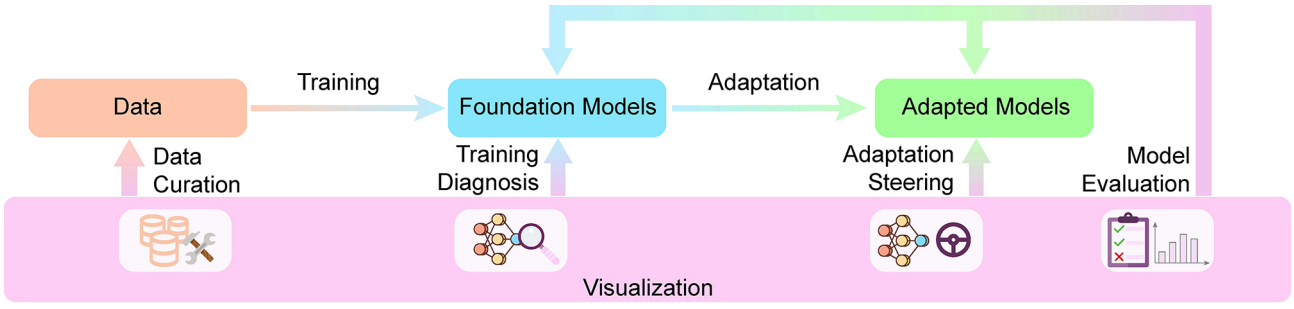

Fig. 2 How visualizations enhance foundation models along the learning pipeline.

scope and have the correct annotations. Second, visualization is useful for integrating heterogeneous data from multiple sources into a coherent and high-quality dataset. This is required for training successful foundation models. Third, visualization assists in data selection by providing a visual representation of the dataset. This simplifies the identification of high-quality samples. The feedback provided by users through visualization is used to further refine the dataset. Fourth, visualization discloses anomalies and biases in the data and enables more targeted corrections. This improves both the efficiency and accuracy of the data correction process.

Training is the initial phase in building foundation models. The models are trained on vast datasets that often contain diverse and general information. This allows the models to learn a wide range of features, patterns, and knowledge from the data. During this phase, visualization is essential for **training diagnoses** [50, 51]. First, a model explanation task is conducted to reveal the working mechanism of the foundation models. Second, the model developers conduct a performance diagnosis to identify the root causes of low performance and make necessary refinements. Finally, an efficiency diagnosis is conducted to identify bottlenecks that impair the training speed or waste resources during training.

Foundation models are typically adapted using task-specific datasets to optimize their performance on specific downstream tasks. This adaptation process refines the general knowledge of the models to better align them with the desired task outputs. In this phase, visualizations are employed to facilitate the **adaptation steering** process in three manners: model finetuning, prompt engineering, and alignment via human feedback. In model finetuning, visualizations help in understanding the knowledge learned by the models and in analyzing whether the model is suitable for the downstream tasks. With a more comprehensive understanding, model developers can then compare multiple finetuned models and select the optimal model. In prompt engineering, visualization streamlines the trial-and-error process of crafting effective prompts that lead to desired outputs. In alignment via human feedback, the model is steered toward human preferences based on human feedback. Visualizations serve two functions: (1) aid in collecting human feedback to improve the training data and (2) offer an interactive platform to iteratively refine the model outputs.

In addition, visualization is a useful technique for enhancing the **model evaluation** process for both foundation and adapted models [45]. For quantitative evaluations with clear metrics, visualizations offer users a comprehensive and intuitive understanding of the model performance. In addition, given the adaptability of foundation models to various downstream applications, evaluating their performance across multiple tasks is important. Well-designed visualizations facilitate efficient comparative analyses based on different metrics, thereby enabling users to select the optimal model or obtain insights for additional refinements. For a qualitative evaluation lacking clear metrics, visualization serves as a valuable tool for incorporating human judgment into the evaluation process. For example, consider open-ended questions that lack definitive ground-truth answers; visualizations can summarize frequent patterns in model-generated answers and provide an informative overview. This enables users to evaluate the quality of the responses more efficiently. Once low-quality responses are identified, various strategies can be employed to enhance their quality. One such method involves enriching the dataset using various instances of the associated problematic questions.

Building on the above discussion, Table 1 summa-

Table 1 Overview of the four main processes in VIS4FM

Process	Tasks supported by VIS	Description	Examples	# Examples
Data curation	Data generation	Use visualizations to help create or augment datasets	[20]	1
	Data integration	Use visualizations to help integrate data from multiple sources	—	0
	Data selection	Interactively select representative samples that align well with the tasks	—	0
	Data correction	Interactively improve the quality of datasets	[21–26]	6
Training diagnosis	Model explanation	Understand the working mechanism of models	[27–30]	4
	Performance diagnosis	Troubleshoot issues where models do not perform as expected	[31, 32]	2
	Efficiency diagnosis	Identify efficiency bottlenecks in the training process	[33]	1
Adaptation steering	Model finetuning	Analyze what knowledge the models learn during finetuning	[34, 35]	2
	Prompt engineering	Facilitate the construction of effective prompts	[36–40]	5
	Alignment via human feedback	Utilize human feedback to steer model outputs	[41]	1
Model evaluation	Quantitative evaluation	Use visualizations to present quantitative measures	[42–44]	3
	Qualitative evaluation	Use visualizations to evaluate and interpret model capability and behaviors	[45]	1

rizes the four main processes in VIS4FM. This table outlines existing initiatives and highlights areas where future research can be beneficial, particularly where few investigations have been conducted to date.

2.2 FM4VIS

FM4VIS leverages the power of foundation models to create more adaptive, user-friendly, and intelligent visualization techniques and systems. These efforts aim to advance the field of visualization. As illustrated in Fig. 3, the visualization pipeline transforms raw data into an interpretable visual representation that allows users to interact with and derive insights from the presented information [78]. FM4VIS focuses on enhancing each phase in this pipeline, from data transformation and visual mapping to view transformation and visual perception.

Data transformation converts raw data into a more suitable format for visualization and analysis purposes. Because foundation models are trained on diverse datasets, they can be used to perform **feature extraction**. This is the extraction of meaningful features from complex data for visualization purposes. It is particularly useful for unstructured data such as text and images, where traditional feature engineering methods often produce less informative

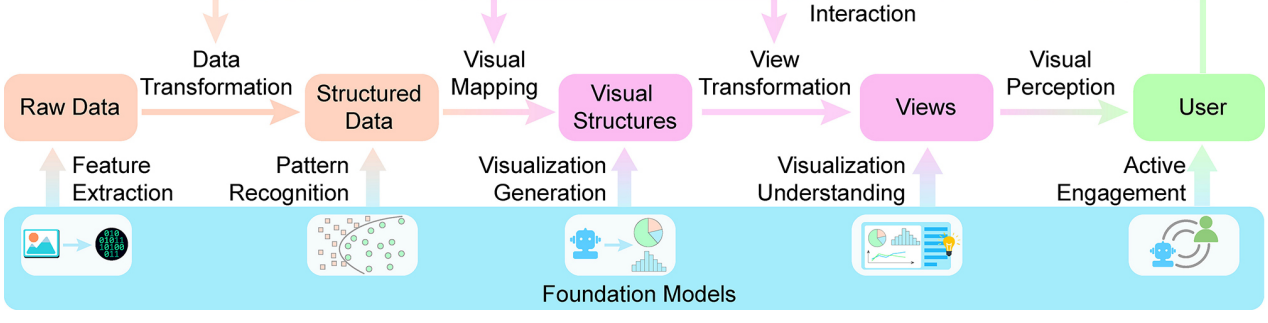

Fig. 3 How foundation models enhance visualizations along the visualization pipeline.

features [79]. Foundation models can perform tasks such as classification, relationship extraction, and object detection to extract various patterns such as relationships, trends, and anomalies. These tasks provide visualization tools with richer pattern data, thereby enabling a multi-faceted understanding and analysis.

Visual mapping determines the way to visually represent underlying data. It involves mapping data and their values to certain marks (e.g., points, lines, or areas) and visual channels (e.g., positions, colors, or sizes, respectively). Foundation models can enrich this phase by facilitating **visualization generation**, including automatic content generation, style generation, and interaction generation. These models can learn patterns and user preferences from datasets. Therefore, they can recommend or generate optimal layouts that highlight important data trends. Moreover, they can understand the context of the data and suggest appropriate marks and visual channels. For example, these models can determine which color palettes best differentiate data categories and which shapes represent specific data points more effectively. By leveraging foundation models, we can generate more insightful and contextually relevant visual representations of the data. With code generation capabilities, foundation models can augment visualization with rich interactions.

View transformation involves converting abstract visual representations into concrete pixels on a screen. It is crucial to ensure that the final visual representation is effectively communicated to users. During this phase, foundation models play an important role in **visualization understanding**, which aims to enhance the understanding of visualization content and communicate the underlying information to users. First, the models contribute to the distillation and abstraction of key information from visual presentations. For example, a foundation model can be finetuned to extract an adaptable visualization template from a set of complex timeline visualizations [73]. This involves recognizing visual elements and understanding their hierarchical and relational significance. Second, they amplify the users' comprehension of visualizations by conveying key information in an engaging, multi-modal format, such as using a combination of natural language and visual elements. For example, the models can provide clear and accurate captions that the visualization designers aim to communicate through the visualizations [75].

Visual perception is a cognitive process that occurs in the mind. It interprets the visual representation and translates the colors, shapes, and patterns back into an understanding of the underlying data. Moreover, users can interact with visualizations, such as by zooming, panning, or selecting specific data points. These interactions promote a deeper understanding and reveal further insights. Here, foundation models can achieve **active engagement**. Active engagement enhances user interactions in two aspects: direct and predictive. Direct interaction enhancement employs foundation models to directly simplify user interactions. For example, in the context of three-dimensional (3D) scatterplots, foundation models can refine the shape of a lasso selection to make them more precise and contextually relevant [80]. In addition to visual selections, these models can interpret text descriptions provided by users. For instance, when a user describes a specific pattern or attribute, the models can process the description and highlight the corresponding visual patterns on display. Predictive interaction enhancement uses foundation models to predict and enhance user interactions for immediate responses and broader data exploration insights. The predictive capabilities of these models can be leveraged to predict user actions within the visualizations. For example, after observing user interactions with a scatter plot, the models can be used to predict where users are likely to click next, streamlining the exploration process [81]. A more advanced application of these models involves analyzing user interactions. Based on how users interact with visualizations, the models can predict the imminent actions of users as well as their broader attributes, such as their likely performance on a specific task or even specific aspects of their personality [82].

Based on the aforementioned discussion, Table 2 summarizes the four main processes in FM4VIS. In addition to overviewing existing efforts, Table 2 indicates potential research directions that few studies have addressed.

3 Existing VIS4FM efforts

This section discusses recent works on VIS4FM with a focus on data curation, training diagnosis, adaptation steering, and model evaluation (Fig. 2). Table 1 lists typical examples of each category.

