

City Research Online

City St George's, University of London

Citation: Madhyastha, P. & Hutchinson, M. (2026). On Interpretability Methods for Vision Language Models in Visualization. Paper presented at the IEEE VIS VISxGenAI Workshop, 9-13 Nov 2026, Boston, USA.

This is the accepted version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/38604/

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to City Research Online policy.

City Research Online: http://openaccess.city.ac.uk/ publications@citystgeorges.ac.uk

On Interpretability Methods for Vision Language Models in Visualization

Pranava Madhyastha*
City, University of London

Maeve Hutchinson
City, University of London

Abstract

The visualization community has invested substantially in deploying Vision-Language Models (VLMs) for visual analytics. However, the complementary question of *how* exactly do these models reason about visualizations, and how do we know when they are doing so correctly has received comparatively little rigorous attention. In this position paper, we argue that the interpretability methods currently applied to VLMs in the visualization domain inherit a cascade of structural mismatches. Methods ported from computer vision assume a crisp notion of objectness, spatial locality, and to a large extent, context-independence that natural images possess but data visualizations do not. Human gaze studies, often proposed as ground truth to compare and contrast interpretability methods, carry their own fragilities rooted in the reproducibility crisis in perception research. Finally, and perhaps more importantly, VLMs themselves are not vision models as their visual inputs merely serve as adaptors into a language model's (linguistic) representational space, fundamentally altering the notion of what "visual attention" means. To probe these questions, we introduce a dual causal-chain account of visualization comprehension, distinguishing the designer's intended reading path from the reader's actual comprehension trace, we propose a set of necessary conditions for valid interpretability evaluation in this domain. We posit that the visualization community is uniquely positioned to develop novel interpretability methods which can significantly impact both theoretical and empirical advances in the broad field of visual analytics.

Index Terms: Vision-language models, interpretability, saliency, referential chain, shortcut learning.

1 Introduction

There has been a recent surge in an enthusiastic integration of large language and vision-language models into visualization research and practice. Models are now applied to chart question answering [28, 30, 19], automated chart captioning [39], visualization design assistance [13, 18], and the broader visual analytics pipeline [23, 21]. Specialized architectures such as ChartGemma [29] and ChartLlama [17] have been fine-tuned specifically for visualization tasks, and many contributions leverage such models to produce new visualization tools and systems [45].

This momentum is indeed necessary as VLMs demonstrate impressive capabilities and their integration into analytical workflows holds genuine promise. However, given that deploying these models at scale in high-stakes analytical contexts, where a misread trend or a confabulated value may have consequential downstream effects, we must perhaps also ask a harder question: *do we understand what these models are actually doing when they interpret a visualization?* We argue that the answer is, at present, *no*, and that we currently lack the tools to rigorously answer this question. The interpretability methods currently applied to VLMs in the visualization domain are borrowed, from computer vision, from natural language processing, and from human perception research. Each borrowing brings productive insights but also structural mismatches that compound one another. The result is an interpretability apparatus that can produce plausible-looking outputs while measuring something importantly different from what we need to know.

This paper makes four contributions. First, we characterise the specific ways in which data visualizations are *semiotically distinct* from natural images, and why this distinction undermines the assumptions of standard interpretability methods. Second, we introduce a dual causal-chain account of visualization comprehension, specifically distinguishing the *designer's intended reading path* from the *reader's actual comprehension trace*, and argue that valid interpretability for VLMs must be able to diagnose the relationship between these two chains and the model's actual behaviour. Third, we discuss the broader landscape of interpretability critiques across computer vision and NLP, situating our argument within this wider discourse and identifying what is specifically novel about visualization in these contexts. Fourth, we derive a set of necessary conditions for valid interpretability evaluation in this domain and specify what a next-generation framework satisfying these conditions would require.

Our aim in this paper is to emphasise that the visualization community possesses exactly the theoretical vocabulary, be it theoretical foundations of visual encodings [44, 31], process models [7, 34], task taxonomies and typologies [1, 6], and empirical study of human visual cognition for analytical tasks [9], that borrowed interpretability methods are currently missing. Building these interpretability methods can perhaps help us reflect the fundamental explanatory causal pathways of both visualization design and comprehension, and help in significantly pushing both theoretical and empirical advances in the field.

