Do Vision-Language Models See Visualizations Like Humans? Alignment in Chart Categorization

Péter Ferenc Gyarmati*
University of Vienna

Manfred Klaffenböck†
TU Wien
University of Vienna

Laura Kösten‡
MBZUAI
Austrian Institute of Technology
University of Vienna

Torsten Möller§
University of Vienna

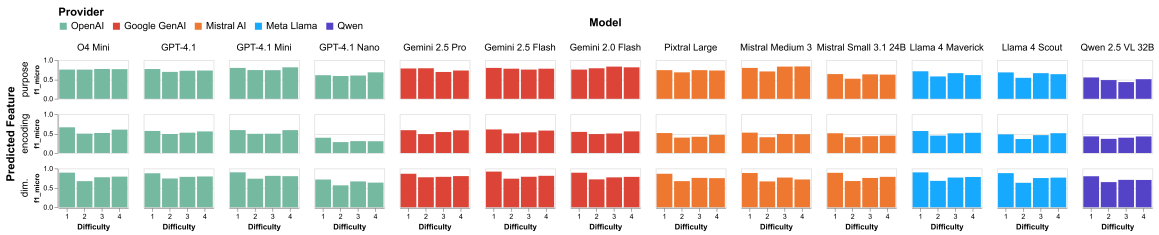

Figure 1: Micro-averaged F1-scores of Vision-language models (VLMs) in categorizing visualizations by PURPOSE, ENCODING, and DIMENSIONALITY, by human-assessed difficulty (1=easiest, 4=hardest). Higher bars mean better alignment with human expert perception. VLMs generally perform well on purpose and dimensionality but struggle with specific encoding types.

ABSTRACT

Vision-language models (VLMs) hold promise for enhancing visualization tools, but effective human-AI collaboration hinges on a shared perceptual understanding of visual content. Prior studies assessed VLM visualization literacy through interpretive tasks, revealing an over-reliance on textual cues rather than genuine visual analysis [1, 4]. Our study investigates a more foundational skill underpinning such literacy: the ability of VLMs to recognize a chart's core visual properties as humans do. We task 13 diverse VLMs with classifying scientific visualizations based solely on visual stimuli, according to three criteria: PURPOSE (e.g., *schematic*, *GUI*, *visualization*), ENCODING (e.g., *bar*, *point*, *node-link*), and DIMENSIONALITY (e.g., *2D*, *3D*). Using expert labels from the human-centric VisType typology [2] as ground truth, we find that VLMs often identify PURPOSE and DIMENSIONALITY accurately but struggle with specific ENCODING types. Our preliminary results show that larger models do not always equate to superior performance and highlight the need for careful integration of VLMs in visualization tasks, with human supervision to ensure reliable outcomes.

1 INTRODUCTION

The integration of vision-language models (VLMs) into visualization tools promises to enhance data analysis [1]. However, effective human-AI collaboration depends on shared VLM-human perception. A perceptual gap can make VLM assistance unreliable or counterproductive. Measuring this gap is therefore a critical first step toward developing trustworthy AI-integrated visualization systems.

Existing research has evaluated VLM visualization literacy through analytical and interpretive tasks [1, 4], finding that VLMs often lean on textual prompts or pre-existing knowledge. We assess a more foundational skill: recognizing a chart's core visual properties, as human experts do. Before a model can *interpret* a visualization's meaning, it must first accurately *recognize* its fundamental attributes, such as its purpose, perceived dimensionality, and visual encoding. This foundational capability is critical for downstream applications. For instance, a tool for grammar-based editing cannot act on a command like "make the bars thicker" if it fails to recognize the "bars" in the first place. Effective critique depends on the same prerequisite. We therefore explore how closely VLM perception aligns with that of human experts on this foundational parsing task, as a first step toward building more trustworthy and human-centered visualization tools.