Table 2 Overview of the four main processes in FM4VIS

Process	Tasks supported by FM	Description	Examples	# Examples
Feature extraction and pattern recognition	Feature extraction	Extract informative features from unstructured data	[52–58]	7
	Pattern recognition	Automatic identification of patterns in data	[59–66]	8
Visualization generation	Content generation	Generate desired visualization content	[67, 68]	2
	Style generation	Generate desired styles	[69]	1
	Interaction generation	Generate desired interactions	—	0
Visualization understanding	Content extraction	Understand and extract content from visualization	[70–73]	4
	Information communication	Summarize and communicate underlying information	[74–76]	3
Active engagement	Direct interaction enhancement	Directly enhance user interactions	[77]	1
	Predictive interaction enhancement	Understand user intent to predict the next interaction	—	0

3.1 Data curation

Visualization can simplify the data curation process in four aspects: data generation, integration, selection, and correction. Existing efforts have primarily focused on data generation and correction.

3.1.1 Data generation

Data generation involves creating new data based on existing data using large-scale machine learning models trained on a large amount of data from different domains. It plays a crucial role in improving machine learning datasets by employing techniques such as filling in missing values, balancing class distributions, and augmenting sparse data collections. Based on their content generation capabilities, foundation models boost the efficiency and effectiveness of generating datasets that can be used to train, finetune, and test models. However, these automatically generated datasets typically contain quality issues, such as the presence of undesirable repetitions and incorrect information (e.g., incorrect annotations, out-of-range values, and untrue relationships). Undesirable repetitions refer to samples that are either highly similar or identical to the seed samples used in dataset generation. These repetitions may hinder the diversity of the generated datasets. To address this, Reif et al. [20] developed LinguisticLens, a visualization tool for identifying potentially undesirable repetitions in a generated dataset. This tool organizes similar sentences into clusters based on syntactic and lexical information. The clustering results allow users to analyze the linguistic patterns and individual sentences in each cluster more efficiently. Based on this comprehensive understanding, users can determine whether similar sentences are valuable enhancements or undesirable repetitions.

3.1.2 Data correction

Data correction refers to the process of correcting noisy annotations and untrue correlations between inputs and outputs (shortcuts) within training datasets. In the context of traditional deep learning models, many visual analysis methods have been developed to improve both the effectiveness and efficiency of data correction processes, including improving the instance representativeness [21–23] and enhancing the annotation quality [24–26]. Given the emphasis on data-centric issues, these methods are readily transferable to enhancing the data quality in the adaptation of foundation models. For example, ShortCutLens [21] facilitates the identification of shortcuts in natural language datasets. It overviews potential shortcuts and allows users to analyze samples associated with specific shortcuts. Once identified, these shortcuts can be addressed by constructing new samples, modifying existing ones, or removing misleading ones. Another exemplary work is DataDebugger, which was developed by Xiang et al. [25]. DataDebugger employs a hierarchical visualization to facilitate the examination and correction of annotations (Fig. 4). Users can navigate through the dataset, identify samples of interest, and provide accurate annotations. These annotations are

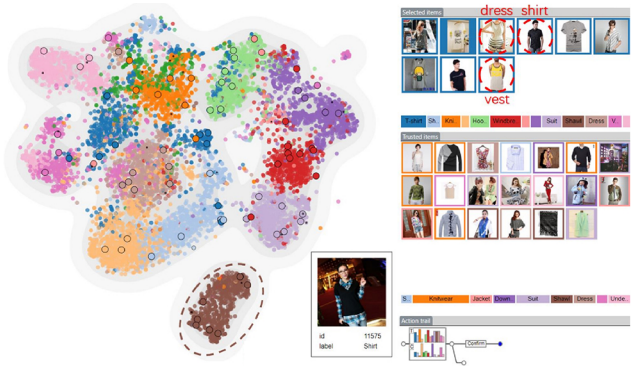

Fig. 4 DataDebugger interface. Reproduced with permission from Ref. [25], © IEEE 2019.

then propagated to correct other noisy annotations using an annotation correction algorithm, thereby reducing human efforts.

3.2 Training diagnosis

Based on the main analytical focus, existing VIS4FM efforts in training diagnosis can be divided into three categories: model explanation, performance diagnosis, and efficiency diagnosis.

3.2.1 Model explanation

Model explanation refers to the process of interpreting the working mechanism of machine learning models and how they make decisions. Recently, transformer-based foundation models such as BERT and VisionTransformer have achieved remarkable performance across various tasks [27–30]. Although the success of these models is often attributed to the self-attention mechanism, the working mechanism remains somewhat unclear. To address this, DeRose et al. [28] developed Attention Flows to interpret how attention flows across tokens and how it contributes to the final prediction results. In addition, it supports the comparison of attention flows between two models to enable an analysis of their similarities and differences. Li et al. [29] proposed a visual analysis tool tailored for analyzing VisionTransformer. This tool offers a multi-faceted examination of attention, including the importance of different attention heads, attention strengths across different image patches, and attention patterns learned by individual heads. While these methods are effective in interpreting working mechanism based on individual samples, analyzing patterns across multiple samples provides a more comprehensive perspective. To this end, Yeh et al. [30] introduced AttentionViz, which is a tool designed to simultaneously examine self-attention patterns across multiple input samples. First, it projects queries and key vectors used by the transformers into a shared space. By examining these query–key interactions in the shared space, model developers can better understand the behavior of different attention heads.

3.2.2 Performance diagnosis

Performance diagnosis aims to troubleshoot issues where models do not perform as expected and understand the reasons for this. Compared with model explanation, performance diagnosis focuses more on diagnosing performance issues than explaining the working mechanism of the model. Visualization techniques provide an interactive and intuitive environment for streamlining the performance diagnosis process. For example, Li et al. [31] developed DeepNLPVis to identify and diagnose performance issues in deep natural language processing models. DeepNLPVis introduces an information-based sample interpretation method to extract intra- and inter-word information. Corpus-, sentence-, and word-level visualizations are tightly integrated to visually explain the model behavior. With a comprehensive understanding of how the model processes inputs, model developers can efficiently identify and address performance issues. Moreover, SliceTeller [32] allows model developers to diagnose model performance on different subsets of validation data. First, SliceTeller automatically constructs several subsets of data with potential performance issues and presents them for performance diagnosis. After the model developers identify the critical subsets for further optimization, SliceTeller estimates the performance changes across different subsets. This enables developers to compare trade-offs and decide whether to accept the optimization.

3.2.3 Efficiency diagnosis

Unlike the performance diagnosis, an efficiency diagnosis focuses on identifying bottlenecks that slow the training speed or consume unnecessary resources during training. As foundation models continue to increase in scale, the importance of the efficiency diagnosis becomes increasingly critical. A widely used strategy for accelerating the training of a foundation model is to parallelize the process in a distributed cluster. Despite the effectiveness, diagnosing the parallel training process is challenging due to the intricate nature of parallelization strategies and the

large volume of profiling data, such as execution time, resource utilization, and communication overhead. To address these issues, Wei et al. [33] proposed a visual analysis method for diagnosing parallel training processes. This method integrates detailed information regarding the parallelization strategy into a computational graph, which is visualized using a directed acyclic graph layout. To facilitate the analysis of the profiling data, Wei et al. designed an enhanced Marey's graph to visualize the execution time of the network layers, peak memory of different devices, and inter-device communication latency. In addition, an aggregation method is employed to handle the large volume of profiling data within Marey's graph.

3.3 Adaptation steering

Based on the methods used to align models with human preferences, existing VIS4FM efforts in adaptation steering can be divided into three categories: model finetuning, prompt engineering, and alignment via human feedback.

3.3.1 Model finetuning

Model finetuning is a widely used technique for adapting foundation models to downstream tasks by updating the model parameters using task-specific training data. In model finetuning, model developers aim to understand the knowledge that the models learn and whether this knowledge is suitable for downstream tasks. Visualizations have been demonstrated to be effective in providing insights into model behavior [34, 50, 83] and thus serve as a useful method for accelerating the finetuning process. For example, Wang et al. [34] developed CommonsenseVIS to analyze the commonsense knowledge learned by the models and whether the knowledge is used in the models' reasoning. First, it employs a knowledge graph to extract the commonsense knowledge from the input data. The alignment of the model behavior with human reasoning is then achieved using the overlap between the extracted and learned knowledge. Using interactive visualizations for the alignment, model developers can effectively understand and diagnose issues for which the models underperform in terms of learning. In addition to the finetuning of foundation models, a growing trend has been observed toward parameter-efficient methods, such as the adapter [84] and low-rank adaptation (LoRA) [85] methods. These methods add task-specific parameters to the foundation models and train only new parameters. This reduces the training complexity and allows the adapters and LoRA modules to learn task-specific knowledge without modifying the weights of the foundation models. Consequently, many publicly available adapters and LoRA modules have been finetuned for different tasks and datasets [86]. Understanding what task-specific knowledge is acquired can help model developers in selecting an appropriate adapter or LoRA module for their tasks. For example, Sevastjanova et al. [35] proposed a visual analysis method to compare the knowledge learned by different adapters. The method integrates three types of explanation methods: concept embedding similarity, concept embedding projection, and concept prediction similarity. These methods are used to compare the adapters. This method enables developers to make informed decisions regarding which adapter best suits the downstream task of interest.

3.3.2 Prompt engineering

Instead of using traditional finetuning methods, foundation models can be adapted for downstream tasks using prompting techniques. A prompt is a natural language description of a task that makes the task suitable for foundation models. The prompt can significantly influence the model performance, and designing a high-performing prompt requires deep expertise. To alleviate the burden of manually crafting prompts, Strobelt et al. [36] developed PromptIDE, which allows users to construct different prompts, compare their performance, and interactively refine them. Figure 5 illustrates the basic workflow. First, the range of variables in a prompt template is specified, and a comprehensive set of prompts that spans all potential combinations can then be generated. The

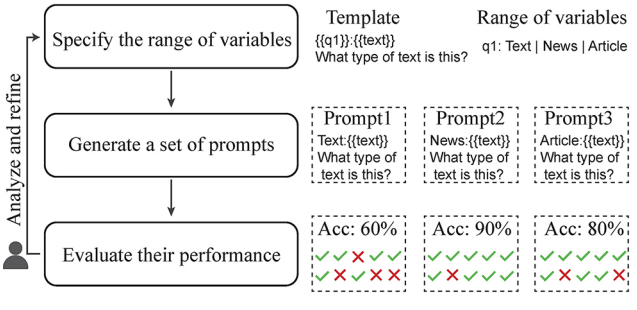

Fig. 5 Prompt engineering workflow in PromptIDE.

generated prompts are evaluated using a small set of validation data with ground-truth labels to provide quantitative measures. Users can then compare their performance and refine the prompt template or a single prompt. Similarly, ScatterShot [37] focuses on helping users interactively select informative samples and add them to the prompts. It employs a clustering technique to organize samples into clusters based on task-specific key phrases and offers a performance estimation for each cluster. Low-performance clusters are prioritized for further exploration and sample selection. For tasks without clear quantitative measures, such as text-to-image generation, visualization can assist in exploring the relationships between the input prompts and output results. For example, PromptMagician [38] streamlines the interactive refinement of text prompts in text-to-image generation tasks. It employs a prompt-recommendation model to retrieve prompt-image pairs that are similar to the input prompt from a preexisting database. The retrieved pairs are visualized in a two-dimensional (2D) space using t-distributed stochastic neighbor embedding (t-SNE) and organized using hierarchical clustering for efficient exploration. Important and relevant prompt keywords are extracted to facilitate prompt refinement. Recently, the chain-of-thought technique has emerged as an effective strategy to enhance the performance of foundation models for handling complex tasks [87]. A chain of thought is a series of prompts that breaks down a complex task into a sequence of more manageable sub-tasks. Visual analysis tools can aid users with limited experience in authoring their own chains [39, 40]. For example, Wu et al. [40] developed AIChains, which supports eight primitive operations that are well suited for language models. An interactive interface was designed to facilitate the examination and the analysis of the chain structure and model outputs. Based on the analysis, users can adjust different granularities, ranging from refinement within an individual prompt to modifying the intermediate model outputs and even restructuring the entire chain.