2 A Cascade of Structural Mismatches

2.1 A Brief Background

The most widely adopted interpretability approach for vision models is saliency mapping, which attributes a model's output to spatial regions of the input image. Methods range from perturbation-based approaches such as LIME [32] to to gradient-based techniques including Grad-CAM [36] and its broad derivatives. Boggust et al. [4] provide a systematic characterization of saliency methods across a range of properties, that includes output type, model access requirements, and human alignment, offering a framework for principled method selection. Within the visualization domain, saliency maps have a parallel tradition grounded in human gaze behaviour. Eye-tracking studies have produced gaze-based saliency maps for tasks including chart memorability [5] and question answering [41], which have been used both to study human cognition and as a reference for evaluating model attention. Recent work has applied gradient-based saliency methods directly to VLMs on visualization tasks [14], and explored alignment between model saliency and human gaze [35].

These contributions represent genuine progress. But we argue they inherit, and in some cases compound, a set of structural assumptions that make them inadequate for the specific demands of visualization interpretability. We develop this argument in the following section. A related but distinct line of work has pursued concept-based explanation methods, including TCAV [22], network dissection, and CLIP-based probing approaches, which attribute

*correspondences for both: {firstname.lastname}@city.ac.uk

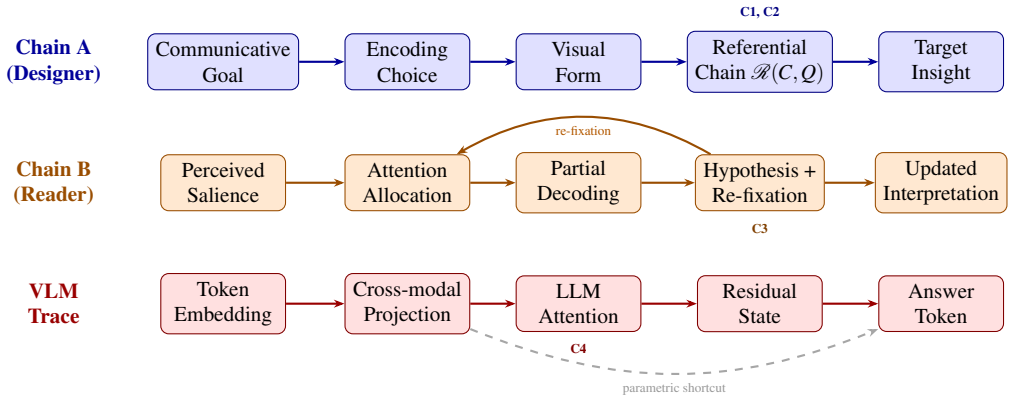


Figure 1: The dual causal chain account. **Chain A** (blue) represents the designer's intended reading path, formalized as the referential chain $\mathscr{R}(C,Q)$. **Chain B** (orange) represents the reader's actual comprehension trace, which is iterative and involves re-fixation. The **VLM trace** (red) follows neither chain faithfully: visual information is projected into a joint embedding space before reasoning occurs in the language backbone, and parametric shortcuts (dashed arc) allow correct answers to be generated without traversing the referential chain at all. The four necessary conditions (C1–C4) are annotated at the chain elements they govern.

model behaviour to human-interpretable semantic concepts rather than pixel regions. These methods represent genuine progress over pixel-saliency by operating in embedding rather than image space, and some recent mechanistic interpretability work has begun probing the internal circuits of transformer models at a fine-grained level [2]. While we acknowledge important work in this trajectory, these methods, however, still lack the domain-specific vocabulary needed to define what concepts are relevant for visualization reading. Without a formal ontology, of which the visualization field has many, there is no principled basis for specifying which concepts should be probed , what their relationships are, or what joint activation patterns constitute correct visual grounding for a given visualisation task. In the following sections, we develop and expand on this specific characterisation.

