A Review and Analysis of Evaluation Practices in VIS Domain Applications

Yiwen Xing, Gabriel D. Cantareira, Rita Borgo, and Alfie Abdul-Rahman

*Abstract***—This article presents a review and analysis of evaluation practices within the visualization and visual analytics (VIS) domain, with a focus on domain application work accepted at the IEEE VIS conference from 2018 to 2022. Through the analysis of 140 pertinent article, we establish a detailed classification principle for evaluation practices, using the *Who, When, What,* and *How* indicators. This principle covers facets such as analysis methods, targets, scenarios, participant expertise, and stages of occurrence. By systematically categorizing the application domains presented in these works, we apply our established classification principle to discern and categorize the evaluation practices within them, identifying the prevailing characteristics and trends. The article explores the variety of evaluation methods employed across different application domains and observes the distinctions in their usage. In conclusion, we provide insights and highlight concerns for conducting evaluations in upcoming domain application research. Our findings are intended to inform and guide subsequent studies in a similar context.**

*Index Terms***—Evaluation methods, visualization application.**

I. Introduction

Visualization and visual analytics (VIS) play an instrumental role in various domains, assisting analysts in interpreting and navigating complex data and phenomena. With the escalating demand for domain-specific visual solutions, there is an expanding body of VIS application research dedicated to devising visualizations and visual analytics applications specifically tailored to meet distinct domain challenges. Beyond the core application, the evaluation process is a critical component. Such evaluations are essential in validating domain requirements, tasks, and design choices, ensuring that the developed applications resonate with the set objectives and stand out in terms of their usability and utility. Evaluation is indispensable throughout the design and development phases of applications. As the landscape of visualization broadens, collaborating with professionals from diverse domains, each bearing distinct characteristics, may naturally lead to varied evaluation practices. This diversification is further amplified as the contemporary VIS application research landscape evolves, incorporating trends like the integration of Artificial Intelligence (AI) algorithms and immersive technologies. Given this backdrop, we are keen to explore the nuances in evaluation practices across application domains and how emerging trends might shape these practices, thereby offering a timely review of application papers from the last five years accepted at the IEEE VIS conference.

Given the pivotal role of evaluation in VIS applications, it becomes imperative to observe and comprehend the prevailing evaluation practices and trends across various application domains. Extracting insights from these trends can aid in sidestepping potential pitfalls, paving the way for a more streamlined evaluation blueprint for future researchers conducting research in a similar context. This realization was the catalyst behind our decision to dive into these evaluation practices. Consequently, this paper embarks on a journey to examine the evaluation practices dominant in VIS domain application research. Our objective is twofold: 1) to offer a comprehensive perspective on contemporary evaluation methodologies; and 2) to provide insights that can guide future research in similar contexts.

To achieve our objectives, we adopted a systematic review methodology [1]. Our approach was anchored by a classification principle, demarcated by four indicators: *Who, When, What,* and *How*. These indicators were instrumental in delineating the evaluation practices, allowing for a meticulous categorization of evaluation sessions. Following a rigorous paper selection process, we delved into understanding the interconnections, correlations, and disparities among the indicators, as well as the distinctions spanning various application domains. Our exploration encompassed the identification of trends, the characterization of unique features, and the extraction of dominant evaluation methodologies.

In conclusion, this paper provides a detailed analysis of evaluation practices in the VIS application domain over the past five years. Our primary contribution lies in the establishment of a classification principle, enabling a structured review of evaluation sessions. We also identified and discussed the relationships and differences among the evaluation practices in various application domains. Our discussion aims to inspire future research directions in evaluation methodologies within the VIS application domain. Through this work, we aspire to contribute a reference that supports the advancement of evaluation practices in VIS domain application research.

Received 9 October 2023; revised 12 July 2024; accepted 9 September 2024. Date of publication 13 September 2024; date of current version 1 August 2025. This work was supported in part by King's-China Scholarship Council PhD Scholarship programme (K-CSC) and in part by EPSRC under Grant EP/V028871/1. Recommended for acceptance by R. Chang. *(Corresponding author: Yiwen Xing.)*

The authors are with King's College London, WC2R 2LS London, U.K. (e-mail: yiwen.xing@kcl.ac.uk; gabriel.dias_cantareira@kcl.ac.uk; rita.borgo@kcl.ac.uk; alfie.abdulrahman@kcl.ac.uk).

All information about the reviewed article, the coding results, and the detailed data analysis process can be accessed at https://github.com/narrating-complexity/vis_evaluation_analysis.

Digital Object Identifier 10.1109/TVCG.2024.3460181

II. Background and Related Work

This section offers an overview of prior research on evaluation practices within the visualization domain, with a focus on visualization applications and emerging trends in VIS.

A. Evaluation in Visualization

Evaluating visualization techniques and applications holds critical importance in the VIS domain. Over the years, various challenges have been identified in conducting evaluations [2], [3], [4], leading to the development of a range of objectives, scenarios, and methods. As a case in point, Lam et al. [5] introduced seven guiding scenarios specifically tailored for information visualization evaluation. Subsequently, this foundational framework was broadened by Isenberg et al. [6] to encompass the wider spectrum of the VIS domain.

The evaluation studies in VIS have primarily focused on the usability and effectiveness of specific visualization techniques, applications, and systems [7], [8]. There is also a growing emphasis on evaluating user experience goals, which has encouraged further exploration of evaluation practices within this context [9]. Additionally, gaining insights is increasingly recognized as an important aspect of visualization evaluation [10], [11], [12]. Stasko [13] proposed a value-driven evaluation framework, highlighting a system's value through four capabilities: answering questions effectively, fostering insight generation, conveying data essence, and building confidence and knowledge about the data's domain and context.

The VIS community employs a diverse range of evaluation methods. Quantitative and qualitative methods are used, including surveys, interviews, focus groups, case studies, and controlled experiments. Carpendale [2] provided an overview of various empirical evaluation methods, discussing their advantages and disadvantages. Tory and Möller [14] highlighted the benefits of expert reviews and evaluations involving domain experts, while North [10] proposed an insight-oriented technique using open-ended protocols.

Drawing parallels with other disciplines, the evaluation practices in VIS reveal a blend of commonalities and distinct characteristics. Crisan and Elliott [15] note that while evaluative practices in VIS are insightful, they seldom undergo comparative analysis with practices in fields like psychology or sociology. This underscores the potential for a more integrative approach, drawing parallels and distinctions between VIS and other empirical disciplines. Despite the extensive body of research dedicated to dissecting, consolidating, and categorizing evaluation practices in the VIS domain, there remains a conspicuous void. Specifically, there is a dearth of studies that perform a comparative analysis of evaluation practices across varied VIS application domains. While certain reviews, such as those focusing on medical visualization [16], offer domain-specific insights, holistic cross-domain comparisons are still few and far between.

B. Visualization Domain Applications and Design Study

In recent years, there has been a growing interest in the applications of data visualization in various industries and domains. Recognizing the need for tailored designs, domain-specific analysis techniques, such as Domain Analysis for Data Visualization (DADV), have emerged [17]. Weber et al. [18] examined the evolving trends in visualization application papers, discussing the criteria for assessing application papers and emphasizing the growing engagement in interdisciplinary research. Their work underscores the importance of assessing how these applications meet cross-disciplinary challenges, reflecting visualization's growing complexity and diversity in practical contexts.

The IEEE VIS community has been at the forefront of this movement, championing the integration of visualization across diverse domains. They not only encourage discussions centered on tangible real-world visualization applications, but also prominently feature "Applications" as one of the five core areas in VIS [19]. Overall, VIS application domain focuses on developing effective visualization techniques that can be applied in different domains to support a variety of data analysis tasks.

Many visualization applications are recognized as outputs from design studies, aligning with the definition by Sedlmair et al. [20]: "a project in which visualization researchers analyze a specific real-world problem faced by domain experts, design a visualization system that supports solving this problem, validate the design and reflect about lessons learned in order to refine visualization design guidelines." These studies are essential to integrating domain knowledge and active collaboration with domain experts. Additionally, various frameworks and models have been developed to provide guidance and support rigorous design studies [20], [21], [22], [23]. Viewing design as a method of inquiry and considering its diverse, subjective, and socially constructed contributions, validating activities during the design process has gained importance [24], [25]. Historically, design studies and related evaluations have focused mainly on deployed, operational applications. However, there is increasing recognition of the value of leveraging complex and nuanced learning through deep engagement with people, data, and technology.