3.3.3 Alignment via human feedback

Unlike model finetuning and prompt engineering, model alignment directly utilizes human feedback to steer the model outputs toward human preferences. Visualization techniques are suitable for collecting human feedback and communicating the associated changes in the model output. Through this human-in-the-loop process, users can iteratively align the model outputs with their preferences. Recently, TaleBrush [41] was developed to support writers in iteratively crafting stories. TaleBrush employs line-sketching interactions along with a GPT-based language model to support writers in dictating character fortune plots in line with the creative goals of the writers. Writers can refine the generated narrative by editing the text and modifying the initial sketches.

3.4 Model evaluation

Foundation models can be evaluated quantitatively and qualitatively.

Quantitative evaluation. Quantitative evaluation employs predefined quantitative measures to evaluate the model performance. Various visualization techniques have been developed to enrich the presentation of these quantitative measures, thereby offering a comprehensive and intuitive understanding of the model performance [42–44]. For example, Görtler et al. [44] developed Neo, which extends traditional confusion matrices to facilitate the evaluation of classification tasks with complex label structures. Users can efficiently explore confusion matrices related to hierarchical or multi-output labels and inspect model confusion patterns.

Qualitative evaluation. Qualitative evaluation lacks clear metrics and often rely on visualizations to integrate human judgment into the evaluation process. For example, Chen et al. [45] developed Uni-Evaluator, a unified evaluation method suitable for various tasks in computer vision, including image classification, object detection, and instance segmentation (Fig. 6). In addition to revealing class-

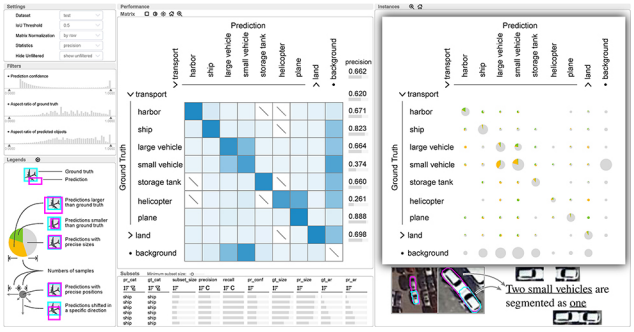

Fig. 6 Uni-Evaluator interface. Reproduced with permission from Ref. [45], © IEEE 2024.

level confusion patterns, Uni-Evaluator facilitates fine-grained examinations of the model capabilities and behaviors at the sample level. For example, when users visually compare model-generated segmentation masks with ground-truth masks, they tend to observe inadequate segmentations of the helicopter rotors. These rotors, due to their thin and limited surface area, are often overlooked or inadequately segmented in the model output. This observation has guided the enhancement of model performance by incorporating a boundary-based loss specifically for helicopter segmentations.

4 Existing FM4VIS efforts

This section introduces recent efforts on FM4VIS with a focus on feature extraction and pattern recognition, visualization generation, visualization understanding, and active engagement (Fig. 3). Table 2 lists typical examples of each category.

4.1 Feature extraction and pattern recognition

4.1.1 Feature extraction

Feature extraction transforms unstructured data, such as text and images, into semantic feature vectors. Foundation models pretrained on vast datasets often outperform traditional models in this task [1]. These high-quality semantic feature vectors facilitate the advancement of visualization techniques. Methods for enhancing visualizations include querying relevant data [52–57] and enriching metadata [58]. For example, Erato [52] is a human–machine cooperative system for generating data stories (Fig. 7). Once users determine key data facts for the story that they want to focus on, Erato utilizes an interpolation algorithm to generate intermediate data facts that smoothly connect the different key data facts. To achieve this, a BERT model is finetuned to generate high-quality fact embeddings for fact interpolation. Similarly, MetaGlyph [53] utilizes a pretrained sentence-BERT to transform both the descriptions of data attributes and data topics into semantic features. MetaGlyph then calculates the distances between these features and ranks the attributes according to the distances between the attribute descriptions and data topics. Attributes with smaller distances are prioritized for selection and subsequently visualized.

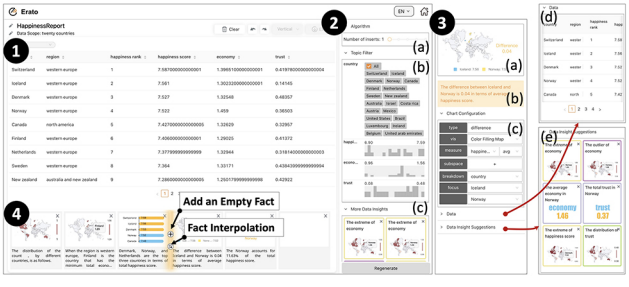

Fig. 7 Erato interface. Reproduced with permission from Ref. [52], © IEEE 2022.

4.1.2 Pattern recognition

Pattern recognition utilizes the extracted features to identify a range of patterns that enhance both understanding and analysis. Similar to existing methods that employ traditional machine learning models, foundation models are used to perform various tasks, such as classification [59–63], object detection [64, 65], and relationship extraction [66]. For example, LegalVis [59] employs a finetuned Longformer model to identify binding precedents (past legal decisions made by higher courts) in legal documents. Similarly, Teddy [60] utilizes a finetuned BERT model to extract fine-grained opinions (e.g., cleanliness and service) from review text and convey them to data scientists.

4.2 Visualization generation

Foundation models have been used to facilitate the visualization generation process by either directly generating visualization content (e.g., visualization types, data encodings, and annotations) [67, 68] or generating visualization styles (e.g., color schemes, layout styles, and typographies) [69].

4.2.1 Content generation

Content generation uses foundation models to produce desired visualization content. For example, Liu et al. [67] developed ADVISor to generate visualizations with annotations given tabular data and natural language questions. In ADVISor, a BERT model is first finetuned to extract the features of both the questions and table heads. Subsequently, several lightweight models are trained to determine the selected attributes, aggregation types, visualization types, and annotations that best address target questions. A corresponding visualization is generated based on this information. Data Player [68] is another representative method designed to simplify the creation of data videos based on static input

visualizations and corresponding narrative text. As illustrated in Fig. 8, Data Player uses OpenAI gpt-3.5-turbo and a large language model (LLM) to establish semantic connections between the visualization components and narrative entities. These semantic connections are then used to generate narration–animation interplay in the resulting data videos.

4.2.2 Style generation

Foundation models have been leveraged to produce desired visualization styles. Xiao et al. [69] developed ChartSpark to simplify the generation of pictorial chart visualizations. ChartSpark employs a text-to-image diffusion model to generate the corresponding visualization style for given semantic text prompts. In addition, it can take a chart image as an additional input to ensure that the generated visualization approximates the given chart. To further enhance the quality of the final output, users can utilize image-to-image generation techniques to improve the harmony and consistency of the generated charts.

4.3 Visualization understanding

Existing efforts on visual understanding can be classified into two categories: content extraction and information communication.

4.3.1 Content extraction

Content extraction focuses on extracting important content from visualizations, including data content [70–72] and visualization templates [73]. In terms of extracting data content from visualizations, Ma et al. [71] finetuned several models to classify chart types, analyze legends, and detect different visual elements such as boxes and points. The detected elements are converted back into data values based on the legend information. To extract visualization templates, Chen et al. [73] utilized deep learning models to segment and extract visual elements from timeline infographics. The extracted graphical elements are used as visualization templates to create similar infographics using different data.

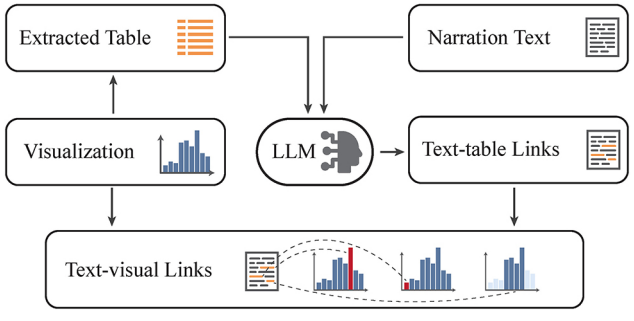

Fig. 8 Text-visual linking process in Data Player.

4.3.2 Information communication

With the capability of content generation, foundation models serve as valuable tools for communicating extracted content and underlying information to users [74–76]. For example, Sultanum and Srinivasan [74] proposed DataTales to create data-driven articles based on data visualizations. DataTales uses charts as input and leverages OpenAI gpt-3.5-turbo to generate corresponding narratives and titles. These generated narratives are then linked back to the original chart to improve the readability and overall comprehension of the given data. Liu et al. [75] developed AutoTitle, an interactive tool designed to generate meaningful titles for visualizations. It first extracts the underlying data from the visualizations and then computes high-level facts through operations such as aggregation and comparison. Based on the computed facts, a T5 [88] foundation model is finetuned to generate fluent and informative natural language descriptions.

4.4 Active engagement

Foundation models offer a promising way for understanding user intent and refining interaction results. For example, entering text without input devices in a virtual environment is challenging and typically involves many errors. By leveraging a BERT model to re-rank possible word alternatives in a user's text input, the word error rate can be significantly reduced [77]. In addition to refining the interaction results, some efforts have been made to simplify the interaction process, for example, by employing natural language [89].

5 Research opportunities

This section explores potential avenues for research on VIS4FM and FM4VIS. In particular, we focus on identifying underexplored, potential, and new challenges to offer a straightforward roadmap for future studies.

5.1 VIS4FM

5.1.1 Data curation

Data generation. Foundation models have demonstrated a capability of generating training datasets for specific tasks. Automatically generated datasets may contain several quality issues, including undesirable

repetition, low coverage, and incorrect annotations. Although an initial effort to address undesirable repetition has been made [20], the issues of low coverage and incorrect annotations remain underexplored. For the issue of low coverage, visualizations offer a useful manner of exploring the distribution of generated datasets and identifying regions with insufficient training samples. Based on the findings, users can interactively steer the data generation strategies to generate more samples in those regions. For the issue of incorrect annotations, visualizations serve as a powerful tool for users to enhance the data quality. For example, with appropriate visualizations, specific subsets in which the samples tend to contain noisy annotations can be easily identified. These corrections provide valuable feedback for the foundation models and contribute to the generation of more accurate data. In addition, incorrect annotations can be addressed via data selection, which is facilitated by visualizations and is discussed in the following.

Data integration. Foundation model training typically requires the collection and preprocessing of vast amounts of data from multiple sources. Merging these heterogeneous data into a coherent and high-quality dataset poses considerable complexities, such as handling data inconsistencies and resolving semantic differences across different sources. These issues often lead to improvements in human feedback during the integration process. In this context, visualization techniques are typically crucial in facilitating more efficient data integration and governance processes. One interesting avenue for future research is the development of a visualization-guided preprocessing framework that enables interactive adjustments to the preprocessing procedure and continuous monitoring of data integrity. Another promising avenue is the investigation of visualization techniques that can simultaneously handle the large-scale and heterogeneous natures of training data. These techniques would facilitate comparisons of data distributions from different sources and the identification of inconsistencies.