2.2 Visualizations are Not Natural Images

The most fundamental mismatch is between the semiotic structure of data visualizations and the assumptions built into computer vision interpretability methods. Standard saliency approaches, including the more popular Grad-CAM [37] and its variants, were developed for natural images, where meaning resides in coherent objects with spatial continuity, texture, and scale invariance. The implicit notion of *objectness*, that important image regions correspond to semantically coherent, spatially compact entities, underlies how these methods define and evaluate salience.

Data visualizations violate these assumptions systematically. To see why, it helps to make the formal structure of visualization explicit. Wilkinson's grammar of graphics [43], for instance, describes a visualization as a compositional pipeline:

$$\textsc{Data} \xrightarrow{\textsc{Trans}} \textsc{Stat} \xrightarrow{\textsc{Scale}} \textsc{Coord} \xrightarrow{\textsc{Element}} \textsc{Mark} + \textsc{Guide}$$

where TRANS applies statistical transformations (aggregation, binning), SCALE maps data values to visual ranges, COORD establishes the coordinate system, ELEMENT specifies the mark geometry, and GUIDE provides axes, legends, and labels, the *inverse-mapping infrastructure* that makes values recoverable from marks. Other foundational work into visualisation specifies that each encoding channel (position, size, colour hue, shape, orientation) has distinct perceptual properties such as position is quantitative and selective, hue is categorical and associative [3, 26, 33]. These properties determine what data relationships each channel can faithfully represent.

On the other hand, reading a visualization means traversing this pipeline *in reverse*, that is, from perceived mark, through its encoding channel and scale, via the guide infrastructure, back to a data value. We define the *referential chain* $\mathscr{R}(C,Q)$ for chart type C and question type Q as the minimal ordered set of $\langle$mark, channel, scale, guide$\rangle$ tuples that must be jointly processed to correctly answer Q from C. To make this concrete: for a bar chart value retrieval question, the grammar yields:

$$\mathscr{R}(\text{bar},\text{value}) = \langle \text{target bar} \xrightarrow{\text{height}\in\text{position}_y} \text{y-scale} \xrightarrow{\text{inverse}} \text{y-axis tick} \rangle$$

The target bar encodes a value via the *position* retinal variable, the y-scale maps pixel height to data range, the y-axis tick labels the inverse. All three elements must be jointly attended to, that is, (and we emphasise here that attending to the bar alone recovers nothing), again attending to the axis alone is uninformative without the bar. This joint necessity is what *referential completeness* means, and it is a property that pixel-overlap metrics such as mIoU are structurally unable to measure.

Drawing on this formalism, and prior research in visualization, we can characterize visualizations as *compositional, relational, and context-dependent* in ways that natural images are usually not (at least in the framing of usual computer vision tasks):

Meaning is relational. A bar in a bar chart carries no numerical information in isolation. Its meaning is constituted by its relationship to the y-axis scale, the relevant gridline or reference, and the legend if colour encodes a categorical variable [38]. Saliency over a single mark, evaluated independently, is mostly uninterpretable. What matters importantly is whether the full referential chain $\mathscr{R}(C,Q)$ was traversed coherently.

Salient regions are task-relative. What constitutes the important region of a visualization is a joint function of the image and the analytical question being posed. A trend question and a value retrieval question over the same chart require attention to entirely different regions and referential chains [1]. Any interpretability method that evaluates saliency without conditioning on task type is measuring something undefined.

Visualizations are sparse and symbolic. Unlike natural images with continuous textures, visualizations frequently encode critical information in high-frequency, minimal-area elements, such as, a single data point, the gap between two bars, the crossing of two lines, or the endpoint of a range mark. Methods that assume spatial smoothness or that smooth importance scores across regions will systematically miss these elements [8].

Reading a chart is iterative and recursive not a unique single recognition event. Natural image recognition approximates a single forward inference, say for example the scene contains a cat. Visualization understanding is on the other hand a process where the reader scans for structure, identifies a legend, forms a hypothesis, re-examines an axis, revises the interpretation [34]. This iterative, multi-fixation process cannot be captured by a single saliency map over a single forward pass.

2.3 On Objectivity in Human Gaze studies

A natural response to the limitations of model-derived saliency is to turn to human gaze data as a reference. Eye-tracking studies of chart reading provide empirical maps of where human readers attend, and several works have used these as targets against which to evaluate VLM attention [41, 35].