Our study leverages two key resources. First, the VIS30K dataset [3] contains nearly 30,000 figures from 30 years of IEEE Visualization conference publications, offering a rich source of real-world scientific visualizations. Second, we leverage the VisType typology [2] as a human-annotated ground truth. This framework was developed by analyzing a subset of VIS30K, focusing on the *essential stimuli* of each image. It provides expert labels for an image's primary PURPOSE (e.g., a conceptual *schematic*, a *GUI* screenshot, or a data *visualization*), its constituent visual ENCODING (e.g., *bar*, *line*, *node-link*), its perceived DIMENSIONALITY (e.g., *2D*, *3D*), and human-assessed difficulty of the categorization task itself.

Our experiment tests VLM perceptual alignment by restricting models to only two inputs: the raw figure image—which may include embedded text like axis labels, but excludes surrounding captions or body text—and the VisType category definitions provided via system prompt. This visual-only approach forces reliance on image analysis, allowing direct comparison with human experts and assessment of performance across models, scale, and human-perceived task difficulty.

2 EXPERIMENTAL SETUP

Our experiment has two phases: VLM inference on images and evaluation against expert labels.

Dataset and Models We use the dataset from Chen et al.'s VisImageNavigator application[1], derived from VIS30K. The labels, based on the VisType typology [2], define images by PURPOSE, ENCODING, DIMENSIONALITY, and perceived HARDNESS. A stratified sample of 305 images ensures representation across these categories. We test 13 diverse VLMs from five providers, spanning flagship and efficient models, as shown in Fig. 1.

*e-mail: peter.ferenc.gyarmati@univie.ac.at
†e-mail: klaffenboeck@cg.tuwien.ac.at
‡e-mail: laura.koesten@mbzuai.ac.ae
§e-mail: torsten.moeller@univie.ac.at

[1]Labeled dataset: VisImageNavigator/VisImageNavigator.github.io

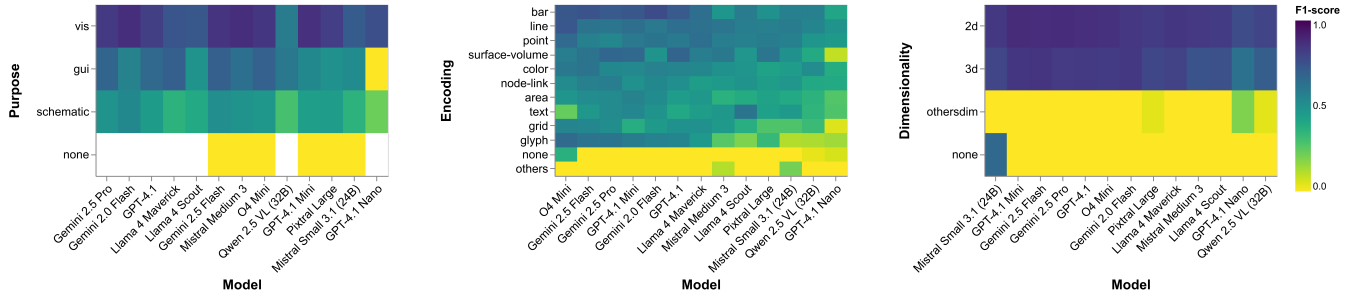

Figure 2: F1-score heatmap comparing VLM classification to expert judgment for PURPOSE, ENCODING, and DIMENSIONALITY. Models (x-axis) and classification labels (y-axis) are sorted in descending order by their average F1-score. Darker colors mean higher agreement, yellow cells (F1-score ≈ 0) indicate very low alignment, and white cells show where a model never produced the corresponding label at all. A *none* label indicates cases where experts intentionally omitted a category. VLMs generally perform better at identifying *2D* dimensionality, simpler encodings like *bar*, *line*, *point*, and *vis* (visualization example) purpose. However, classifying specific encoding types remains challenging. The prevalence of yellow and white cells reveals that models tend to offer an incorrect label rather than abstain, a behavior underscored by their difficulty predicting the *none* category.