Data selection. The training and adaptation of foundation models are computationally intense processes and typically require millions or even billions of training data [8]. This large-scale data requirement introduces several complexities, including data storage, computational power, and processing time. Furthermore, the training of foundation models is becoming a serious source of carbon emissions that threaten our environment [90]. Recent studies have shown that selecting a subset of data for training can achieve comparable or even better performance [88, 91]. These findings suggest the possibility of reducing computational and environmental costs associated with model training. Visualization is a valuable tool for exploring large-scale datasets and selecting high-quality training data [92, 93]. However, two major challenges must be addressed.

The first challenge is scalability. This is particularly important in the context of foundation models. The large amount of data for training and finetuning these models is too large to fit in memory, increasing the difficulty of the simultaneous processing and visualizing of all the data. This not only calls for out-of-memory sampling techniques but also poses real-time interaction challenges for visualization. Out-of-memory sampling techniques can be used to present an overview of the data distribution. This allows users to examine the general landscape quickly and identify regions that warrant closer inspection. Users can then zoom in on these targeted regions for a more granular analysis. Because new data are not loaded from memory, studying how to support real-time interactions is worthwhile.

The second challenge stems from the unannotated and unstructured nature of training data. Most training data for foundation models, such as images or text crawled from websites, are unstructured without annotations. Their unannotated nature increases the difficulty in evaluating the quality of training data and selecting high-quality samples for training. One possible solution is to design multiple metrics to visually summarize the data characteristics from different perspectives. The unstructured nature of the data poses difficulties for users in quickly understanding the content of the samples. Innovative visualizations of the data are then required to alleviate the cognitive load. In addition, multi-modal data have been widely used in training foundation models. However, the visualization of alignments between different modalities remains underexplored and deserves further investigation.

The selection of test data shares challenges with the selection of training data, including scalability

and the unstructured nature of the data. However, some differences are noteworthy. The test data are primarily intended to faithfully convey the performance of the foundation models while exposing their potential weaknesses. Therefore, the test data must cover both the common samples that models process regularly and "edge case" samples where the models may fail. Visualization techniques are suitable for examining the selection balance between the two types of samples. Therefore, the integration of visualization techniques with the subset selection method is worth exploring for a well-balanced selection.

5.1.2 Training diagnosis

Model explanation. The intrinsic nature of foundation models is defined by their vast number of parameters. Although this vastness is the source of the models' capabilities, it also makes model interpretation difficult. Understanding the complex interactions, transformations, and computations within these parameters is challenging. When a foundation model produces an output, the output is the result of a cascade of intricate operations influenced by millions, or even billions, of parameters. Tracing back these operations to identify the exact reasoning or mechanism is similar to navigating a vast, complex maze without a map. As a model increases in size and complexity, understanding the specific factors or processes contributing to the output becomes increasingly difficult.

The aforementioned challenge posed by the scale and complexity of foundation models requires innovative visualization solutions to incorporate human knowledge into the analysis process. These visualization tools can serve as "lenses" that allow users to investigate the intricacies of these models and offer insights that can be understood intuitively. In addition, exploration based on rich interaction techniques is important for explaining foundation models. These exploration methods aim to distill the complex behaviors of foundation models into more understandable forms without compromising their essence. This might involve developing multi-level interpretation mechanism where users can select the granularity of the explanation, leverage unsupervised techniques to automatically identify the most salient features or operations driving the model decisions, and present them for further analysis. Multi-level interpretation mechanism is tailored to offer explanations at varying levels of detail, from high-level overviews to detailed, granular insights. At the highest level, these explanations provide a general summary of the models' decision-making logic. This is a surface-level interpretation. For example, for a text generation task, a surface-level explanation might state, "the model generated this sentence based on the overall sentiment of the input". In addition, it can summarize associated statistics, such as confidence and bias scores. The next level provides a component-level interpretation that aims to explain the role of specific model components, such as particular layers or attention heads. For example, "the 10th attention head focused primarily on the relationships between the subject and object in the sentence". The deepest potential level can provide a parameter-level interpretation. This enables the examination of the influence and interactions of specific parameters or groups of parameters. This can involve visualizing the weights, gradients, or activations associated with particular tokens or features. Given the vast amount of data present at each level, an effective sampling method that can easily capture human interest and display the corresponding data is in demand. This has motivated studies on interactive sampling strategies, which require the development of interactive visualizations to facilitate the detection of different user intents and provide tailored data subsets. These strategies enable users to seamlessly navigate through complex data layers. For example, they probe deeper into specific areas of interest or approach the issue by taking a step back for a broader perspective to enhance the overall understanding of the model functioning.

Online training diagnosis. With the increasing complexity of foundation models, their training time typically requires weeks or even months for high-end GPUs. Traditional offline methods gather relevant data after the training process and then feed them to an analysis tool. This is less effective in reducing unnecessary training trials. Moving the visual analysis earlier in the model development workflow can save vast amounts of time and computational resources, such as by halting ineffective and inefficient training. Therefore, visualization techniques suitable for monitoring results in real time and identifying performance and/or efficiency issues

must be developed. Two interesting avenues warrant further exploration.

The first promising avenue is to support an in-depth analysis of model performance during model training. Although some existing methods, such as Tensorboard [94], have supported the online monitoring of the training process, they consider only high-level performance metrics, such as the loss and prediction accuracy. These metrics are too abstract to effectively troubleshoot why the model does not perform as expected. To address this, it is necessary to integrate advanced data and model analysis modules into the visualizations to provide richer information. By analyzing the sample content and how the model processes it, model developers can obtain more insights into the performance issues and address them accordingly.

The second promising avenue lies in the management of large-scale profiling data for online diagnoses. Given the rapid generation of profiling data and input/output overhead associated with transferring data from GPU to memory or even disk storage, storing all the data and then transferring them to a visualization tool for analysis is an impractical approach. In-situ visualization is a promising method for addressing this issue [95]. It generates visualizations directly within the computational environment in which the data are generated. Although in-situ visualization has been demonstrated to be useful for scientific visualizations [96, 97], whether it can be employed to streamline the efficiency diagnosis during model training remains unexplored.

5.1.3 Adaptation steering

Model finetuning. After a foundation model is finetuned for a specific task, it deviates from its pretrained version. The changes can be in terms of performance metrics as well as in model behavior, such as in processing different types of inputs and developing new input–output associations. By analyzing these behavior changes, model developers can understand how generic knowledge evolves into task-specific knowledge and identify where the model does not function as expected. Therefore, a promising research opportunity lies in using visualizations to effectively monitor behavioral changes and identify abnormal behavior during the finetuning process. With a deep understanding of behavioral changes, model developers can identify when the model begins to exhibit biases or vulnerabilities that downgrade its performance. Subsequently, visualizations can be leveraged as an efficient manner of interactively steering the finetuning process, for example, by adding more balanced or targeted data. This method enhances the model performance as well as its reliability and robustness.

Prompt engineering. Recent studies have shown that providing high-quality examples within prompts can significantly enhance the model performance. This is known as the in-context learning ability [98]. In-context learning is a valuable component of prompt engineering. In this setup, prompt engineering is critical for curating and structuring examples that can effectively guide the model. To fully leverage the capabilities of foundation models and achieve satisfactory performance, the examples provided should be well suited for the downstream task. However, generating high-quality examples requires expertise and often involves iterative refinement. This typically involves trial and error. Visualizations offer an efficient method to facilitate this refinement process by integrating humans into the analysis loop [11, 13, 99]. One promising solution involves employing visualizations to illustrate model responses across different in-context examples. The insights derived from the visualizations enable users to evaluate the effectiveness of the constructed examples and identify those most suitable for the current task. Once informed, the users can then refine the examples for improved performance. In addition to interactively refining examples for each task, another promising direction lies in using visualizations to summarize the general principles for in-context example selection [100]. In exploring different subsets of examples and comparing them, users can summarize the principles that determine which types of examples are beneficial and which are not. These principles contribute to a more systematic and informed example selection to craft effective prompts for the downstream task.

Alignment via human feedback. In the model adaptation process, aligning the model behavior with human preferences is essential. This alignment improves the user experience by generating more relevant responses and addresses ethical and societal concerns [7]. Recently, reinforcement learning from

human feedback has been shown to be effective in aligning model behavior with human preferences [7]. This method first trains a reward model directly from human feedback, which predicts whether the response aligns with human preferences (high reward) or not (low reward). Subsequently, this reward information guides the optimization of foundation models through reinforcement learning. In this process, the key lies in collecting high-quality human feedback and using this data to train a reward model that accurately captures human preferences. Visualization techniques are suitable for both tasks. Interactive visualizations have already demonstrated their value in enhancing the process of collecting human feedback. For example, existing research on interactive data labeling has demonstrated the effectiveness of employing visualization techniques to facilitate the collection of human-generated data [101–103]. Moreover, visualizations offer an efficient method for diagnosing the training process of reward models and interactively refining them through additional human feedback. A tight integration of human feedback into this process better aligns the reward models with actual human preferences. This integration leads to more accurate and reliable reward information for the ongoing optimization of the foundation model.

The primary challenges in this context are rooted in the collection of high-quality human feedback and the complexities of integrating visualization techniques into reinforcement learning pipelines. First, collecting high-quality human feedback is difficult, and this difficulty is amplified when the data must be fed to the reward model that drives the reinforcement learning. Any errors or biases in the feedback collection can result in skewed training or unreliable models. Second, although visualization techniques offer the opportunity to collect human-generated data more effectively, seamlessly integrating these techniques with reinforcement learning pipelines presents additional complexities. Balancing real-time interactions with computational efficiency in a complex training process is another challenge that must be overcome.

Model selection. Recently, there has been an increasing trend among model developers to upload their models with metadata (e.g., descriptions, model architectures, and resource requirements) to learnware markets [86, 104, 105]. The increasing availability of publicly finetuned foundation models has opened new avenues for the efficient development of AI systems. When confronted with an AI task, users can search for and select a preexisting model that fits their needs from a learnware market. However, without sufficient expertise, navigating the expansive model space to determine the most suitable foundation model can be challenging [106]. The challenge lies in facilitating user exploration by capturing user requirements and recommending high-performance models. One potential solution is to employ visualization techniques to illustrate the model space. Using these visualizations, users can navigate the complex model space more easily, understand model behaviors, identify model limitations, and compare models from multiple perspectives, such as performance scores and resource requirements. Such a comprehensive understanding and comparison enable the identification of an optimal model for specific tasks.

5.1.4 Model evaluation

The field of visualization has extensively covered quantitative evaluations. Therefore, we discuss the research challenges and opportunities related to qualitative evaluations.

Evaluating free-form outputs. Recently, foundation models have achieved impressive performance in various tasks, particularly in answering open-ended questions without definitive ground-truth answers. However, evaluating the quality of free-form model responses remains challenging because of the high variability in possible responses and the absence of clear ground-truth answers. Addressing this challenge requires human involvement during the evaluation process. However, users are unable to manually inspect and assess each model response because of the huge volume of data. One possible solution is to semi-automatically create rules for evaluating model responses using active learning methods. Visualizations enhance this process by offering a comprehensive overview of the evaluation rules and their associated model responses. Subsequently, users can iteratively refine these rules based on their preferences. This ultimately leads to more accurate and reliable evaluations. Another potential solution involves using visualizations to highlight responses that are difficult for semi-automatic evaluation methods and present them to users for manual review.