This approach is indeed valuable but, we argue that this carries its own structural fragilities. The reproducibility crisis in psychology and cognitive science [10] has had significant impact on small-N perceptual studies of exactly the kind that gaze research on visualizations typically involves. Studies are highly sensitive to task framing, display conditions, participant expertise, and individual differences in reading strategy. Expert readers compress the referential chain, that is, they skip steps that novices must traverse explicitly, meaning that a single gaze map cannot characterize "correct" chart reading across populations. Furthermore, recent empirical investigations into perceptual alignment [24] confirm that VLMs often exhibit a severe over-reliance on explicit textual cues, structurally bypassing the visual encodings entirely, struggling most precisely with the visual encoding types that sit at the heart of the referential chains we describe. Attempting to benchmark these models against human perceptual traces when the underlying mechanisms of extraction are categorically distinct is therefore a fragile ground truth.

More fundamentally, using human gaze as a ground truth conflates two distinct questions of whether a VLM reads a chart *as humans do*, and whether it reads it *correctly*. These indeed are not the same question. A model might traverse the referential chain correctly via a path that differs from any human's gaze trajectory, or might produce human-like gaze patterns while deriving its answer from parametric memory rather than visual extraction. Human gaze alignment is at best a weak proxy for genuine visual grounding.

2.4 VLMs are Not Vision Models

The third and deepest mismatch concerns the architecture of VLMs themselves. A persistent error in applying vision interpretability methods to VLMs is treating them as vision models with language output. They are not, and this distinction and appreciation of modelling choice matters precisely for interpretability.

Consider a representative architecture in the LLaVA family [15], which underlies many chart-focused models including those evaluated on ChartQA. A Vision Transformer encodes the chart image as a sequence of N patch embeddings $\{v_1, \ldots, v_N\}$. A lightweight projection layer (typically a two-layer MLP) maps these into the language model's token embedding space, producing visual tokens $\{\hat{v}_1, \ldots, \hat{v}_N\}$ that are prepended to the text token sequence. From this point forward, the language model, which is a decoder-only transformer, processes visual and textual tokens identically. There is no architectural distinction between attending to a patch embedding and attending to a word embedding as both are vectors in the same high-dimensional space, routed by the same attention mechanism. The model does not "look at" the chart during answer generation, instead it routes information already encoded in its residual stream, in a representational space where the geometric relationship between visual tokens and their original spatial positions in the image has been substantially distorted by projection and multiple layers of self-attention.

This has several important consequences for interpretability. First, attention in the language layers is not looking at image regions in any meaningful sense as it is routing information through a joint embedding space in which visual and textual tokens have been fused. Standard gradient-based saliency methods applied to the visual encoder's outputs are measuring something that precedes the model's actual reasoning process, that is they probe the encoder, not the reasoner.

Second, the debate initiated by Jain and Wallace [20] and contested by Wiegreffe and Pinter [42] on whether attention weights constitute faithful explanations of model decisions in language models, this becomes substantially more complex in the cross-modal setting. In a unimodal language model, attention over tokens at least operates in a space where tokens have stable semantic identities. In a VLM, the visual tokens $\hat{v}_i$ are projections of patch embeddings whose semantic content is not directly interpretable, so a high attention weight on $\hat{v}_i$ does not straightforwardly indicate that the model is attending to the spatial region of the image that patch i covered, because the projection layer has mixed spatial and semantic information across the patch neighbourhood. The resolution of the attention related debates within NLP does not easily transfer to this setting, even if attention *is* a faithful explanation in a language model, it is not a *spatially grounded* one when the attended tokens are projected patch embeddings.

Third, there is a practical infrastructure barrier that deserves more than passing mention. Modern large-scale VLMs are compiled and deployed using fused attention kernels, implementations such as FlashAttention [12] that compute the attention output in a single hardware-efficient pass without ever materialising the $N \times N$ attention weight matrix. Extracting attention weights for interpretability therefore requires disabling these optimisations and reverting to eager execution mode. The fused and eager implementations can differ in numerical precision, in which operations are computed in what order, and for models that use techniques such as speculative decoding or KV-cache manipulation in the effective computation graph. The model whose attention weights are inspected, most commonly, is not identical to the model that is deployed. Interpretability, under current infrastructure norms, requires breaking the system being interpreted. So almost all interpretation techniques are giving us a very crude notion, sometimes unhelpful, of the internals of the model behaviour.