C. New Trends in Visualization Applications

Recent advancements have seen visual analytics technologies increasingly augmented with Artificial Intelligence (AI) techniques and immersive approaches. The interplay between AI and VIS unfolds in two key areas [26]: AI for VIS (AI4VIS) using AI technologies in visualization applications to enhance data analysis and visualization capabilities [27], and VIS for AI (VIS4AI) focuses on improving the explainability of AI models through visualization techniques [28], thus establishing AI as an application domain within VIS, comparable to traditional domains. Do emerging trends carry unique properties that may impact evaluation practices? For instance, AI-enhanced VIS applications may require assessments centered on algorithmic performance, while immersive technologies call for evaluations prioritizing user experience in three-dimensional contexts. This review briefly investigates these trends' effects on evaluation methods, highlighting areas for future in-depth research.

III. Methodology and Approach

To obtain a systematic understanding of the current state of evaluation in visualization domain application research, we performed a qualitative literature review. Qualitative literature reviews [29] are common methods in various scientific fields, employed to summarize and analyze existing studies to gain insights into specific topics, phenomena, or issues. The methodology and approach of our literature review are detailed in the subsequent section.

A. Choice of Literature

Visualization, with its inherent applicability and propensity for interdisciplinary collaboration, possesses a unique character. Research focused on domain applications is particularly prominent within the VIS domain due to these attributes. Evaluative activities, especially those that involve human participants, are pivotal in such research, acting as a gold standard for gauging the quality of the research output. This understanding formed the foundation for our emphasis on domain application research. To understand recent evaluation practices in visualization domain application research, we began with all 620 full papers accepted at the IEEE VIS: Visualization & Visual Analytics Conference over the past five years (2018-2022). This venue was selected due to its prestigious status as an annual conference focused on scientific visualization, information visualization, and visual analytics. The chosen timeframe of five years is intended to capture emerging trends. In the area model of the IEEE VIS [19], a specific category is designated for application research, aligning closely with the theme of our review. Although this model primarily aids authors during submission and does not enable the direct retrieval of papers classified under the "Application" category from published works, it still serves as a valuable guide for our study. Consulting the area model, we initiated the screening process by conducting keyword searches within the abstracts of these articles. We adopted a targeted approach to pinpoint research emphasizing applications, insisting that these studies not only engage with specific domains but also incorporate evaluation practices. From this procedure, 303 application-centric papers were identified that fulfilled both domain involvement and evaluation criteria, forming our preliminary dataset. This collection underwent further manual iterative filtering and coding, narrowing to a refined set of 140 articles designated for detailed examination. The methodology behind our screening, including keywords used and filtering rationale, is detailed in Fig. 1.

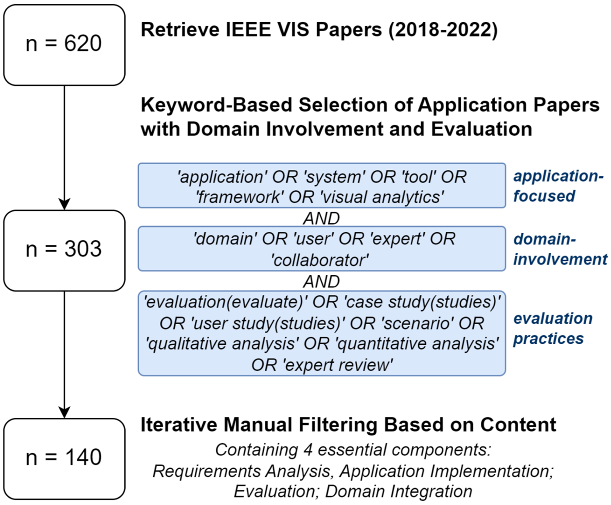
n = 620
Retrieve IEEE VIS Papers (2018-2022)
Keyword-Based Selection of Application Papers with Domain Involvement and Evaluation
'application' OR 'system' OR 'tool' OR 'framework' OR 'visual analytics'
application-focused
AND
n = 303
'domain' OR 'user' OR 'expert' OR 'collaborator'
domain-involvement
AND
'evaluation(evaluate)' OR 'case study(studies)' OR 'user study(studies)' OR 'scenario' OR 'qualitative analysis' OR 'quantitative analysis' OR 'expert review'
evaluation practices
n = 140
Iterative Manual Filtering Based on Content
Containing 4 essential components:
Requirements Analysis, Application Implementation; Evaluation; Domain Integration

Fig. 1. The literature screening process.

As a primary outcome of a design study, VIS application research encompasses key components such as conducting domain requirements analysis, implementing visualization applications, and evaluating these applications, which are subsequently reported in scholarly articles [20]. Our review focuses explicitly on domain-focused applications, where we assessed the presence of domain integration within the papers. The criteria for paper selection were iteratively refined during the manual coding process. It is important to emphasize that our selection was based exclusively on the content presented within the papers. This approach might inadvertently omit some studies that fulfilled our criteria but did not explicitly detail them in their publications. The filters we ultimately employed are as follows:

- *Requirements Analysis:* This criterion assesses whether the paper addresses domain requirements or task analysis. It encompasses refined requirements derived from interactions with domain experts, challenges identified through extensive literature reviews, and the exploration of the design space by pinpointing gaps in existing solutions. The goal is to ensure that the research is grounded in real-world needs.
- *Application Implementation:* This filter evaluates whether the paper provides a detailed introduction to the application's implementation. It looks for a detailed presentation and description of the application, be it a system, or a framework. The emphasis is on understanding the technical underpinnings, design choices, and the overall functionality of the developed application.
- *Evaluation:* This filter examines if the paper introduces any form of evaluation. An evaluation, in this context, refers to a set of designed evaluation practices crafted for what needs to be assessed. It can manifest itself as a combination of diverse evaluation activities or can be presented in the form of a case study. The essence is to determine the rigor and depth of the evaluation process undertaken in the research.
- *Domain Integration:* This criterion assesses the extent of domain integration within the paper. For quintessential domain application articles, specific domain integration is vital. However, for data-driven or domain-agnostic articles, inclusion in our scope depends on the application's presence in one or multiple domains, especially during evaluation. The goal is to ensure that the research is both theoretically robust and practically applicable in real-world domains.

B. Choice of Codes

Within the scope of VIS application research, it is common for a study to include multiple evaluation sessions, each designed with precision to validate specific objectives and content. For example, evaluation activities during the requirements analysis phase often focus on gaining insights and involving domain knowledge. In contrast, post-deployment evaluations might concentrate more on assessing the application's usability. During our review process, we aimed to distinctly categorize different evaluations to explore the nuances in evaluation practices across various sessions. To achieve this, we adopted methods from other fields to define events precisely. Originating from the field of journalism, the "Five Ws and How" framework offers a comprehensive approach to capturing the essence of any event [30]. This approach guarantees a thorough examination of all pivotal facets of an event, offering a comprehensive perspective. Mirroring practices in User-Centered Design [31], this framework has found frequent application in framing evaluation sessions (events), as detailed below, focusing solely on three of the original five Ws.

- *Who:* Refers to the participant, pinpointing the target demographic for the evaluation.
- *When:* Explores the context or timing, identifying the specific phase of the project or research during which the evaluation takes place.
- *What:* Centers on the specific task or objective of the evaluation, elucidating what precisely is under assessment or validation.
- *How:* Encompasses the methods, techniques, or approaches employed for the evaluation.

Given this framework, we define an evaluation session (event) in each individual research based on the *Who*, *When*, and *What* indicators. We consider *a set of evaluation practices conducted with the same or different types of participants, during a specific phase of project development, and aimed at a particular evaluation task* as a singular evaluation session. Within each of these sessions, the specific evaluation methods (*How*) are central to our data coding and analysis, derived directly from the content of the literature.

Broadly, our coding scheme bifurcates into two segments:

Basic Information and Filter Codes: This encompasses the collection of fundamental details about the paper, including its year of acceptance, application domain, and data-driven nature. Additionally, it includes the four filters previously mentioned in Section III-A – the integration of requirements analysis, implementation, evaluation, and domain. Recognizing the importance of certain nuances, we also extracted information on the provision of an evaluation protocol and the involvement of domain experts during the requirement and task analysis stages. Regarding the emerging trends and their potential impact on evaluation practices, we categorized papers based on their alignment with AI4VIS and VIS4AI, or their focus on applications integrated with immersive techniques.