To minimize redundancy and simplify this process, it is essential to cluster a massive volume of responses and summarize the clustering results in an intuitive visual form.

Robustness. Many foundation models, such as those in the GPT series [6, 8], are generative models. Although these models demonstrate impressive generation abilities, they can misinterpret inputs or generate off-target or incorrect outputs. Such inconsistencies pose challenges for the reliable deployment of these models, particularly in scenarios where a single error can have significant consequences. Therefore, clearly understanding their robustness is an urgent need. With this information, users can assess the performance of these models in different situations and identify weak areas that require finetuning to improve their performance [107, 108].

One possible solution is to construct a set of input samples with perturbations and compare the corresponding model responses with well-designed visualizations. This method effectively illustrates how small changes in the input can affect the model output. This provides insights into the robustness and sensitivity of the model. Visualizations provide an important method used to identify critical samples for closer examination, interactively construct perturbated samples for deeper behavioral insight, and summarize multiple model responses for efficient analysis. Another solution involves analyzing numerous input samples collected in real-world scenarios to identify potential robustness issues. Models are often deployed in complex environments, where they encounter a wide range of inputs. The manual examination of each robustness issue is overwhelming. Visualizations offer an effective means of exploring and filtering a set of similar inputs that produce diverse results, which frequently indicates robustness issues. Once these issues are identified, visualization tools help enable "what–if" analyses. These analyses examine how the model behaves under various conditions and then identify specific areas where its robustness could be improved.

Fairness. Given that foundation models are increasingly being deployed in diverse cultural contexts and used by diverse user groups, it is crucial to prioritize culturally sensitive, ethically sound, and socially aligned explanations provided by VIS4FM techniques. Consequently, how VIS4FM techniques can effectively navigate cross-cultural differences, address ethical dilemmas, and assess broader societal impacts are essential avenues of exploration. These research directions are essential for advancing the area of VIS4FM and ensuring responsible model deployments.

First, cross-cultural differences significantly affect how individuals perceive and interpret information. Cultural factors such as language, beliefs, values, and norms influence the understanding and acceptance of foundation models and their explanations. Therefore, how VIS4FM techniques account for and adapt to cross-cultural differences in explanation generation and presentation applications must be investigated. This involves studying cultural biases in foundation models, developing culture-aware explanation methods, and conducting user studies in diverse cultural contexts to assess the effectiveness and appropriateness of VIS4FM techniques.

Second, ethical considerations are important for the development and application of adapted models. Visualization techniques should adhere to ethical principles such as transparency, fairness, privacy, and accountability. This includes addressing issues such as algorithmic bias, discrimination, and the potential impact of VIS4FM explanations on vulnerable populations. Research on specific ethical frameworks and guidelines for VIS4FM can help ensure that adapted models with visual explanations are deployed in a responsible and ethical manner.

5.2 FM4VIS

5.2.1 Feature extraction and pattern recognition

Foundation models offer two notable opportunities that are unavailable with traditional machine learning models. First, because of their training on more diverse and extensive datasets, foundation models typically generate features of higher quality than those obtained from traditional machine learning models. These features better reveal the underlying patterns in the data, such as clusters [5, 60, 109] and important insights [52, 110, 111]. These high-quality features and patterns facilitate the design of suitable visualizations used to analyze the data. Second, previous feature extraction methods have primarily focused on single-modality data, such as latent Dirichlet allocation for textual data [112] and the scale-invariant feature transform for image data [113]. Recent research efforts have been made to train multi-modality foundation models, such as

CLIP [5], to map multi-modality data into one unified feature space. This enables researchers to design a unified visualization of multi-modal data to facilitate the disclosure of inter-modality relationships within the data.

5.2.2 Visualization generation

Prompted content generation. As widely studied foundation models, LLMs have demonstrated a capability to generate source code given natural language prompts. For example, Code LLAMA has exhibited state-of-the-art performance on several public code generation benchmarks [114]. An interesting avenue for future research could be to democratize visualization designs by extending these capabilities to automatically generate advanced visualizations. By integrating well-known engines, such as D3 [115] and matplotlib [116], this method simplifies the process for individuals without prior experience in visualization design. They can be used to create advanced visual data representations and address complex challenges. Although the execution of this concept seems intuitive using existing public APIs, it has not been fully implemented. Several research efforts are still underway to improve the quality of generated visualizations. First, the development of a visualization-related instruction-tuning dataset is critical. Currently, visualization codes such as the D3 code comprise only a small portion of the training corpus of LLMs. Therefore, developing a dataset containing both instructions and accompanying visualization code is necessary to improve the performance of creating different visualization components with LLMs. The importance of visualization-specific datasets has been demonstrated using existing automatic graph layout methods [117]. Using such datasets and leveraging advanced finetuning techniques, such as reinforcement learning from human feedback, can significantly enhance the code-generation capabilities of a model in the visualization field. Second, prompt engineering is essential to ensure that the generated visualizations align with user intent. Existing research has shown that different prompts substantially influence the output generated by LLMs [118]. Therefore, effective prompts are critical. To alleviate human efforts in the tedious prompt curation process, recent techniques, such as automatic prompt optimization [119], can be leveraged.

Style generation. In computer vision, style transfer refers to the technique of applying the visual style of one image to the content of another image [120]. This often involves a content and a style image. The algorithm reconfigures the content image to assume the artistic style of the style image. For instance, StyleGAN [121] leverages generative adversarial networks to distill the style cues from reference images. The incorporation of style-based generator layers offers fine-grained control over the image attributes. This improves the quality and versatility of the generated images. Currently, these style-transfer models remain within the domain of natural image generation. However, the principles behind style transfer offer potential applications beyond the visual arts. They open avenues to other fields, such as visualization. This remains an open but important research avenue for effectively harnessing style transfer techniques in the field of visualization. This extension would allow users to easily transfer stylistic elements from one visualization to another. Moreover, it serves as a valuable resource for users with limited programming skills and facilitates the creation of user-centric visualizations with minimal efforts. This makes complex data more accessible and understandable to a broader audience. A critical challenge in this endeavor is preserving the data integrity in transferred visualizations. Unlike natural images, visualization is a visual form of data. Therefore, a reliable representation of these data is critical. Current style transfer techniques, when applied to visualization, may introduce subtle changes in visual elements, such as line-length adjustments. This may lead to perceptual errors. A promising research opportunity lies in adapting style transfer models to incorporate the original data used to generate visualizations, thereby ensuring data integrity when transferring styles. Another challenge is the automatic recommendation of styles, which is complicated by the multifaceted intricacies of human perception and divergent individual preferences. For example, one user might prioritize clarity and simplicity, whereas another might focus on intricate details and vibrant color schemes. Additionally, cultural background, professional training, and mood can influence what a user finds engaging or easy to interpret. These varying factors make the automatic process of recommending styles a complex endeavor

because the system must account for a wide range of subjective preferences.

Interaction generation. Interaction enables users to tailor their views according to specific information requirements. This serves as a cornerstone for effective data exploration and analysis. However, creating intuitive and responsive interactions is a challenge that requires expertise in both visualization techniques and programming. The code-generation capabilities of foundation models offer significant opportunities. An interesting avenue for research is the simplified interaction design. As with the aforementioned prompted content generation, users can implement basic interactions by describing their intent using natural language. The challenge lies in the ambiguities that natural languages often present [18]. This increases the difficulty of describing complex interactive functionalities clearly. Therefore, extending foundation models to accept other types of inputs, such as sketches and video examples, is an exciting opportunity for producing more accurate interaction designs. At a more advanced level, foundation models have the potential to simplify the programming of complex interactions such as multi-stage animation scheduling and sophisticated visual effects. However, ensuring that the generated code satisfies quality standards remains an issue. Hence, a potential avenue for future research is the development of automatic quality assurance mechanisms that can evaluate and refine the code generated by foundation models.

5.2.3 Visualization understanding

Content extraction. Previous research has highlighted the enhanced reasoning capabilities inherent in foundation models [6]. Using these capabilities, visualization researchers can adapt foundation models to comprehend complex visualizations, such as node-link diagrams and tree maps, and extract key information for in-depth analyses [122]. For example, when presented with a node-link diagram representing a complex social network, foundation models can effectively identify key information such as influential users, sub-communities, and their connections. Descriptive captions and concise summaries of this information can be generated and presented alongside visualizations. This greatly facilitates comprehension. A critical challenge in adapting current foundation models to understand complex visualizations is the lack of domain-specific data. Currently, existing public datasets in the visualization field often focus on simple charts such as bar and line charts [15]. Therefore, the creation of a public dataset that contains complex visualizations and extracted insights is critical. Another challenge lies in identifying contextually relevant information that matches the analytical focus. Interactive visualizations often excel at conveying useful patterns embedded in large amounts of data. For example, the visualization of a social network may present multiple interesting sub-communities that deserve exploration. A tailored summary of the sub-communities of interest is often more beneficial than a generic overview of the entire network. Consequently, the task of capturing the analytical focus of users and dynamically extracting relevant patterns and tailored summaries for visualization has emerged as a promising avenue for future investigation.

Visual-question-answering-based communication. In computer vision, the development of machine learning models to answer questions about an image is an active research topic called visual question answering [123]. Using foundation models, users can engage in free-form and open-ended dialog regarding visualizations. This alleviates the cognitive load of understanding the visualizations. To achieve this, two key aspects must be considered. First, the model must have a robust linguistic comprehension capability and possess a large amount of knowledge to effectively address open-ended questions regarding the visualizations. While some foundation models have achieved remarkable accuracy rates exceeding 90% on the CommonsenseQA benchmark dataset [124], the ability to answer open-ended questions regarding visualization remains a topic for further study. Second, contextual awareness is a critical component that enables a smooth, multi-round dialog experience in foundation models. Currently, chat-centric models such as ChatGPT have demonstrated the ability to deliver desired results conditioned on previous user prompts in the dialog [7]. Adding the underlying data to the prompts can help the foundation model understand the visualizations more precisely and answer numerical questions. However, the incorporation of data into the prompts raises scalability issues. Directly incorporating all

the data into the prompts is inefficient, as well as unfeasible given the large volume of data. To solve this, the development of data abstraction techniques (e.g., sampling [125, 126] and statistical summary) is necessary to extract the most important data closely linked to the generated visualizations.

5.2.4 Active engagement

Direct interaction enhancement. Currently, several widely used interactions such as brushing and zooming have been overlooked in the training of foundation models. Consequently, these models struggle to understand and enhance such user interactions. Two potential solutions exist to address this gap. A straightforward solution is to convert these interactions into formats that current foundation models can readily understand. For example, mouse-click interactions can be converted into textual descriptions and fed to LLMs. A more promising solution involves training or adapting foundation models to understand these interactions directly. Encouragingly, initial efforts have been made to enhance model capabilities in this direction. For example, DragGAN enables users to manipulate objects within images using drag-and-drop interactions [127]. These efforts are notable steps toward expanding the capabilities of interaction-aware foundation models.