3 A Dual Causal Account of Chart Comprehension

We propose that adequate interpretability for VLMs on visualizations requires distinguishing two causal chains that govern chart reading, and diagnosing where a model's behaviour stands in relation to each. Figure 1 illustrates the three trajectories and the diagnostic space they define.

Chain A: The designer's intended reading path. A visualization encodes a communicative intent. The designer has selected data, applied transformations, chosen encoding channels, and arranged guides (axes, legends, labels) to make a specific analytical insight perceptually accessible [31]. This intent can be partially formalized using the vocabulary of visualization algebra, that is, for a given chart type and question type, there is a *referential chain* which is the minimal ordered set of mark elements, guide elements, and scale mappings that must be jointly processed to recover the correct answer.

Chain B: The reader's actual comprehension trace. Human readers do not follow the designer's intended path mechanically. They are driven by perceptual salience, prior knowledge, task framing, and cognitive load. Their comprehension is iterative where fixations cluster on visually prominent elements, hypotheses are formed and tested, and attention returns to guides when a value must be resolved. This trace is specialised and individual, variable, and incompletely captured by gaze studies. But it is systematically

structured, and its systematic deviations from Chain A are the subject of a substantial visualization perception literature [5].

Readers over-attend to visually salient but informationally redundant elements (prominent marks, large text) and under-attend to informationally critical but perceptually modest elements (axis scales, legends). This mismatch between salience and informational importance is a known failure mode of human chart reading, and it is precisely the failure mode that VLMs may inherit, amplify, or circumvent via parametric shortcuts.

This dual-chain account enables three diagnostic questions: a) **Chain A alignment**: Does the model attend to the elements of the referential chain specified by the chart's encoding grammar?; b) **Chain B alignment**: Does the model's attention trace resemble a human's iterative comprehension process? c) **Shortcut detection**: Does the model produce correct answers while aligning with the chain?

We believe the third question to be the most important and having a substantial consequence. A model that consistently answers correctly while attending to neither the referential chain nor any plausible human reading trajectory is, almost certainly, retrieving answers via parametric memory, text-based heuristics, or training data contamination [16, 11]. This vulnerability is specifically demonstrated by recent hallucination-specific benchmarks such as ChartHal [40], which reveal a catastrophic collapse in accuracy when models are confronted with queries that contradict their parametric priors or lack explicit text cues, with state-of-the-art proprietary (popularly referred as frontier) models achieving a substantially lower accuracy on such cases. We emphasise that this is a diagnosis that standard accuracy metrics cannot naturally make. More importantly, Chain A can be derived from the visualization's encoding grammar without any human data, it is a theory-derived baseline, not an empirical one, and thus sidesteps the reproducibility problems of gaze research entirely.

4 Necessary Conditions for Valid Interpretability

We draw from our analyses in the preceding sections and specify four necessary conditions that any valid interpretability framework for VLMs on data visualizations must satisfy. Each condition is motivated by a specific property of the dual causal chain account or the architectural analysis above. These conditions are intended as a target for future framework development, and as a diagnostic test against which existing methods can actually be carefully and rigorously evaluated.

C1: Context-Relativity. An interpretability measure must be conditioned on the analytical task and we must not evaluate independently of the task itself. A saliency map over a chart is undefined without a question. This condition follows directly from the task-relativity of Chain A where the referential chain $\mathscr{R}(C,Q)$ is a function of both chart type *and* question type, so any evaluation that does not condition on Q is evaluating against an undefined target. This rules out any image-level saliency metric computed independently of task type, including current applications of mIoU and the pointing game metric to chart interpretability.