Evaluation Practices Information and Codes: This segment delves into the specifics of the evaluation practices adopted in the papers, branching out from the four indicators mentioned above. For the *Who* indicator, we established two categories: domain expertise and number of participants. Referencing Burns et al.'s [32] taxonomy of participant roles, we categorize participants by their expertise in the application domain. A separate code was also assigned to VIS or UX researchers who participated in the evaluations. Under the *When* indicator, considering the core stages in a design study [20], we broadly segmented the application development process into three milestones: design, development, and deployment. We documented the phase during which various evaluation sessions transpired in the papers, noting if any iterative evaluations were followed up with tangible refinements and improvements. Additionally, preliminary user studies in the early stages were also considered as evaluation. For the *What* indicator, we adopted the eight VIS evaluation scenarios proposed by Isenberg et al. [6], which were initially derived from the seven scenarios presented by Lam et al. [5]. During our coding process, we expanded this list with two additional codes: ESG (Evaluating Scalability and Generalizability) and EUV (Extended User Validation). Under the *How* indicator, we listed a variety of methods that are frequently employed in both qualitative and quantitative evaluation studies.

A detailed enumeration of our codes and their specific meanings is provided in Table I.

C. Coding Procedure

Four coders, all co-authors, were involved in the literature assessment process. Initially, we held a brainstorming session to discuss paper filtering criteria, key evaluation factors, and coding methodology. During our kickoff meeting, we conducted a trial coding exercise with four randomly selected papers to align our perspectives and ensure understanding. Subsequently, two of the authors took the lead in coding and categorizing all 303 papers. To assess inter-coder reliability, we randomly selected 20% of the papers (n=61) and had the two lead coders independently code these papers. The agreement rate was calculated as the proportion of matching coding items to the total number of coding items using the formula

$$\text{Agreement Rate} = \left(\frac{\text{Number of Matching Items}}{\text{Total Number of Items}}\right) \times 100\%.$$

This resulted in a 97.25% agreement rate, indicating robust reliability. In cases of coding disagreements or uncertainties, the issues were brought up in the weekly collaborative meetings for discussion, where the other two coders intervened. The four coders reached a consensus on the results for each paper being discussed through these discussions. The discussion meetings also served as platforms for iterative calibration, allowing us to refine our criteria continuously. After each meeting, we revisited and updated previously coded papers to maintain uniformity and accuracy across our dataset. This rigorous approach ensured our coding methodology was both robust and adaptable. Coding was conducted using Microsoft Excel.

D. Analysis Methods

Upon completing the coding process, we embarked on a comprehensive analysis and discussion of the data obtained.

- *Trend Analysis:* To capture the evolution of evaluation practices, we used line graphs and other time-series charts. These charts enabled us to effectively track and showcase the characteristics and trends of the data over the years.

TABLE I
A List of Codes With Descriptions, Categorized by Four Key Indicators That Characterize the Evaluation Process

	Who
Domain expert	Participant with a high level of knowledge and understanding of the domain.
Domain user	Participant with medium or lower level of domain knowledge.
General public	Participant with no background knowledge about the domain.
Visual expert	Expert or researcher with visualization, design, and UX background.
No participant	Evaluation conducted with no participant involved.
	When
Before/during design stage	Evaluations conducted during the initial design conceptualization, focusing on aligning the design with domain requirements.
During development stage	Evaluations conducted as the visualization is developed and refined, involving iterative testing and feedback cycles.
Other pre-deployment	Evaluations conducted at any point before the deployment but do not specify the exact pre-deployment stage.
After deployment	Evaluations conducted after the visualization is deployed or used in its intended environment without further refinement.
	What *The first eight codes are derived from the evaluation scenarios outlined in [6]. The last two are our new additions.*
UWP	Understanding Environments and Work Practices
VDAR	Visual Data Analysis and Reasoning
CTV	Evaluating Communication Through Visualization
CDA	Evaluating Collaborative Data Analysis
UP	User Performance
UE	User Experience
AP	Algorithm Performance
QRI	Qualitative Result Inspection
ESG	Evaluating Scalability and Generalizability: This category evaluates an application's scalability and generalizability, both crucial for assessing performance across varied conditions. Scalability focuses on the application's ability to manage increasing data volumes or complexity without losing performance or user experience, while generalizability examines its effectiveness across different data types and user needs. Both require testing the application on larger or more diverse datasets than those initially used for development to measure how well it adapts to extended datasets in real-world scenarios.
EUV	Extended User Validation: This category centers on validating if an application, crafted in collaboration with specific domain experts, can meet the broader needs of analysts within that same domain.
	How
Observation *(Obs)*	Observing analysts in their natural environment to gain insight into their behavior and how they interact with visualization systems in context.
Workshop *(Wshp)*	Engaging stakeholders collaboratively in the design and refinement of visualization applications, facilitating brainstorming and feedback.
Focus group *(FGrp)*	Collecting qualitative feedback on their experiences and challenges with a visualization application through open-ended discussions.
Think-aloud protocol *(Tap)*	Participants verbalize their thoughts during interaction with a visualization, revealing their decision-making and experience.
Interview *(Int)*	Conducting structured or semi-structured discussions with experts to understand their experiences and insights with a visualization system.
Survey&Questionnaire *(SnQ)*	Using quantitative and qualitative questions to gauge satisfaction and gather feedback on a visualization system.
Qualitative-task-based evaluation *(Tsk_Ql)*	Assessing a visualization system's capability to support specific tasks, often measuring satisfaction through qualitative feedback.
Interaction with the tool *(IwT)*	Determining if participants directly interact with the application or only receive visual representations such as screenshots or demos.
Content analysis *(CntAn)*	Examining textual data, such as feedback or comments, to discern patterns or themes related to user experience.
Controlled experiments *(CExp)*	Comparing a visualization system's performance to baseline or control, measuring aspects like accuracy or completion time.
A/B testing *(AB)*	Comparing two versions of a visualization system to determine effectiveness, often through metrics such as satisfaction.
Performance evaluation *(PerfEval)*	Assessing a visualization system's speed and efficiency, measuring aspects like rendering speed or memory usage.
Quantitative-task-based evaluation *(Tsk_Qn)*	Evaluating a visualization system's support for specific tasks, focusing on quantitative measures such as task completion time.
User behavior analysis *(UsrBA)*	Analyzing interactions, such as clicks or keystrokes, to understand their engagement with a visualization system.
Case study *(CStudy)*	A detailed examination of a real-world scenario using visualization, with domain experts providing insights on the solution's effectiveness.
Expert feedback *(ExpFdbk)*	Gathering domain expert opinions as part of the evaluation, often through informal discussions, without specifying the exact method of collection.

- *Correlation Analysis:* To explore the relationships between various factors, such as the interplay among the five evaluation indicators or the correlation between domains and evaluation practices, we calculated the correlation coefficient. This was then visualized using heatmaps, providing an intuitive representation that bolstered our analytical insights.
- *Association Analysis:* To discern patterns of co-occurrence, for instance, which evaluation scenarios or objectives frequently pair with specific evaluation methods, we turned to Association Rules [33], [34]. By calculating frequent itemsets and association coefficients, we were able to substantiate our observations and hypotheses about prevalent combinations in evaluation practices.
- *Differential Analysis:* To ascertain disparities, such as the potential variations in evaluation practices across different application domains, we initiated our analysis with visual aids like histograms and heatmaps. These visual aids provided an initial overview of the distribution and trends. Depending on the nature of the data and variable types, we then employed statistical analysis methods like Logistic Regression [35] with L1 (Lasso) regularization [36] to explore significant differences.

Building on data visualization and statistical analysis, we explored the nuances of evaluation practices using Jupyter Notebook. Our focus spanned across application domains, tracking the evolution of practices over time. We identified patterns, emerging trends, and intriguing phenomena that warrant further consideration and exploration in the VIS domain application research.

IV. Results and Discussion

In this section, we present and discuss the findings derived from our coding process. Our analysis emphasizes four primary dimensions: temporal trends, variations in evaluation practices across different domains, the relationship between domains and evaluation practices, and the connections both within and between evaluation objectives and methods.