Predictive interaction enhancement. Recently, several initiatives have been implemented to enhance the capabilities of foundation models by creating foundation-model-based AI agents [128]. These AI agents are designed to mimic human behaviors and typically include various modules, such as perception, memory, planning, and reflection, each of which is often supported by a foundation model. Such agents can actively identify human feedback and incorporate it into their reflection module, which adapts their actions to subsequent steps based on this feedback [129]. Employing these AI agents is feasible for visual analyses. Traditional approaches require domain experts to manually examine data through visualizations and identify patterns through sequences of interactions. This process is time-consuming and expertise-dependent. By contrast, AI agents may help simplify this analysis process by generating similar interaction sequences based on the interaction sequences performed by domain experts. However, achieving productive collaboration between humans and AI agents poses two challenges.

The first challenge lies in finetuning a foundation model capable of automatically generating interaction sequences to extract useful patterns. To alleviate the efforts in interacting with different visual analysis tools, foundation models can be used to generate interaction sequences, which are then used to automatically extract pattern candidates. Domain experts need only examine these candidates and find the most relevant patterns for further analysis. The second challenge is the efficient adaptation of the foundation model to specific visual analysis tools and domain experts. To achieve this, the capacity of the model must be boosted for in-context learning. The foundation model should be able to learn from a few example interaction sequences performed by experts and then extract more patterns from similar interactions.

6 Conclusions

The intersection of foundation models and visualizations represents a substantial step in the advancement of AI systems. On the one hand, VIS4FM is crucial in explaining the complexities of foundation models. This highlights the growing need for transparency, explainability, fairness, and robustness in the expanding role of AI. On the other hand, FM4VIS provides new pathways for further advances in visualization techniques. Although integrating these two fields presents certain challenges, their potential benefits and advancements are undeniable. The challenges must be confronted directly while embracing the vast opportunities that lie ahead. This confluence not only promises a brighter future for AI and visualization but also encourages a sustained journey of discovery and innovation in this emerging research topic.

Fundings

This work was supported by the National Natural Science Foundation of China (Grant Nos. U21A20469 and 61936002), the National Key R&D Program of China (Grant No. 2020YFB2104100), and grants from the Institute Guo Qiang, THUIBCS, and BLBCI.

Author contributions

Weikai Yang: Conceptualization, Writing - Original

Draft, Writing - Review Editing. Mengchen Liu: Conceptualization, Writing - Original Draft, Writing - Review Editing. Wang Zheng: Writing - Original Draft, Writing - Review Editing. Shixia Liu: Conceptualization, Supervision, Writing - Original Draft, Writing - Review Editing, Funding acquisition.

Acknowledgements

The authors thank Dr. Xiting Wang, Dr. Changjian Chen, Jun Yuan, Yukai Guo, Jiangning Zhu, and Duan Li for their valuable comments.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. On the opportunities and risks of foundation models. *arXiv preprint* arXiv:2108.07258, 2021.

[2] Devlin, J.; Chang, M. W.; Lee, K.; Toutanova. K. BERT: Pretraining of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186, 2019.

[3] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations, 2021.

[4] Wang, W.; Dai, J.; Chen, Z.; Huang, Z.; Li, Z.; Zhu, X.; Hu, X.; Lu, T.; Lu, L.; Li, H.; et al. InternImage: Exploring large-scale vision foundation models with deformable convolutions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 14408–14419, 2023.

[5] Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning, 8748–8763, 2021.

[6] Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. In: Proceedings of the 34th Conference on Neural Information Processing Systems, 1877–1901, 2020.

[7] Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. In: Proceedings of the 36th Conference on Neural Information Processing Systems, 27730–27744, 2022.

[8] OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S. GPT-4 technical report. *arXiv preprint* arXiv:2303.08774, 2023.

[9] Eloundou, T.; Manning, S.; Mishkin, P.; Rock, D. GPTs are GPTs: An early look at the labor market impact potential of large language models. *arXiv preprint* arXiv:2303.10130, 2023.

[10] Liu, S.; Wang, X.; Liu, M.; Zhu, J. Towards better analysis of machine learning models: A visual analytics perspective. *Visual Informatics* Vol. 1, No. 1, 48–56, 2017.

[11] Choo, J.; Liu, S. Visual analytics for explainable deep learning. *IEEE Computer Graphics and Applications* Vol. 38, No. 4, 84–92, 2018.

[12] Hohman, F.; Kahng, M.; Pienta, R.; Chau, D. H. Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE Transactions on Visualization and Computer Graphics* Vol. 25, No. 8, 2674–2693, 2019.

[13] Yuan, J.; Chen, C.; Yang, W.; Liu, M.; Xia, J.; Liu, S. A survey of visual analytics techniques for machine learning. *Computational Visual Media* Vol. 7, No. 1, 3–36, 2021.

[14] Sacha, D.; Kraus, M.; Keim, D. A.; Chen, M. VIS4ML: An ontology for visual analytics assisted machine learning. *IEEE Transactions on Visualization and Computer Graphics* Vol. 25, No. 1, 385–395, 2019.

[15] Wang, Q.; Chen, Z. T.; Wang, Y.; Qu, H. A survey on ML4VIS: Applying machine learning advances to data visualization. *IEEE Transactions on Visualization and Computer Graphics* Vol. 28, No. 12, 5134–5153, 2022.

[16] Wu, A.; Wang, Y.; Shu, X.; Moritz, D.; Cui, W.; Zhang, H.; Zhang, D.; Qu, H. AI4VIS: Survey on artificial intelligence approaches for data visualization. *IEEE Transactions on Visualization and Computer Graphics* Vol. 28, No. 12, 5049–5070, 2022.

[17] Wang, J.; Liu, S.; Zhang, W. Visual analytics for machine learning: A data perspective survey. *arXiv preprint* arXiv:2307.07712, 2023.

[18] Shen, L.; Shen, E.; Luo, Y.; Yang, X.; Hu, X.; Zhang, X.; Tai, Z.; Wang, J. Towards natural language interfaces for data visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics* Vol. 29, No. 6, 3121–3144, 2023.

[19] Liu, S.; Wang, X.; Collins, C.; Dou, W.; Ouyang, F.; El-Assady, M.; Jiang, L.; Keim, D. A. Bridging text visualization and mining: A task-driven survey. *IEEE Transactions on Visualization and Computer Graphics* Vol. 25, No. 7, 2482–2504, 2019.

[20] Reif, E.; Kahng, M.; Petridis, S. Visualizing linguistic diversity of text datasets synthesized by large language models. *arXiv preprint* arXiv:2305.11364, 2023.

[21] Jin, Z.; Wang, X.; Cheng, F.; Sun, C.; Liu, Q.; Qu, H. ShortcutLens: A visual analytics approach for exploring shortcuts in natural language understanding dataset. *IEEE Transactions on Visualization and Computer Graphics* doi: 10.1109/TVCG.2023.3236380, 2023.

[22] Chen, C.; Yuan, J.; Lu, Y.; Liu, Y.; Su, H.; Yuan, S.; Liu, S. OoDAnalyzer: Interactive analysis of out-of-distribution samples. *IEEE Transactions on Visualization and Computer Graphics* Vol. 27, No. 7, 3335–3349, 2021.

[23] Yang, W.; Li, Z.; Liu, M.; Lu, Y.; Cao, K.; Maciejewski, R.; Liu, S. Diagnosing concept drift with visual analytics. In: Proceedings of the IEEE Conference on Visual Analytics Science and Technology, 12–23, 2020.

[24] Liu, S.; Chen, C.; Lu, Y.; Ouyang, F.; Wang, B. An interactive method to improve crowdsourced annotations. *IEEE Transactions on Visualization and Computer Graphics* Vol. 25, No. 1, 235–245, 2019.

[25] Xiang, S.; Ye, X.; Xia, J.; Wu, J.; Chen, Y.; Liu, S. Interactive correction of mislabeled training data. In: Proceedings of the IEEE Conference on Visual Analytics Science and Technology, 57–68, 2019.

[26] Bäuerle, A.; Neumann, H.; Ropinski, T. Classifier-guided visual correction of noisy labels for image classification tasks. *Computer Graphics Forum* Vol. 39, No. 3, 195–205, 2020.

[27] Li, R.; Xiao, W.; Wang, L.; Jang, H.; Carenini, G. T3-Vis: Visual analytic for Training and fine-Tuning Transformers in NLP. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 220–230, 2021.

[28] DeRose, J. F.; Wang, J.; Berger, M. Attention flows: Analyzing and comparing attention mechanisms in language models. *IEEE Transactions on Visualization and Computer Graphics* Vol. 27, No. 2, 1160–1170, 2021.

[29] Li, Y.; Wang, J.; Dai, X.; Wang, L.; Yeh, C. C M.; Zheng, Y.; Zhang, W.; Ma, K. L. How does attention work in vision transformers? A visual analytics attempt. *IEEE Transactions on Visualization and Computer Graphics* Vol. 29, No. 6, 2888–2900, 2023.

[30] Yeh, C.; Chen, Y.; Wu, A.; Chen, C.; Viégas, F.; Wattenberg, M. AttentionViz: A global view of transformer attention. *IEEE Transactions on Visualization and Computer Graphics* Vol. 30, No. 1, 262–272, 2024.

[31] Li, Z.; Wang, X.; Yang, W.; Wu, J.; Zhang, Z.; Liu, Z.; Sun, M.; Zhang, H.; Liu, S. A unified understanding of deep NLP models for text classification. *IEEE Transactions on Visualization and Computer Graphics* Vol. 28, No. 12, 4980–4994, 2022.

[32] Zhang, X.; Ono, J. P.; Song, H.; Gou, L.; Ma, K. L.; Ren, L. SliceTeller: A data slice-driven approach for machine learning model validation. *IEEE Transactions on Visualization and Computer Graphics* Vol. 29, No. 1, 842–852, 2023.

[33] Wei, Y.; Wang, Z.; Wang, Z.; Dai, Y.; Ou, G.; Gao, H.; Yang, H.; Wang, Y.; Cao, C. C.; Weng, L.; et al. Visual diagnostics of parallel performance in training large-scale DNN models. *IEEE Transactions on Visualization and Computer Graphics* doi: 10.1109/TVCG.2023.3243228, 2023.

[34] Wang, X.; Huang, R.; Jin, Z.; Fang, T.; Qu, H. CommonsenseVIS: Visualizing and understanding commonsense reasoning capabilities of natural language models. *IEEE Transactions on Visualization and Computer Graphics* Vol. 30, No. 1, 273–283, 2024.

[35] Sevastjanova, R.; Cakmak, E.; Ravfogel, S.; Cotterell, R.; El-Assady, M. Visual comparison of language model adaptation. *IEEE Transactions on Visualization and Computer Graphics* Vol. 29, No. 1, 1178–1188, 2023.

[36] Strobelt, H.; Webson, A.; Sanh, V.; Hoover, B.; Beyer, J.; Pfister, H.; Rush, A. M. Interactive and visual prompt engineering for ad-hoc task adaptation with large language models. *IEEE Transactions on Visualization and Computer Graphics* Vol. 29, No. 1, 1146–1156, 2023.

[37] Wu, S.; Shen, H.; Weld, D. S.; Heer, J.; Ribeiro, M. T. ScatterShot: Interactive In-context example curation for text transformation. In: Proceedings of the Proceedings of the 28th International Conference on Intelligent User Interfaces, 353–367, 2023.

[38] Feng, Y.; Wang, X.; Wong, K. K.; Wang, S.; Lu, Y.; Zhu, M.; Wang, B.; Chen, W. PromptMagician: Interactive prompt engineering for text-to-image creation. *IEEE Transactions on Visualization and Computer Graphics* Vol. 30, No. 1, 295–305, 2024.