C2: Referential Completeness. An interpretability method must be able to assess whether the full referential chain $\mathscr{R}(C,Q)$ was traversed (or what the information flow was). This condition is the operationalisation of Chain A on attending to a bar in isolation is not equivalent to attending to a bar in relation to its axis, because the chain $\langle \text{bar} \xrightarrow{\text{position}_y} \text{y-scale} \xrightarrow{\text{inverse}} \text{axis} \rangle$ requires joint co-attention across all three elements. Region-level metrics usually only measure elements where as C2 requires chain-level evaluation.

C3: Process Sensitivity. In an autoregressive model, correct visual grounding may occur at different points in the generation process, during prompt processing, during intermediate token generation, or at the moment of answer token production. This condition is motivated by the architectural analysis of Section 2 where the decoupling of visual encoding from answer generation in LLM-backbone VLMs means that a model may "solve" a chart query in early layers during prompt processing, caching the result in its residual stream, and generate the correct answer token without any active attention to the visualization at generation time. An interpretability method must therefore be sensitive to *when* and in what *sequence* attention was allocated

C4: Architectural Relevance. Any framework must specify what it is measuring given the model's actual architecture. For VLMs, this means stating explicitly which layer's attention is being inspected or what the relationship is between that attention and the joint visual-textual embedding space described in Section 2 or whether the inspection mode (eager vs. optimised execution) alters model behaviour and whether attention weights over projected patch tokens constitute spatial grounding in the original image. Frameworks that apply saliency vocabulary which were perhaps well suited in the context of convolutional neural networks to transformer architectures [8] or that inspect the visual encoder while the reasoning occurs in the language backbone perhaps violate this condition by failing to declare what they are actually measuring.

No current method satisfies all four conditions. Gradient-based saliency methods fail C1, C2, C3, and C4. Human gaze alignment fails C3 and provides only partial satisfaction of C2. Attention-guided methods of the kind proposed for ViTs [8] make progress on C4 but do not address C1 or C2. There seems to be a structural gap in the field's interpretability vocabulary.

5 Discussion and Call to the Community

The necessary conditions we have articulated that include context-relativity, referential completeness, process sensitivity, and architectural relevance require two important ingredients: a formal theory of how visualizations encode meaning, and a rigorous empirical tradition of studying how humans read them. This very community is perhaps the best suited as the community has both, in the form of visualization algebra and the perception literature. The AI interpretability community has made significant strides toward more structured accounts, such as mechanistic interpretability and concept-based methods [22]. These represent genuine efforts to move beyond pixel attribution, but indeed lack the domain-specific vocabulary to define what semantic concepts are relevant for chart reading, what their causal relationships are, or what joint activation patterns constitute correct visual grounding. We believe that the vocabulary is explicitly grounded in visualization theory and is situated precisely in the visualization community. As VLMs are increasingly deployed in analytical contexts where the stakes of misinterpretation are high, the absence of valid interpretability methods to probe, understand and fix is of utmost urgency. The community that studies how meaning is encoded in visualizations is perhaps the community best positioned to specify what it means for a model to read one correctly. In this paper, we have made an effort to present three components that a potential framework (theoretical or empirical) would require where we present a chart-type and task-type taxonomy of referential chains derived from previous prominent work by the community. VLMs also provide something that prior visualization research has rarely had access to, in the form a large-scale, systematically probed computational system whose successes and failures in chart reading and interpreting can be interrogated across tens of thousands of chart-question pairs and diverse encoding conditions. Cognitive science has used computational models as theoretical probes [27] where a model that fails in the same ways as humans under similar conditions provides evidence for shared computational structure, while a model that fails differently sharpens the account of what makes human cognition distinctive [25].