A. VIS Domain Application Research Over Time

Fig. 2 offers an overview of the distribution of application domains, categorized by year. The category with the highest stacked bar comprises data-driven papers. This prominence does not necessarily signify their standalone importance but becomes meaningful when compared collectively against the sum of typical domain-specific application papers. These data-driven works, focusing on specific datasets and challenges, also involve domain expert evaluations and integrate domain-specific insights, aligning with our article selection criteria. Hence, we draw a deliberate distinction between these and projects directly addressing domain-specific problems for analytical clarity. Setting aside data-driven articles to focus on domain-specific applications reveals that AI-related domains – works aimed at enhancing AI model explainability or supporting AI research (predominantly VIS4AI) – dominate the research landscape over the observed period. This trend consistently places AI-related work as the most popular domain annually, peaking between 2018 and 2021 before a slight decrease in 2022. The healthcare and medical imaging domain emerges as the second most prominent area in VIS application research, initially experiencing a decline in accepted articles by 2019 yet stabilizing in subsequent years. This trend may reflect the initial impact of the COVID-19 pandemic, followed by sustained research engagement. Other domains like journalism & media, sports & entertainment, and biology & life science have also maintained steady attention over the years. The result generally indicates the diversity and breadth of VIS application research domains.

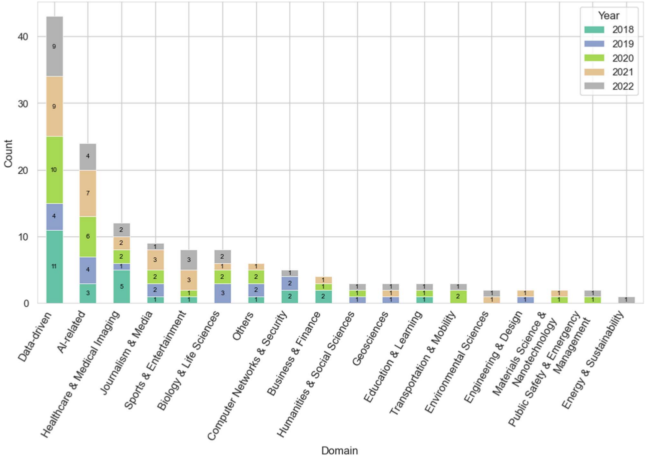

Fig. 2. The distribution of application domains by year.

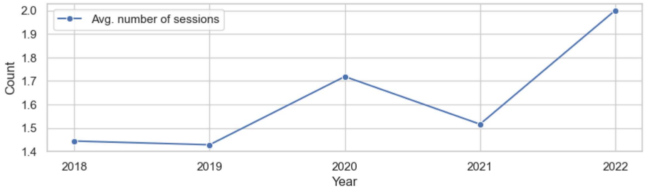

Fig. 3. Yearly average of evaluation sessions per paper.

Using the three indicators: *Who*, *When*, and *What*, we differentiated between various evaluation sessions within a single research study. Each session was then individually coded according to the methods used (the *How* indicator). Of the 140 articles that we examined, a total of 224 evaluation sessions were conducted. By calculating the average number of sessions documented per paper, we observed a clear upward trend in evaluation sessions over the five-year span (Fig. 3). This trend indicates a growing emphasis among researchers on conducting and documenting evaluations for different purposes or at different stages. It also signifies that evaluations in VIS domain application research are becoming more systematic and distinctly categorized. However, when exploring whether there have been significant changes in evaluation practices over the past five years, we inferred from Figs. 4, 5, and 6 that the *Who*, *When*, and *What* indicators have remained relatively consistent over time, without any pronounced trends. This suggests that researchers have established patterns in their evaluation objectives, methods,

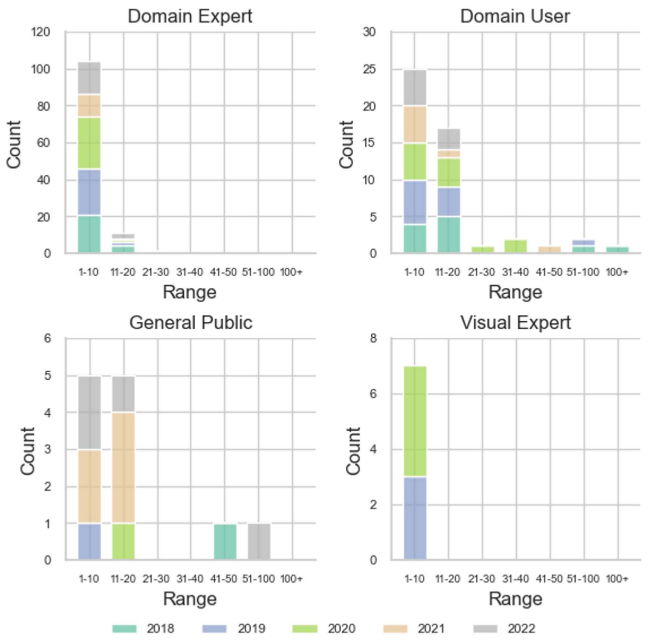

Fig. 4. Stacked bar chart depicting the count of four distinct participant types, with time represented as stacked layers within each bar. *Note: Each chart has a different y-scale.*

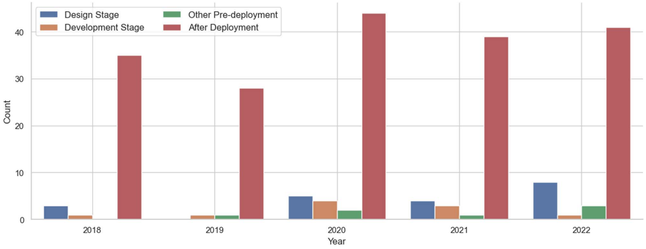

Fig. 5. Bar chart illustrating the frequency of evaluation sessions conducted across various stages over the years.

stages, and participant compositions, with minimal variations over the years.

B. Evaluation in VIS Domain Application Research

Given that evaluation is the central focus of this paper, we delve into the patterns of evaluation in VIS domain application research. Our exploration is structured around the four pivotal elements of evaluation: *Who*, *When*, *What*, and *How*.

1) Who: Concerning the involvement of participants in the evaluation, we scrutinized both their expertise and their numbers. Turning our attention once again to Fig. 4, a discernible trend emerges: as the level of domain expertise decreases (from *Domain Expert* to *Domain User* to *General Public*), the number of participants increases. This observation may suggest that expertise can compensate for a smaller sample size, however, the expertise and number of participants are influenced by many factors such as the depth and duration of their involvement, time and financial costs, and the objects of the evaluation. Furthermore, we observed a decline in the participation of visual experts over the past three years compared to the preceding two. Visual experts typically assess the usability of an application and the efficacy of its visual design. Several hypotheses might explain this reduction in visual expert involvement:

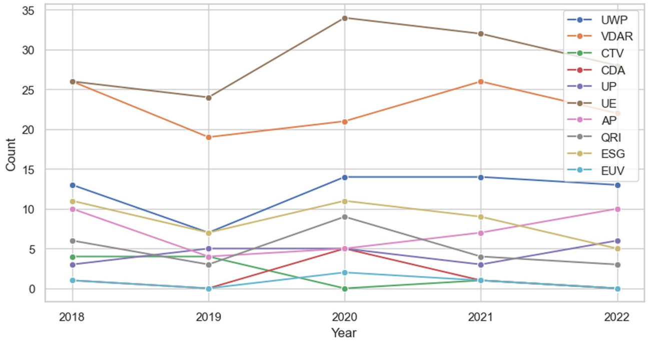

Fig. 6. Multi-line chart illustrating the count of distinct evaluation scenarios across the years.

Priority of Domain Expertise: When there is a conflict between visual efficiency and domain-specific objectives, the preferences of domain experts often take precedence. Adjustments to the visualization design are made accordingly.

Self-evaluation: Many researchers in this field have a background in visualization, which allows them to self-assess and refine the visual aspects of their work to some extent.

Prototyping Focus: When the application is in a prototype form, its primary goal is to meet domain-specific requirements. Unlike finished products, which might require UX (User Experience) experts to evaluate usability rigorously, prototypes prioritize domain needs over polished user experience.

2) When: Regarding the typical stages of evaluation, we observed that the vast majority of articles conducted evaluations only after the application was fully developed. As illustrated in Fig. 5, the number of evaluations categorized under *After Development* significantly outnumbers other categories.

From a procedural perspective, we believe that in VIS domain application research, evaluations should be multifaceted, iterative, and accompanied by refinements. This ensures the rigor of the research process–from the collection and analysis of domain requirements to the selection of design alternatives, and finally, to the assessment of the application's usability and utility. However, from an outcome standpoint, conducting evaluations after the application's deployment and receiving positive feedback can be indicative of the research's success and significance.

One possible reason for this trend could be the constraints on article length. Authors might prioritize detailing the final evaluation based on the completed application, potentially due to space limitations. This focus might inadvertently overshadow the presentation of earlier evaluations in the research process (excluding design study articles that emphasize the design process). While we argue that introducing evaluations from the initial and intermediate stages can enhance the article's persuasiveness, credibility, and reproducibility, we also recognize that each article has its unique focus. Including such information, where space permits, would undoubtedly be a commendable choice.