[39] Wu, T.; Jiang, E.; Donsbach, A.; Gray, J.; Molina, A.; Terry, M.; Cai, C. J. PromptChainer: Chaining large language model prompts through visual programming. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, Article No. 359, 2022.

[40] Wu, T.; Terry, M.; Cai, C. J. AI chains: Transparent and controllable human-AI interaction by chaining large language model prompts. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, Article No. 385, 2022.

[41] Chung, J. J. Y.; Kim, W.; Yoo, K. M.; Lee, H.; Adar, E.; Chang, M. TaleBrush: Sketching stories with generative pretrained language models. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, Article No. 209, 2022.

[42] Alsallakh, B.; Hanbury, A.; Hauser, H.; Miksch, S.; Rauber, A. Visual methods for analyzing probabilistic classification data. *IEEE Transactions on Visualization and Computer Graphics* Vol. 20, No. 12, 1703–1712, 2014.

[43] Ren, D.; Amershi, S.; Lee, B.; Suh, J.; Williams, J. D. Squares: Supporting interactive performance analysis for multiclass classifiers. *IEEE Transactions on Visualization and Computer Graphics* Vol. 23, No. 1, 61–70, 2017.

[44] Görtler, J.; Hohman, F.; Moritz, D.; Wongsuphasawat, K.; Ren, D.; Nair, R.; Kirchner, M.; Patel, K. Neo: Generalizing confusion matrix visualization to hierarchical and multi-output labels. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, Article No. 408, 2022.

[45] Chen, C.; Guo, Y.; Tian, F.; Liu, S.; Yang, W.; Wang, Z.; Wu, J.; Su, H.; Pfister, H.; Liu, S. A unified interactive model evaluation for classification, object detection, and instance segmentation in computer vision. *IEEE Transactions on Visualization and Computer Graphics* Vol. 30, No. 1, 76–86, 2024.

[46] Liu, S.; Andrienko, G.; Wu, Y.; Cao, N.; Jiang, L.; Shi, C.; Wang, Y. S.; Hong, S. Steering data quality with visual analytics: The complexity challenge. *Visual Informatics* Vol. 2, No. 4, 191–197, 2018.

[47] Jiang, L.; Liu, S.; Chen, C. Recent research advances on interactive machine learning. *Journal of Visualization* Vol. 22, No. 2, 401–417, 2019.

[48] Chen, C.; Wang, Z.; Wu, J.; Wang, X.; Guo, L. Z.; Li, Y. F.; Liu, S. Interactive graph construction for graph-based semi-supervised learning. *IEEE Transactions on Visualization and Computer Graphics* Vol. 27, No. 9, 3701–3716, 2021.

[49] Chen, C.; Wu, J.; Wang, X.; Xiang, S.; Zhang, S. H.; Tang, Q.; Liu, S. Towards better caption supervision for object detection. *IEEE Transactions on Visualization and Computer Graphics* Vol. 28, No. 4, 1941–1954, 2022.

[50] Liu, M.; Shi, J.; Li, Z.; Li, C.; Zhu, J.; Liu, S. Towards better analysis of deep convolutional neural networks. *IEEE Transactions on Visualization and Computer Graphics* Vol. 23, No. 1, 91–100, 2017.

[51] Liu, M.; Shi, J.; Cao, K.; Zhu, J.; Liu, S. Analyzing the training processes of deep generative models. *IEEE Transactions on Visualization and Computer Graphics* Vol. 24, No. 1, 77–87, 2018.

[52] Sun, M.; Cai, L.; Cui, W.; Wu, Y.; Shi, Y.; Cao, N. Erato: Cooperative data story editing via fact interpolation. *IEEE Transactions on Visualization and Computer Graphics* Vol. 29, No. 1, 983–993, 2023.

[53] Ying, L.; Shu, X.; Deng, D.; Yang, Y.; Tang, T.; Yu, L.; Wu, Y. MetaGlyph: Automatic generation of metaphoric glyph-based visualization. *IEEE Transactions on Visualization and Computer Graphics* Vol. 29, No. 1, 331–341, 2023.

[54] Guo, Y.; Han, Q.; Lou, Y.; Wang, Y.; Liu, C.; Yuan, X. Edit-history vis: An interactive visual exploration and analysis on wikipedia edit history. In: Proceedings of the IEEE 16th Pacific Visualization Symposium, 157–166, 2023.

[55] Tu, Y.; Qiu, R.; Wang, Y. S.; Yen, P. Y.; Shen, H. W. PhraseMap: Attention-based keyphrases recommendation for information seeking. *IEEE Transactions on Visualization and Computer Graphics* Vol. 30, No. 3, 1787–1802, 2024.

[56] Li, X.; Wang, Y.; Wang, H.; Wang, Y.; Zhao, J. NBSearch: Semantic search and visual exploration of computational notebooks. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, Article No. 308, 2021.

[57] Narechania, A.; Karduni, A.; Wesslen, R.; Wall, E. VITALITY: Promoting serendipitous discovery of academic literature with transformers & visual analytics. *IEEE Transactions on Visualization and Computer Graphics* Vol. 28, No. 1, 486–496, 2022.

[58] Shi, C.; Nie, F.; Hu, Y.; Xu, Y.; Chen, L.; Ma, X.; Luo, Q. MedChemLens: An interactive visual tool to support direction selection in interdisciplinary experimental research of medicinal chemistry. *IEEE Transactions on Visualization and Computer Graphics* Vol. 29, No. 1, 63–73, 2023.

[59] Resck, L. E.; Ponciano, J. R.; Nonato, L. G.; Poco, J. LegalVis: Exploring and inferring precedent citations in legal documents. *IEEE Transactions on Visualization and Computer Graphics* Vol. 29, No. 6, 3105–3120, 2023.

[60] Zhang, X.; Engel, J.; Evensen, S.; Li, Y.; Demiralp, Ç.; Tan, W. C. Teddy: A system for interactive review analysis. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, Article No. 108, 2020.

[61] Wu, Y.; Xu, Y.; Gao, S.; Wang, X.; Song, W.; Nie, Z.; Fan, X.; Li, Q. LiveRetro: Visual analytics for strategic retrospect in livestream E-commerce. *IEEE Transactions on Visualization and Computer Graphics* Vol. 30, No. 1, 1117–1127, 2024.

[62] Ouyang, Y.; Wu, Y.; Wang, H.; Zhang, C.; Cheng, F.; Jiang, C.; Jin, L.; Cao, Y.; Li, Q. Leveraging historical medical records as a proxy via multimodal modeling and visualization to enrich medical diagnostic learning. *IEEE Transactions on Visualization and Computer Graphics* Vol. 30, No. 1, 1238–1248, 2024.

[63] Tu, Y.; Li, O.; Wang, J.; Shen, H. W.; Powałko, P.; Tomescu-Dubrow, I.; Slomczynski, K. M.; Blanas, S.; Jenkins, J. C. SDRQuerier: A visual querying framework for cross-national survey data recycling. *IEEE Transactions on Visualization and Computer Graphics* Vol. 29, No. 6, 2862–2874, 2023.

[64] Chen, Z.; Yang, Q.; Shan, J.; Lin, T.; Beyer, J.; Xia, H.; Pfister, H. IBall: Augmenting basketball videos with gaze-moderated embedded visualizations. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, Article No. 841, 2023.

[65] Chen, Z. T.; Yang, Q.; Xie, X.; Beyer, J.; Xia, H.; Wu, Y.; Pfister, H. Sporthesia: Augmenting sports videos using natural language. *IEEE Transactions on Visualization and Computer Graphics* Vol. 29, No. 1, 918–928, 2023.

[66] Tu, Y.; Xu, J.; Shen, H. W. KeywordMap: Attention-based visual exploration for keyword analysis. In: Proceedings of the IEEE 14th Pacific Visualization Symposium, 206–215, 2021.

[67] Liu, C.; Han, Y.; Jiang, R.; Yuan, X. ADVISor: Automatic visualization answer for natural-language question on tabular data. In: Proceedings of the IEEE 14th Pacific Visualization Symposium, 11–20, 2021.

[68] Shen, L.; Zhang, Y.; Zhang, H.; Wang, Y. Data player: Automatic generation of data videos with narration-animation interplay. *IEEE Transactions on Visualization and Computer Graphics* Vol. 30, No. 1, 109–119, 2024.

[69] Xiao, S.; Huang, S.; Lin, Y.; Ye, Y.; Zeng, W. Let the chart spark: Embedding semantic context into chart with text-to-image generative model. *IEEE Transactions on Visualization and Computer Graphics* Vol. 30, No. 1, 284–294, 2024.

[70] Singh, H.; Shekhar, S. STL-CQA: Structure-based transformers with localization and encoding for chart question answering. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 3275–3284, 2020.

[71] Ma, W.; Zhang, H.; Yan, S.; Yao, G.; Huang, Y.; Li, H.; Wu, Y.; Jin, L. Towards an efficient framework for data extraction from chart images. In: *Document Analysis and Recognition – ICDAR 2021. Lecture Notes in Computer Science, Vol. 12821.* Lladós, J.; Lopresti, D.; Uchida, S. Eds. Springer Cham, 583–597, 2021.

[72] Song, S.; Li, C.; Sun, Y.; Wang, C. VividGraph: Learning to extract and redesign network graphs from visualization images. *IEEE Transactions on Visualization and Computer Graphics* Vol. 29, No. 7, 3169–3181, 2023.

[73] Chen, Z. T.; Wang, Y.; Wang, Q.; Wang, Y.; Qu, H. Towards automated infographic design: Deep learning-based auto-extraction of extensible timeline. *IEEE Transactions on Visualization and Computer Graphics* Vol. 26, No. 1, 917–926, 2020.

[74] Sultanum, N.; Srinivasan, A. DATATALES: Investigating the use of large language models for authoring data-driven articles. In: Proceedings of the IEEE Visualization and Visual Analytics, 231–235, 2023.

[75] Liu, C.; Guo, Y.; Yuan, X. AutoTitle: An interactive title generator for visualizations. *IEEE Transactions on Visualization and Computer Graphics* doi: 10.1109/TVCG.2023.3290241, 2023.

[76] Song, S.; Chen, J.; Li, C.; Wang, C. GVQA: Learning to answer questions about graphs with visualizations via knowledge base. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, Article No. 464, 2023.

[77] Adhikary, J.; Vertanen, K. Text entry in virtual environments using speech and a midair keyboard. *IEEE Transactions on Visualization and Computer Graphics* Vol. 27, No. 5, 2648–2658, 2021.

[78] Card, S. K.; Mackinlay, J. D.; Shneiderman, B. *Readings in Information Visualization: Using Vision to Think.* San Francisco, CA, USA: Academic Press, 1999.

[79] Zhou, C.; Li, Q.; Li, C.; Yu, J.; Liu, Y.; Wang, G.; Zhang, K.; Ji, C.; Yan, Q.; He, L.; et al. A comprehensive survey on pretrained foundation models: A history from BERT to ChatGPT. *arXiv preprint* arXiv:2302.09419, 2023.

[80] Chen, Z. T.; Zeng, W.; Yang, Z.; Yu, L.; Fu, C. W.; Qu, H. LassoNet: Deep lasso-selection of 3D point clouds. *IEEE Transactions on Visualization and Computer Graphics* Vol. 26, No. 1, 195–204, 2020.