REFERENCES

[1] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *INFOVIS*, pp. 111–117, Oct. 2005. doi: 10.1109/INFVIS.2005.1532136 1, 2

[2] E. Ameisen et al. Circuit Tracing: Revealing Computational Graphs in Language Models, 2025. 2

[3] J. Bertin. *Semiology of Graphics: Diagrams, Networks, Maps*. ESRI Press, 2011. 2

[4] A. Boggust, H. Suresh, H. Strobelt, J. Guttag, and A. Satyanarayan. Saliency Cards: A Framework to Characterize and Compare Saliency Methods. In *FAaCT 2023*, FAccT '23, pp. 285–296. ACM, New York, NY, USA, June 2023. doi: 10.1145/3593013.3593997 1

[5] M. A. Borkin, Z. Bylinskii, N. W. Kim, C. M. Bainbridge, C. S. Yeh, D. Borkin, H. Pfister, and A. Oliva. Beyond memorability: Visualization recognition and recall. *TVCG*, 22(1):519–528, 2015. 1, 4

[6] M. Brehmer and T. Munzner. A Multi-Level Typology of Abstract Visualization Tasks. *TVCG*, 19(12):2376–2385, Dec. 2013. doi: 10.1109/TVCG.2013.124 1

[7] S. K. Card, J. D. Mackinlay, and B. Shneiderman. *Readings in Information Visualization: Using Vision to Think*. The Morgan Kaufmann series in interactive technologies. Morgan Kaufmann Publishers, San Francisco, Calif, 1999. 1

[8] H. Chefer, S. Gur, and L. Wolf. Transformer interpretability beyond attention visualization. In *CVPR*, pp. 782–791, 2021. doi: 10.1109/CVPR46437.2021.00084 2, 4

[9] W. S. Cleveland, , and R. McGill. Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association*, 79(387):531–554, Sept. 1984. doi: 10.1080/01621459.1984.10478080 1

[10] O. S. Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015. 3

[11] C. Dancette, R. Cadene, D. Teney, and M. Cord. Beyond Question-Based Biases: Assessing Multimodal Shortcut Learning in Visual Question Answering, Sept. 2021. doi: 10.48550/arXiv.2104.03149 4

[12] T. Dao, D. Fu, S. Ermon, A. Rudra, and C. Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359, 2022. 3

[13] V. Dibia. LIDA: A Tool for Automatic Generation of Grammar-Agnostic Visualizations and Infographics using Large Language Models. In D. Bollegala, R. Huang, and A. Ritter, eds., *ACL*, pp. 113–126. ACL, Toronto, Canada, July 2023. doi: 10.18653/v1/2023.acl-demo.11 1

[14] L. Dong and A. Crisan. Probing the visualization literacy of vision language models: the good, the bad, and the ugly. *arXiv preprint arXiv:2504.05445*, 2025. 1

[15] A. Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[16] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, Nov. 2020. doi: 10.1038/s42256-020-00257-z 4

[17] Y. Han et al. Chartllama: A multimodal llm for chart understanding and generation. *arXiv preprint arXiv:2311.16483*, 2023. 1

[18] M. Hutchinson, R. Jianu, A. Slingsby, J. Wood, and P. Madhyastha. Capturing Visualization Design Rationale, July 2025. doi: 10.48550/arXiv.2506.16571 1

[19] M. Hutchinson, R. Jianu, A. Slingsby, J. Wood, and P. Madhyastha. Chart Question Answering from Real-World Analytical Narratives. In J. Zhao, M. Wang, and Z. Liu, eds., *ACL SRW*, pp. 760–773. ACL, Vienna, Austria, July 2025. doi: 10.18653/v1/2025.acl-srw.50 1

[20] S. Jain and B. C. Wallace. Attention is not Explanation. In J. Burstein, C. Doran, and T. Solorio, eds., *NAACL*, pp. 3543–3556. ACL, Minneapolis, Minnesota, June 2019. doi: 10.18653/v1/N19-1357 3

[21] R. Jianu et al. Mind-Mapping Data Analysis with LLMs: From Vision to First Steps. In *CGVC*. The Eurographics Association, Sept. 2025. doi: 10.2312/cgvc.20251221 1

[22] B. Kim et al. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *ICML*, pp. 2668–2677. PMLR, July 2018. 1, 4

[23] X. Lei et al. Scaffolding coordinates to promote vision-language coordination in large multi-modal models. In *COLING*, pp. 2886–2903, 2025. 1

[24] F. Lin et al. Do Vision-Language Models Measure Up? Benchmarking Visual Measurement Reading with MeasureBench, Mar. 2026. doi: 10.48550/arXiv.2510.26865 3