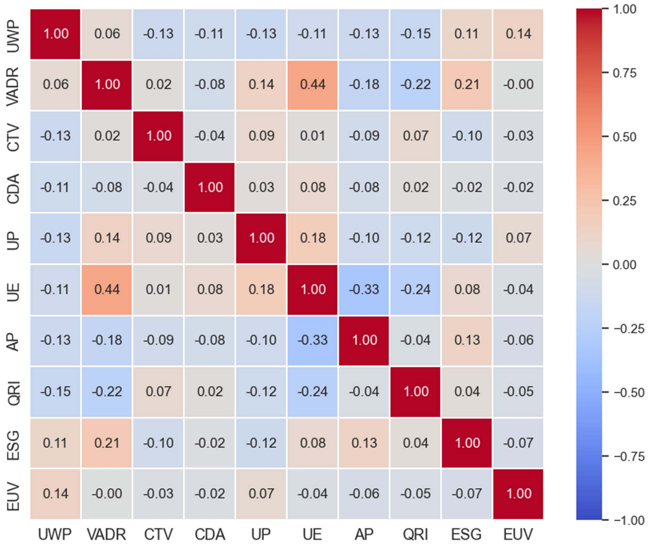

Fig. 7. Heatmap depicting the relationships between different evaluation scenarios.

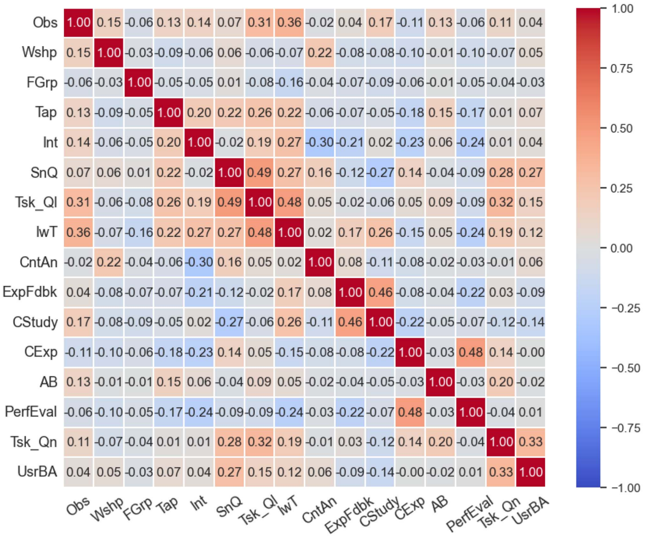

Fig. 8. Heatmap depicting the relationships between different evaluation methods.

3) What: Regarding the objectives of the evaluation, we delved deeper based on the summarized evaluation scenarios [6], focusing solely on VIS domain application research. During the coding process, we encountered two scenarios that could not be directly adopted from the previous classifications, leading us to introduce two new codes: 1) **ESG** (Evaluating Scalability and Generalizability), which targets the evaluation of an application's capability to efficiently manage and represent escalating data volumes or complexity without notable performance or user experience degradation. It also assesses the adaptability of a visualization application across varied use cases, gauging its proficiency in accommodating different data types, domain prerequisites, and user necessities, and 2) **EUV** (Extended User Validation), which concentrates on determining if an application, crafted in tight collaboration with a select group of domain experts, caters to the broader needs of users within the same domain. Following Isenberg et al.'s approach [6], we plotted a line chart (Fig. 6). The results indicate that UE (User Experience) and VDAR (Visual Data Analysis and Reasoning) dominate the evaluation objectives. This contrasts with prior studies on broader VIS research [6], where QRI (Qualitative Result Inspection) and AP (Algorithm Performance) were the primary focuses, highlighting the unique nature of domain application research within the VIS field. The prominence of UE and VDAR aligns with the emphasis on assessing the usability of applications and their effectiveness in meeting domain-specific requirements.

In addition to comparing with the results from previous studies, we also explored the interrelationships between different evaluation scenarios within the *What* indicator (Fig. 7). Echoing our earlier findings, we observed a strong correlation between UE and VDAR, with a coefficient of 0.44. The correlation between ESG and VDAR came in second, registering at 0.21. Based on our coding notes, we attribute this correlation primarily to the shared evaluation practices observed in the "Data-Driven with Domain Integration" category of articles. Specifically, articles in this category often undergo evaluations related to scalability or generalizability. Given their inherent domain integration feature, these studies also typically involve domain experts or users in assessing the achievement of requirements and evaluating how the application aids them in their data analysis endeavors.

We also employed association rules [34] to identify frequent item(sets) that commonly appear together. We utilized the *support* metric, a statistical measure that quantifies the regularity with which specific combinations of items co-occur in a dataset. From our findings, four item(sets) emerged prominently with $support \geq 0.25$: UE (0.64), VDAR (0.51), the combination of UE & VDAR (0.43), and UWP (0.27). These results underscore that, besides the individual and combined emphasis on UE and VDAR, UWP (Understanding Environments and Work Practices) also garners significant attention in evaluations in VIS domain application research. This focus on UWP is pivotal as it directly influences domain requirements analysis, ensuring a comprehensive grasp of the domain in question.

4) How: Regarding the *How* indicator, the factors primarily encompass the quantitative or qualitative evaluation methods that we observed in the examined articles. Naturally, these methods are intrinsically linked to the determinants of evaluation: the execution time, the category of participants, and most crucially, the evaluation objectives. Our analysis unfolds in two segments: first, we look into the interrelationships among the methods themselves, and subsequently, we explore the associations between the objectives and methods, with the former serving as the premise.

The interrelationships among evaluation methods are meticulously illustrated in Fig. 8. By broadly categorizing the methods into qualitative and quantitative data collection and analysis techniques, it is evident that commonly used qualitative methods such as observation, think-aloud protocol, interview, survey & questionnaire, and qualitative task-based evaluation are highly correlated. This suggests that these techniques often coexist within a single evaluation session to achieve a comprehensive feedback collection. Notably, there is a slight negative correlation between interviews and surveys & questionnaires, hinting at a potential substitutive relationship between the two. For instance, when conditions preclude face-to-face interviews, disseminating questionnaires emerges as a compensatory measure. However, given the absence of direct interaction in surveys, they often need to be supplemented with other methods to make up for the lack of face-to-face communication. For quantitative analysis, there is a notable correlation between controlled experiments and performance evaluation (0.48). This phenomenon is attributed to many domain application works that incorporate algorithms and necessitate an assessment of their performance. Additionally, user behavior and qualitative task-based evaluation also exhibit a high correlation (0.33), elucidating that qualitative tasks are often evaluated through user behaviors such as mouse clicks and task completion times.

Based on our coding process, another prevalent evaluation presentation is the case study supplemented with expert feedback. In this format, authors typically assess the application's usefulness through case scenarios while simultaneously showcasing its usage. They might not explicitly mention the specific evaluation study protocol, but they can effectively demonstrate the functionality, elucidate how each feature supports the requirements, and depict how domain experts interact with the application in each case.

In terms of the associations between evaluation objectives and methods, we again employed association rules [34], using *lift* as the statistical metric. *Lift* is used to quantify the strength of association between two itemsets, indicating how much more often antecedents and consequents co-occur than if they were independent. This helps identify closely linked evaluation scenarios and methods, highlighting connections stronger than random chance.

Fig. 9 displays the frequently occurring item(sets) with high degrees of association. Consistent with the results mentioned earlier, UE, VDAR, and their combinations, due to their high occurrence, are also extracted in the frequently associated item(sets). The method item(sets) associated with them predominantly fall within the qualitative evaluation techniques. UE, emphasizing interaction with the application, is often assessed through methods like interviews and observations. On the other hand, VDAR, which underscores the usefulness of the application in problem-solving and aiding data analysis, is commonly evaluated using task-based evaluations, case studies, and interviews.