[81] Ottley, A.; Garnett, R.; Wan, R. Follow the clicks: Learning and anticipating mouse interactions during exploratory data analysis. *Computer Graphics Forum* Vol. 38, No. 3, 41–52, 2019.

[82] Brown, E. T.; Ottley, A.; Zhao, H.; Lin, Q.; Souvenir, R.; Endert, A.; Chang, R. Finding Waldo: Learning about users from their interactions. *IEEE Transactions on Visualization and Computer Graphics* Vol. 20, No. 12, 1663–1672, 2014.

[83] Wexler, J.; Pushkarna, M.; Bolukbasi, T.; Wattenberg, M.; Viégas, F.; Wilson, J. The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics* Vol. 26, No. 1, 56–65, 2020.

[84] Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; Gelly, S. Parameterefficient transfer learning for NLP. In: Proceedings of the 36th International Conference on Machine Learning, 2790–2799, 2019.

[85] Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-rank adaptation of large language models. In: Proceedings of the International Conference on Learning Representations, 2021.

[86] AdapterHub. Available at https://adapterhub.ml/

[87] Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of the 36th Conference on Neural Information Processing Systems, 24824–24837, 2022.

[88] Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* Vol. 21, No. 1, 5485–5551, 2020.

[89] Wang, Y.; Hou, Z.; Shen, L.; Wu, T.; Wang, J.; Huang, H.; Zhang, H.; Zhang, D. Towards natural language-based visualization authoring. *IEEE Transactions on Visualization and Computer Graphics* Vol. 29, No. 1, 1222–1232, 2023.

[90] Schwartz, R.; Dodge, J.; Smith, N. A.; Etzioni, O. Green AI. *Communications of the ACM* Vol. 63, No. 12, 54–63, 2020.

[91] Zhou, C.; Liu, P.; Xu, P.; Lyer, S.; Sun, J.; Mao, Y.; Ma, X.; Efrat, A.; Yu, P.; Yu, L.; et al. LIMA: Less is more for alignment. In: Proceedings of the 37th Conference on Neural Information Processing Systems, 2024.

[92] Zhou, Y.; Yang, W.; Chen, J.; Chen, C.; Shen, Z.; Luo, X.; Yu, L.; Liu, S. Cluster-aware grid layout. *IEEE Transactions on Visualization and Computer Graphics* Vol. 30, No. 1, 240–250, 2024.

[93] Yang, W.; Wang, X.; Lu, J.; Dou, W.; Liu, S. Interactive steering of hierarchical clustering. *IEEE Transactions on Visualization and Computer Graphics* Vol. 27, No. 10, 3953–3967, 2021.

[94] Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, S. G.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint* arXiv:1603.04467, 2016.

[95] Ma, K. L. *In situ* visualization at extreme scale: Challenges and opportunities. *IEEE Computer Graphics and Applications* Vol. 29, No. 6, 14–19, 2009.

[96] Rapp, T.; Peters, C.; Dachsbacher, C. Image-based visualization of large volumetric data using moments. *IEEE Transactions on Visualization and Computer Graphics* Vol. 28, No. 6, 2314–2325, 2022.

[97] Richer, G.; Pister, A.; Abdelaal, M.; Fekete, J. D.; Sedlmair, M.; Weiskopf, D. Scalability in visualization. *IEEE Transactions on Visualization and Computer Graphics* doi: 10.1109/TVCG.2022.3231230, 2022.

[98] Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Wu, Z.; Chang, B.; Sun, X.; Xu, J.; Li, L.; Sui, Z. A survey on in-context learning. *arXiv preprint* arXiv:2301.00234, 2022.

[99] Liu, S.; Xiao, J.; Liu, J.; Wang, X.; Wu, J.; Zhu, J. Visual diagnosis of tree boosting methods. *IEEE Transactions on Visualization and Computer Graphics* Vol. 24, No. 1, 163–173, 2018.

[100] Yuan, J.; Liu, M.; Tian, F.; Liu, S. Visual analysis of neural architecture spaces for summarizing design principles. *IEEE Transactions on Visualization and Computer Graphics* Vol. 29, No. 1, 288–298, 2023.

[101] Khayat, M.; Karimzadeh, M.; Zhao, J.; Ebert, D. S. VASSL: A visual analytics toolkit for social spambot labeling. *IEEE Transactions on Visualization and Computer Graphics* Vol. 26, No. 1, 874–883, 2020.

[102] Bernard, J.; Zeppelzauer, M.; Lehmann, M.; Müller, M.; Sedlmair, M. Towards user-centered active learning algorithms. *Computer Graphics Forum* Vol. 37, No. 3, 121–132, 2018.

[103] Yang, W.; Ye, X.; Zhang, X.; Xiao, L.; Xia, J.; Wang, Z.; Zhu, J.; Pfister, H.; Liu, S. Diagnosing ensemble few-shot classifiers. *IEEE Transactions on Visualization and Computer Graphics* Vol. 28, No. 9, 3292–3306, 2022.

[104] Zhou, Z. H.; Tan, Z. H. Learnware: Small models do big. *Science China Information Sciences* Vol. 67, No. 1, Article No. 112102, 2023.

[105] HuggingFace. Available at https://huggingface.co/models

[106] Wang, Q.; Yuan, J.; Chen, S.; Su, H.; Qu, H.; Liu,

S. Visual genealogy of deep neural networks. *IEEE Transactions on Visualization and Computer Graphics* Vol. 26, No. 11, 3340–3352, 2020.

[107] Cao, K.; Liu, M.; Su, H.; Wu, J.; Zhu, J.; Liu, S. Analyzing the noise robustness of deep neural networks. *IEEE Transactions on Visualization and Computer Graphics* Vol. 27, No. 7, 3289–3304, 2021.

[108] Liu, M.; Liu, S.; Su, H.; Cao, K.; Zhu, J. Analyzing the noise robustness of deep neural networks. In: Proceedings of the IEEE Conference on Visual Analytics Science and Technology, 60–71, 2018.

[109] Qiu, R.; Tu, Y.; Wang, Y. S.; Yen, P. Y.; Shen, H. W. DocFlow: A visual analytics system for question-based document retrieval and categorization. *IEEE Transactions on Visualization and Computer Graphics* Vol. 30, No. 2, 1533–1548, 2024.

[110] Shi, D.; Xu, X.; Sun, F.; Shi, Y.; Cao, N. Calliope: Automatic visual data story generation from a spreadsheet. *IEEE Transactions on Visualization and Computer Graphics* Vol. 27, No. 2, 453–463, 2021.

[111] Chen, Q.; Chen, N.; Shuai, W.; Wu, G.; Xu, Z.; Tong, H.; Cao, N. Calliope-net: Automatic generation of graph data facts via annotated node-link diagrams. *IEEE Transactions on Visualization and Computer Graphics* Vol. 30, No. 1, 562–572, 2024.

[112] Blei D. M.; Ng A. Y.; Jordan, M. I. Latent dirichlet allocation. *Journal of Machine Learning Research* Vol. 3, 993–1022, 2003.

[113] Lowe, D. G. Object recognition from local scale-invariant features. In: Proceedings of the 7th IEEE International Conference on Computer Vision, 1150–1157, 1999.

[114] Rozière, B.; Gehring, J.; Gloeckle, F.; Sootla, S.; Gat, L.; Tan, X. E.; Adi, Y.; Liu, J.; Sauvestre, R.; Remez, T.; et al. Code Llama: Open foundation models for code. *arXiv preprint* arXiv:2308.12950, 2023.

[115] Bostock, M.; Ogievetsky, V.; Heer, J. D^3 Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics* Vol. 17, No. 12, 2301–2309, 2011.

[116] Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing in Science and Engineering* Vol. 9, No. 3, 90–95, 2007.

[117] Kwon, O. H.; Ma, K. L. A deep generative model for graph layout. *IEEE Transactions on Visualization and Computer Graphics* Vol. 26, No. 1, 665–675, 2020.

[118] Zamfirescu-Pereira, J. D.; Wong, R. Y.; Hartmann, B.; Yang, Q. Why johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, Article No. 437, 2023.

[119] Pryzant, R.; Iter, D.; Li, J.; Lee, Y. T.; Zhu, C.; Zeng, M. Automatic prompt optimization with "gradient descent" and beam search. *arXiv preprint* arXiv:2305.03495, 2023.

[120] Jing, Y.; Yang, Y.; Feng, Z.; Ye, J.; Yu, Y.; Song, M. Neural style transfer: A review. *IEEE Transactions on Visualization and Computer Graphics* Vol. 26, No. 11, 3365–3385, 2020.

[121] Abdal, R.; Qin, Y.; Wonka, P. Image2StyleGAN: How to embed images into the StyleGAN latent space? In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 4432–4441, 2019.

[122] Chen, Q.; Cao, S.; Wang, J.; Cao, N. How does automation shape the process of narrative visualization: A survey of tools. *IEEE Transactions on Visualization and Computer Graphics* doi: 10.1109/TVCG.2023.3261320, 2023.

[123] Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; Parikh, D. VQA: Visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision, 2425–2433, 2015.

[124] Anil, R.; Dai, A. M.; Firat, O.; Johnson, M.; Lepikhin, D.; Passos, A.; Shakeri, S.; Taropa, E.; Bailey, P.; Chen, Z.; et al. PaLM 2 technical report. *arXiv preprint* arXiv:2305.10403, 2023.

[125] Zhao, Y.; Jiang, H.; Chen, Q. A.; Qin, Y.; Xie, H.; Wu, Y.; Liu, S.; Zhou, Z.; Xia, J.; Zhou, F. Preserving minority structures in graph sampling. *IEEE Transactions on Visualization and Computer Graphics* Vol. 27, No. 2, 1698–1708, 2021.

[126] Yuan, J.; Xiang, S.; Xia, J.; Yu, L.; Liu, S. Evaluation of sampling methods for scatterplots. *IEEE Transactions on Visualization and Computer Graphics* Vol. 27, No. 2, 1720–1730, 2021.

[127] Pan, X.; Tewari, A.; Leimkühler, T.; Liu, L.; Meka, A.; Theobalt, C. Drag your GAN: Interactive point-based manipulation on the generative image manifold. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference, Article No. 78, 2023.

[128] Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; et al. A survey on large language model based autonomous agents. *arXiv preprint* arXiv:2308.11432, 2023.

[129] Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; Bernstein, M. S. Generative agents: Interactive simulacra of human behavior. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, Article No. 2, 2023.

Weikai Yang is a Ph.D. candidate at Tsinghua University. His research interests include visual text analytics and interactive machine learning. He received his B.S. degree from Tsinghua University.

Mengchen Liu is a senior researcher at Microsoft. His research interests include explainable AI and computer vision. He received his B.S. degree in electronics engineering and his Ph.D. degree in computer science from Tsinghua University. He has served as a PC member and reviewer for various conferences and journals.

Zheng Wang is currently working toward a graduate degree at Tsinghua University.

Shixia Liu is a professor at Tsinghua University. Her research interests include visual text analytics, visual social analytics, interactive machine learning, and text mining. She worked as a research staff member at IBM China Research Lab and a lead researcher at Microsoft Research Asia. She received her B.S. and M.S. degrees from Harbin Institute of Technology, her Ph.D. degree from Tsinghua University. She is a fellow of IEEE and an associate editor-in-chief of *IEEE Trans. Vis. Comput. Graph.*