[25] Z. C. Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018. doi: 10.1145/3236386.3241340 4

[26] A. M. MacEachren et al. Visual Semiotics & Uncertainty Visualization: An Empirical Study. *TVCG*, 18(12):2496–2505, Dec. 2012. doi: 10.1109/TVCG.2012.279 2

[27] D. Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010. 4

[28] A. Masry et al. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In S. Muresan, P. Nakov, and A. Villavicencio, eds., *ACL*, pp. 2263–2279. ACL, Dublin, Ireland, May 2022. doi: 10.18653/v1/2022.findings-acl.177 1

[29] A. Masry et al. ChartGemma: Visual Instruction-tuning for Chart Reasoning in the Wild. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, S. Schockaert, K. Darwish, and A. Agarwal, eds., *ACL*, pp. 625–643. ACL, Abu Dhabi, UAE, Jan. 2025. 1

[30] N. Methani, P. Ganguly, M. M. Khapra, and P. Kumar. PlotQA: Reasoning over Scientific Plots. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1516–1525, Mar. 2020. doi: 10.1109/WACV45572.2020.9093523 1

[31] T. Munzner. *Visualization Analysis and Design*. CRC Press, Dec. 2014. 1, 3

[32] M. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In J. DeNero, M. Finlayson, and S. Reddy, eds., *ACL*, pp. 97–101. ACL, San Diego, California, June 2016. doi: 10.18653/v1/N16-3020 1

[33] R. E. Roth. Visual Variables. In *International Encyclopedia of Geography*, pp. 1–11. John Wiley & Sons, Ltd, 2017. doi: 10.1002/9781118786352.wbieg0761 2

[34] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, and D. A. Keim. Knowledge Generation Model for Visual Analytics. *TVCG*, 20(12):1604–1613, Dec. 2014. doi: 10.1109/TVCG.2014.2346481 1, 3

[35] A. Salamatian et al. ChartGaze: Enhancing Chart Understanding in LVLMs with Eye-Tracking Guided Attention Refinement. In C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, eds., *EMNLP 2025*, pp. 12104–12124. ACL, Suzhou, China, Nov. 2025. doi: 10.18653/v1/2025.emnlp-main.607 1, 3

[36] R. R. Selvaraju et al. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. pp. 618–626, 2017. 1

[37] R. R. Selvaraju et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE ICCV*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74 2

[38] C. Stokes, A. Arunkumar, M. A. Hearst, and L. Padilla. An Analysis of Text Functions in Information Visualization, July 2025. doi: 10.48550/arXiv.2507.12334 2

[39] B. J. Tang, A. Boggust, and A. Satyanarayan. VisText: A Benchmark for Semantically Rich Chart Captioning, June 2023. doi: 10.48550/arXiv.2307.05356 1

[40] X. Wang, Y. Cui, X. Yao, S. Wang, G. Hu, and X. Qin. ChartHal: A Fine-grained Framework Evaluating Hallucination of Large Vision Language Models in Chart Understanding, Sept. 2025. doi: 10.48550/arXiv.2509.17481 4

[41] Y. Wang, W. Wang, A. Abdelhafez, M. Elfares, Z. Hu, M. Bâce, and A. Bulling. Salchartqa: Question-driven saliency on information visualisations. In *CHI 2024*, pp. 1–14, 2024. 1, 3

[42] S. Wiegreffe and Y. Pinter. Attention is not not Explanation. In K. Inui, J. Jiang, V. Ng, and X. Wan, eds., *EMNLP-IJCNLP 2019*, pp. 11–20. ACL, Hong Kong, China, Nov. 2019. doi: 10.18653/v1/D19-1002 3

[43] L. Wilkinson. *The Grammar of Graphics*. Statistics and Computing. Springer-Verlag, New York, 2005. doi: 10.1007/0-387-28695-0 2

[44] L. Wilkinson. *The Grammar of Graphics*. Springer Science & Business Media, Mar. 2013. 1

[45] Y. Zhao et al. LEVA: Using Large Language Models to Enhance Visual Analytics. *TVCG*, 31(3):1830–1847, Mar. 2025. doi: 10.1109/TVCG.2024.3368060 1