C. Domain Difference

Another focus of our review is exploring the influence of different application domains on evaluation practices. We categorized the domains and analyzed all variables related to the four indicators of evaluation, as well as the average occurrence frequency of the two soft filters in each domain. Fig. 10(left) shows the two soft filters that we set: 1) whether the article describes the evaluation study protocol, and 2) whether domain experts or users jointly participated in the requirements analysis phase. The former enhances the credibility and reproducibility of the evaluation, while the latter affirms the domain-specific utility of the VIS application. We observed that AI-related domains have a noticeably lower level of domain personnel involvement during the requirements analysis phase compared to other domain categories. This aligns with our impressions during the coding process. We found that VIS applications related to AI often involve more algorithmic content, and human involvement in evaluation is lower than in other domains. Moreover, as an emerging field, AI has a rapidly growing community and

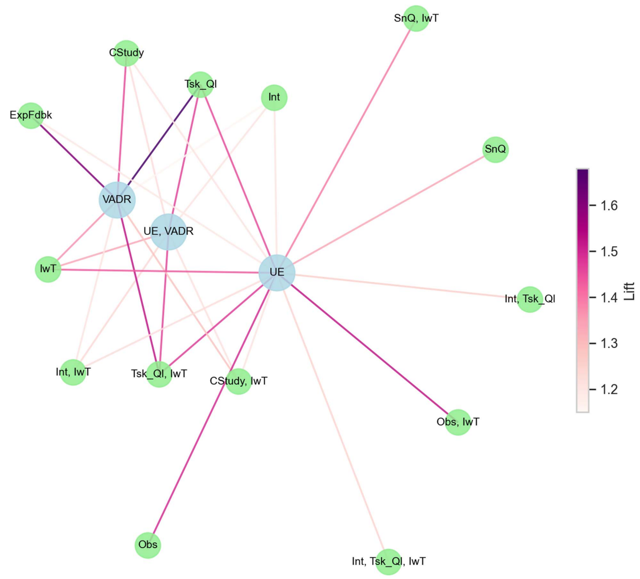

Fig. 9. Network graph illustrating the strength of connections between associated item(sets).

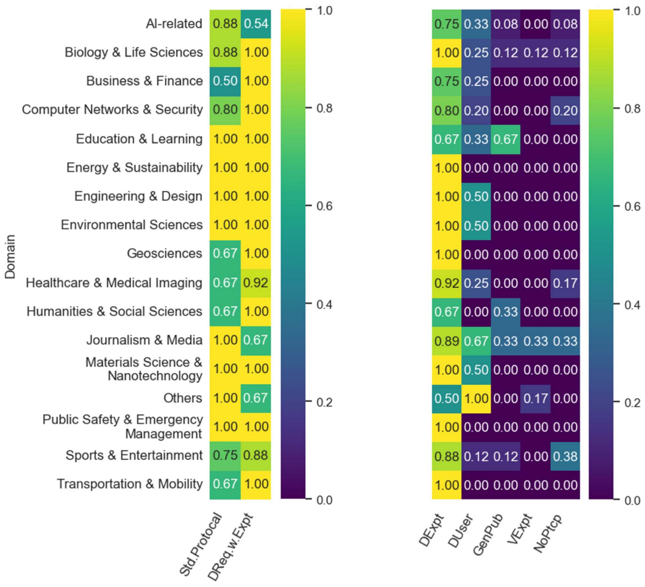

Fig. 10. Heatmap illustrating the avg. frequency of variables within the *Who* indicator (right) and two soft filters: 1) Study protocol provided and 2) Requirements analysis with domain experts/users' involvement (left) across domains.

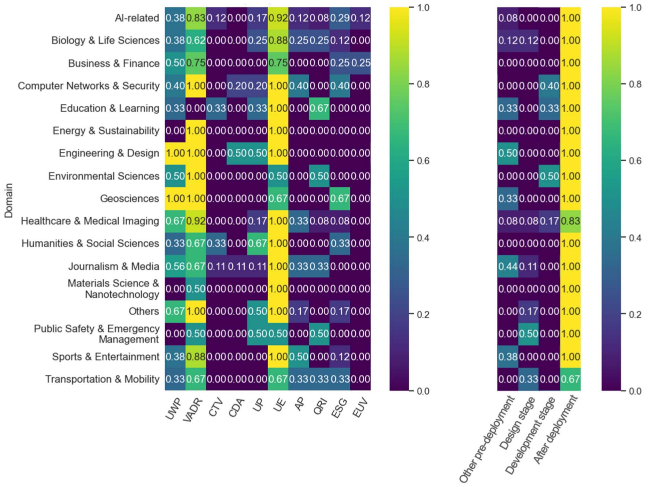

Fig. 11. Heatmap showing the avg. frequency of variables for the *What* (left) & *When* (right) indicators across domains.

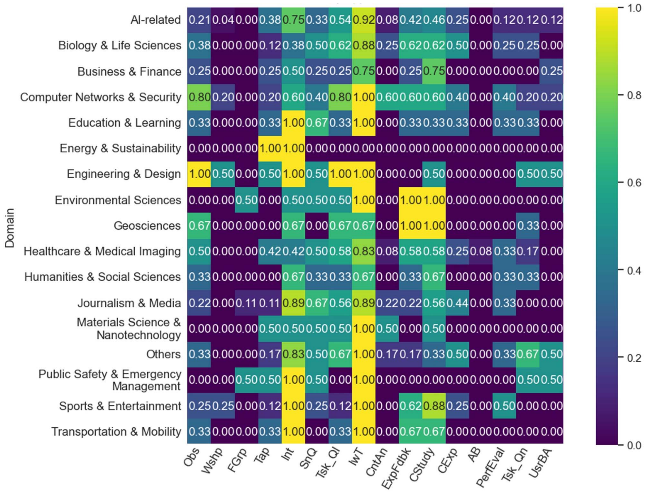

Fig. 12. Heatmap illustrating the avg. frequency of variables for the *How* indicator across domains.

many well-known challenges to overcome. These challenges and requirements can be directly obtained from current research and literature without the involvement of domain experts, ensuring credibility. The provision rate of evaluation protocols in Business & Finance-related VIS application research is unexpectedly low compared to other domains. This emphasizes the importance for researchers in this field to clearly express the rigor of study protocols in future research endeavors. In terms of the expertise of the domain experts involved in the evaluation, Fig. 10(right) shows that in the fields of Education & E-learning and Humanities & Social Science, the expertise is relatively low. This is consistent with the inherent nature of these domains, where their application content emphasizes communication and supports understanding, and the evaluation often involves participants without a strong professional background.

We analyzed the average distribution of evaluation objectives, methods, and execution stages across domains (Figs. 11 & 12), finding subtle but not significant differences.

TABLE II

Variables Significant At P = 0.1 for VIS4AI (n = 34) versus Non-VIS4AI (n = 106) Using Binary Logistic Regression With L1 Regularization

	coef	std err	z	$P > \|z\|$	[0.025	0.975]
CTV	1.8640	0.981	1.900	0.057	−0.059	3.787
UP	1.3955	0.787	1.772	0.076	−0.148	2.939
Obs	−1.5566	0.590	−2.638	0.008	−2.713	−0.400
PerfEval	−1.4987	0.835	−1.795	0.073	−3.135	0.137

A significance level of 0.1 is used for exploratory purposes. The table includes coefficient estimates (coef), standard error (std err), z-values (z), p-values, and 95% confidence intervals, which provide ranges for the coefficient estimates, indicating the likely extent and direction of each predictor variable's effect on the outcome.

D. New Trends in VIS Domain Application Research

Following our previous analysis, we observed differences in the evaluation practices of AI-related VIS applications (VIS4AI) compared to other domains. Additionally, we investigated whether VIS incorporating AI (AI4VIS) and using immersive approaches exhibits unique patterns in evaluation practices. To explore these differences, we used binary logistic regression with L1 (Lasso) regularization [35], [36]. This method allows for effective handling of multicollinearity, which we observed in earlier Section IV-B where some variables exhibited high correlations, and selects the most relevant predictors by shrinking the coefficients of less important variables to zero, thus providing a more interpretable model.

In this model, the response variable y is binary, where $y = 1$ indicates targeted papers (e.g., VIS4AI, AI4VIS, or immersive approaches) and $y = 0$ indicates non-targeted papers. The predictor variables X consist of the variables from the 3W1H indicators, which capture the evaluation practices used in the papers. By fitting the model $y \sim X$, we can determine which evaluation-related variables are significantly associated with being a targeted paper.

We primarily considered the differences between the targeted and non-targeted papers. For the VIS4AI category, the number of targeted papers is higher than those strictly within the AI-related domain because some papers categorized under Data-driven also relate to VIS4AI and were included in this count.

1) VIS4AI: The result shown in Table II reveals distinct evaluation practices between VIS4AI (n = 34) and non-VIS4AI (n = 106) papers. VIS4AI papers tend to emphasize evaluating how effectively visualization communicates (CTV), likely due to the intricate nature of AI models, which necessitates clear and effective visualization techniques. Furthermore, these papers prioritize user performance (UP), suggesting that visualization applications for AI are designed to enhance user comprehension and interaction. However, the VIS4AI papers seem to be less reliant on observational methods (Obs), possibly favoring quantitative metrics given the technical intricacies of AI. While performance evaluation (PerfEval) remains an essential aspect, it appears that the VIS4AI articles may be more focused on the clarity and interpretability of visualizations rather than just their speed and memory performance.

2) AI4VIS: Based on the result shown in Table III, we observe distinct patterns in AI4VIS (n=20) papers compared to

TABLE III
VARIABLES SIGNIFICANT AT P = 0.1 FOR AI4VIS (N=20) VERSUS NON-AI4VIS (N=120) USING BINARY LOGISTIC REGRESSION WITH L1 REGULARIZATION

	coef	std err	z	$P > \|z\|$	[0.025	0.975]
AP	−2.1561	1.021	−2.111	0.035	−4.158	−0.154
Tap	−1.1815	0.698	−1.692	0.091	−2.550	0.187
Tsk_Ql	1.0560	0.590	1.790	0.073	−0.100	2.212
CExp	1.2070	0.659	1.831	0.067	−0.085	2.499
Tsk_Qn	−2.8756	1.173	−2.451	0.014	−5.176	−0.576
UsrBA	2.4937	1.230	2.028	0.043	0.084	4.904
develop	2.2937	1.029	2.230	0.026	0.277	4.310

non-AI4VIS (n=120) papers. They often focus on understanding user interactions with AI-enhanced visualization applications, predominantly using controlled experiments (CExp) and emphasizing behavior analysis (UsrBA). They tend to lean toward qualitative task-based evaluations (Tsk_Ql) and proactive evaluations before full development (develop), likely due to using non-AI-enhanced human workflows as baselines for comparison with AI-enhanced behaviors. Additionally, the infrequent use of think-aloud protocols (Tap) suggests a preference for evaluation methods more suited to the specific characteristics of AI-enhanced visualizations.

3) Immersive VIS: Our analysis found no significant differences in evaluation practices between immersive (n=5) and non-immersive VIS applications (n=135). A crucial factor to consider is the relatively small number of papers on immersive VIS applications included in our review, which presents a promising avenue for future review-oriented studies.

V. CONSIDERATIONS FOR EVALUATION IN DOMAIN APPLICATION RESEARCH

Throughout this review, we have identified several considerations that we believe warrant further discussion and attention within the broader community.

A. Domain Integration

Domain integration was a pivotal filter in our article selection process. We posit that every VIS application research should be permeated with domain-specific knowledge to some extent, to truly elucidate the applicability and value of the research outcomes. For typical VIS applications tailored for a specific domain, this integration is essential. Whether it is the initial phase of domain requirements analysis or the subsequent stages focusing on application utility and usability testing, collaboration and input from domain experts are paramount.

However, there exists another category of VIS application research – those that are data-driven or problem-centric. Such research often proposes innovative visualizations or techniques to address overarching visual challenges or handle specific data types, emphasizing a domain-agnostic approach. During our selection, we excluded certain articles from this category, primarily because their evaluations lacked any form of domain integration, focusing solely on generic user experience.

We argue that even domain-agnostic applications should be validated through concrete application instances. Demonstrating their utility across one or more domains can provide a more comprehensive picture of their versatility. Nevertheless, our stance on the necessity of domain integration as a filter remains flexible. We acknowledge the potential value of VIS applications that operate without deep domain integration, especially when they cater to universal visualization challenges.

B. Significance of Requirements Analysis

Requirements and tasks analysis form an integral part of our filtering criteria. The rationale behind this is simple: applications arise from specific needs. Research on applications without a foundational requirements analysis is akin to constructing a building without a solid foundation. Regardless of the depth of implementation, its utility can easily come under scrutiny. During our coding process, our criteria for requirements analysis were adaptable, recognizing the diverse ways it can manifest in different papers. Descriptions such as design space, motivated usage scenario, and domain challenges were regarded by us as part of this analysis. All these facets contribute to a comprehensive understanding of the workflow and the challenges inherent to it.

While a majority of the papers we reviewed dedicated distinct sections to elucidate how requirements were established or distilled from discussions with domain experts, a notable fraction merely skimmed the surface. Some papers presented the requirements without introducing the process of their derivation or the intermediate validations they underwent. Although we recognize that each paper has its unique narrative style, we advocate a more thorough presentation of the requirements analysis. This not only bolsters the rigor of the research but also provides readers with a clearer understanding of the foundational needs driving the proposed solutions.

C. Evaluation Study versus Case Study

Evaluation studies employ a suite of crafted practices designed to assess specific objectives. These studies typically adhere to strict protocols, providing a formalized and systematic approach. Conversely, case studies and usage scenarios explore the efficacy of applications within the confines of a particular real-world context, focusing on how VIS applications integrate into domain environments and meet domain requirements [6]. We observed that both forms of evaluation are prevalent. Typically, they are presented under sections titled “Evaluation” and “Case Study.” In some cases, the authors integrate both in their research.

Case studies and usage scenarios are sometimes critiqued for lacking a clear protocol, which may affect their reproducibility. However, Meyer and Dykes [24] argue that reproducibility limitations do not necessarily diminish the value of a design study. They propose that thoroughness and transparency in presenting the study can compensate for these limitations. The unique strength of case studies lies in their ability to vividly showcase functionalities in action. For readers, this offers an intuitive correlation between use cases and the application’s features, facilitating a direct understanding of the application’s utility. This is an aspect in which traditional evaluation practices may fail.

We advocate for an inclusive approach to these evaluation presentations. Where space permits, we appreciate the inclusion of both protocols and use cases or including them as supplementary materials. We believe that intriguing observations from system evaluation studies, especially those concerning user interaction with the application, can be elaborated upon in an informal case study format. This approach not only enriches the content but also enhances the reader's comprehension and engagement.

D. Action versus Articulation

In the research process, it is imperative to delineate between what we need to evaluate, the actual evaluation practices undertaken, what we choose to report in the paper, and what readers anticipate from the evaluation section.

In practice, conducting a systematic evaluation study with a clear protocol is paramount, especially for VIS domain application research. However, the content in the paper often leans towards showcasing the developed application, using the evaluation as a supportive and validating element. This implies that the evaluation content presented in the paper represents only a fraction of the actual evaluation conducted. Consequently, the selection and articulation of this content requires careful consideration and discernment from the authors. The reviewers, on the other hand, should not take a negative attitude toward what fails to be presented in the article.

For instance, when presenting evaluator praise, superficial praise should be avoided in favor of emphasizing specific aspects of the appreciation received. At times, it is also beneficial to include negative feedback received during the evaluation. This not only showcases transparency, but also demonstrates how such feedback was addressed, leading to refinements that better align the final product with requirements.

Furthermore, the presentation format of the evaluation results should be tailored to the reader's perspective. For example, as mentioned previously, presenting findings through a case study can offer a more relatable and comprehensive view of the application's utility and effectiveness.

VI. Limitations and Future Works

This review sources papers accepted to the IEEE VIS conference which, while ensuring high-quality research coverage, introduces inherent limitations and potential biases by excluding works not accepted or submitted elsewhere. This may not fully represent the diverse evaluation practices across domain-focused VIS applications, particularly those published in domain-specific journals. A limitation in our methods and analysis is the small number of papers in certain subgroups. For example, the energy and sustainability domain has only one paper, and immersive approaches also have few papers. This leads to potentially biased qualitative and quantitative results. To address these limitations, future research should expand the scope of data collection to include various publication venues beyond IEEE VIS, such as domain-specific journals, workshop papers, and other conferences. This approach will help capture a more diverse and representative set of evaluation practices, enabling a more robust and comprehensive analysis.

Despite these limitations, our goal was to identify prevailing trends and potential features in VIS domain application evaluation practices, acknowledging potential biases while laying the groundwork for future detailed exploration. Researchers and practitioners are encouraged to build upon the results from this work by delving deeper into the diverse evaluation practices within the VIS community. Future studies are encouraged to expand the data collection scope, explore context-specific evaluation needs, and refine the classification method for capturing evaluation practices. We aspire for this work to serve as a cornerstone, steering and elevating the standards of evaluations within the VIS community.

References

[1] S. K. Boell and D. Cecez-Kecmanovic, "On being 'systematic' in literature reviews in is," *J. Inf. Technol.*, vol. 30, no. 2, pp. 161–173, 2015, doi: 10.1057/jit.2014.26.

[2] S. Carpendale, "Evaluating information visualizations," in *Information Visualization: Human-Centered Issues and Perspectives*, A. Kerren, J. T. Stasko, J.-D. Fekete, and C. North Eds., Berlin, Germany: Springer, 2008, pp. 19–45.

[3] P. Isenberg, T. Zuk, C. Collins, and S. Carpendale, "Grounded evaluation of information visualizations," in *Proc. Workshop BEyond Time Errors: Novel EvaLuation Methods Inf. Visual.*, 2008, Art. no. 6.

[4] C. Plaisant, "The challenge of information visualization evaluation," in *Proc. Work. Conf. Adv. Vis. Interfaces*, 2004, pp. 109–116.

[5] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale, "Empirical studies in information visualization: Seven scenarios," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 9, pp. 1520–1536, Sep. 2012.

[6] T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair, and T. Möller, "A systematic review on the practice of evaluating visualization," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 12, pp. 2818–2827, Dec. 2013.

[7] C. Dal et al., "Evaluating usability of information visualization techniques," in *Proc. Braz. Symp. Hum. Factors Comput. Syst.*, 2002.

[8] E. E. Banissi, *Information Visualisation: Techniques, Usability and Evaluation*, 1st ed., Newcastle upon Tyne, U.K.: Cambridge Scholars Publishing, 2014.

[9] B. Saket, A. Endert, and J. Stasko, "Beyond usability and performance: A review of user experience-focused evaluations in visualization," in *Proc. 6th Workshop Beyond Time Errors Novel Eval. Methods Visual.*, 2016, pp. 133–142.

[10] C. North, "Toward measuring visualization insight," *IEEE Comput. Graph. Appl.*, vol. 26, no. 3, pp. 6–9, May/Jun. 2006.

[11] P. Saraiya, C. North, V. Lam, and K. Duca, "An insight-based longitudinal study of visual analytics," *IEEE Trans. Vis. Comput. Graph.*, vol. 12, no. 6, pp. 1511–1522, Nov./Dec. 2006.

[12] M. Smuc, E. Mayr, T. Lammarsch, W. Aigner, S. Miksch, and J. Gärtner, "To score or not to score? Tripling insights for participatory design," *IEEE Comput. Graph. Appl.*, vol. 29, no. 3, pp. 29–38, May/Jun. 2009.

[13] J. Stasko, "Value-driven evaluation of visualizations," in *Proc. 5th Workshop Beyond Time Errors: Novel Eval. Methods Visual.*, New York, NY, USA, 2014, pp. 46–53.

[14] M. Tory and T. Moller, "Evaluating visualizations: Do expert reviews work?," *IEEE Comput. Graph. Appl.*, vol. 25, no. 5, pp. 8–11, Sep./Oct. 2005.

[15] A. Crisan and M. Elliott, "How to evaluate an evaluation study? Comparing and contrasting practices in VIS with those of other disciplines: Position paper," in *Proc. IEEE Eval. Beyond - Methodol. Approaches Visual.*, 2018, pp. 28–36.

[16] B. Preim, T. Ropinski, and P. Isenberg, "A critical analysis of the evaluation practice in medical visualization," in *Proc. Eurographics Workshop Vis. Comput. Biol. Med.*, 2018, pp. 45–56.

[17] O. Espinosa, C. Hendrickson, and J. Garrett, "Domain analysis: A technique to design a user-centered visualization framework," in *Proc. IEEE Symp. Inf. Visual.*, 1999, pp. 44–52.

[18] G. H. Weber et al., "Apply or die: On the role and assessment of application papers in visualization," *IEEE Comput. Graph. Appl.*, vol. 37, no. 3, pp. 96–104, May/Jun. 2017.

[19] IEEE VIS 2023. (2023) IEEE VIS 2023 - Call for participation - Area model, IEEE VIS. Accessed: Sep. 9, 2023. [Online]. Available: https://ieeevis.org/year/2023/info/call-participation/area-model

[20] M. Sedlmair, M. Meyer, and T. Munzner, "Design study methodology: Reflections from the trenches and the stacks," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 12, pp. 2431–2440, Dec. 2012.

[21] T. Munzner, "A nested model for visualization design and validation," *IEEE Trans. Vis. Comput. Graph.*, vol. 15, no. 6, pp. 921–928, Nov./Dec. 2009.

[22] M. Meyer, M. Sedlmair, P. S. Quinan, and T. Munzner, "The nested blocks and guidelines model," *Inf. Visual.*, vol. 14, no. 3, pp. 234–249, 2015.

[23] S. McKenna, D. Mazur, J. Agutter, and M. Meyer, "Design activity framework for visualization design," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 2191–2200, Dec. 2014.

[24] M. Meyer and J. Dykes, "Criteria for rigor in visualization design study," *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 1, pp. 87–97, Jan. 2020.

[25] M. Sedlmair, "Design study contributions come in different guises: Seven guiding scenarios," in *Proc. 6th Workshop Beyond Time Errors Novel Eval. Methods Visual.*, 2016, pp. 152–161.

[26] X. Wang et al., "VIS AI: Integrating visualization with artificial intelligence for efficient data analysis," *Front. Comput. Sci.*, vol. 17, 2023, Art. no. 176709.

[27] A. Wu et al., "AI4VIS: Survey on artificial intelligence approaches for data visualization," *IEEE Trans. Vis. Comput. Graph.*, vol. 28, no. 12, pp. 5049–5070, Dec. 2022.

[28] L. Encarnacao, J. Kohlhammer, and C. A. Steed, "Visualization for AI explainability," *IEEE Comput. Graph. Appl.*, vol. 42, no. 06, pp. 9–10, Nov./Dec. 2022.

[29] A. J. Onwuegbuzie, N. L. Leech, and K. M. T. Collins, "Qualitative analysis techniques for the review of the literature," *Qualitative Rep.*, vol. 17, no. 28, pp. 1–28, 2012.

[30] Ç. Ç. Karaman, S. Yaliman, and S. A. Oto, "Event detection from social media: 5W1H analysis on Big Data," in *Proc. 25th Signal Process. Commun. Appl. Conf.*, 2017, pp. 1–4.

[31] Y. Yu and Y. Bi, "A study on "5W1H," user analysis on interaction design of interface," in *Proc. IEEE 11th Int. Conf. Comput.-Aided Ind. Des. Conceptual Des.*, 2010, pp. 329–332.

[32] A. Burns, C. Lee, R. Chawla, E. Peck, and N. Mahyar, "Who do we mean when we talk about visualization novices?," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2023, Art. no. 819.

[33] E. Duneja and A. Sachan, "A survey on frequent itemset mining with association rules," *Int. J. Comput. Appl.*, vol. 46, pp. 18–24, 2012.

[34] A. Gosain and M. Bhugra, "A comprehensive survey of association rules on quantitative data in data mining," in *Proc. IEEE Conf. Inf. Commun. Technol.*, 2013, pp. 1003–1008.

[35] J. E. King, "Binary logistic regression," in *Best Practices in Quantitative Methods*, Newcastle upon Tyne, U.K.: SAGE, 2008, pp. 358–384.

[36] L. Meier, S. Van De Geer, and P. Bühlmann, "The group lasso for logistic regression," *J. Roy. Stat. Soc. Ser. B: Stat. Methodol.*, vol. 70, no. 1, pp. 53–71, Jan. 2008.

Yiwen Xing is currently working toward the PhD degree in computer science with the Informatics Department, King's College London (KCL). Her research interests encompass data visualization, data science, and design studies.

Gabriel D. Cantareira received the PhD degree from the University of São Paulo. He is a postdoctoral research associate with the Informatics Department, King's College London (KCL). His research interests are focused on data science, visual analytics, and explainable AI.

Rita Borgo is a reader in data visualization with the Informatics Department, King's College (KCL), and head of the Human Centred Computing Research Group. Her main research interests lie in the areas of information visualization, visual analytics, human-AI interaction, human factors in visualization, and urban science. Her research has been awarded support from the Royal Society, EPSRC, and EU. She is a partnership lead for the Centre for Urban Science and Progress (CUSP) – London.

Alfie Abdul-Rahman received the PhD degree from Swansea University. She is a senior lecturer in computer science with the Informatics Department, King's College London (KCL). Before joining KCL, she was a research associate with the University of Oxford e-Research Centre. She worked as a research engineer with HP Labs Bristol on document engineering, and then as a software developer in London, working on multi-format publishing. Her research interests include visualization, computer graphics, and human-computer interaction.