Inference and Evaluation Process Each VLM processes the 305 images in a zero-shot setting using default hyperparameters. The system prompt contains only the exact textual definitions for each category from Table 1 of Chen et al.'s VisType typology [2]. We deliberately omitted visual examples to avoid priming the models' visual recognition. To handle structured output and uncertainty, we used models' native tool-calling to enforce a JSON schema. The schema definition for each category explicitly allowed a *null* value, providing models a direct way to abstain when uncertain. For evaluation, we compare VLM outputs with the expert ground truth using standard multi-label classification metrics (e.g., F1-score). Code, prompts and results are openly available[2].

3 Preliminary Findings & Discussion

Our initial results (Fig. 1–2) reveal trends in VLM visualization perception. We report micro-averaged F1-scores, suitable for this multi-label task.

VLMs generally identify image PURPOSE (e.g., *vis*, *schematic*) and DIMENSIONALITY (especially clear *2D*) with reasonable accuracy. However, discerning specific visual ENCODING types is a significant challenge. Fig. 2 (middle) shows that while simpler encodings like *bar*, *line*, and *point* achieve moderate alignment, complex types like *glyph*, *grid*, or *node–link*–and even *color* or *surface–volume*–are often misidentified. This suggests that VLMs, when relying solely on visual input without additional contextual guidance, struggle to capture the nuanced visual features human experts rely on when categorizing these encodings.

Notably, for this categorization task, larger models do not consistently outperform smaller ones. Some advanced models (e.g., Gemini 2.5 Pro, GPT-4.1) may better identify PURPOSE but show no corresponding gain in recognizing specific ENCODINGS. This implies that for this task, model scale or recency alone does not guarantee better perceptual alignment.

Despite the option for a *null* response for uncertainty, VLMs rarely used it, often assigning an incorrect category instead. This overconfidence is critical: a system misidentifying a line chart as a bar chart will offer nonsensical feedback, eroding trust. The *none* category in Fig. 2 (where experts intentionally omitted a label) was also poorly predicted, underscoring this behavior. For instance, in identifying image PURPOSE, 7 of the 13 tested models never produced a *none* label, always forcing a different classification.

Human-assessed difficulty (Fig. 1, x-axis) generally correlates with VLM performance: images that are harder to label for humans also tend to be harder for models. However, this correlation should not be mistaken for perceptual agreement. The models' poor performance on many charts labeled as 'easy' by human experts indicates that VLMs struggle with nuanced visual features. This reveals a perceptual gap: visual ambiguities that are easy for humans present challenges for VLMs.

4 Conclusion & Next Steps

While modern VLMs make it easy to prompt a complex classifier from definitions, this accessibility masks a significant perceptual gap. Our findings show VLMs identify a chart's PURPOSE and DIMENSIONALITY but struggle with the fine-grained ENCODINGS, showing only partial alignment with expert perception. Model scale does not close this gap, and systems often guess incorrectly rather than express uncertainty. This cautions against blindly trusting VLMs and underscores the need for human-in-the-loop AI support. Next, we will extend the benchmark to all labeled samples in the VIS30K corpus, inject controlled textual cues to probe truly multimodal reasoning, and quantify when a model's uncertainty is reliable enough for iterative design feedback. As VLMs evolve with greater performance and deeper multimodality, such benchmarks become crucial for tracking whether these advances translate to better human-AI alignment and trust. By tracking progress with this open evaluation framework, we aim to guide both VLM development and visualization practice toward more transparent human-AI alignment in visualization systems.

[2]Code and data: peter-gy/AutoVisType

References

[1] A. Bendeck and J. Stasko. An Empirical Evaluation of the GPT-4 Multimodal Language Model on Visualization Literacy Tasks. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):1105–1115, 2025. doi: 10.1109/TVCG.2024.3456155

[2] J. Chen, P. Isenberg, R. S. Laramee, T. Isenberg, M. Sedlmair, T. Möller, and R. Li. An Image-based Typology for Visualization, 2025. arXiv preprint.

[3] J. Chen, M. Ling, R. Li, P. Isenberg, T. Isenberg, M. Sedlmair, T. Möller, R. S. Laramee, H.-W. Shen, K. Wünsche, and Q. Wang. VIS30K: A Collection of Figures and Tables From IEEE Visualization Conference Publications. *IEEE Transactions on Visualization and Computer Graphics*, 27(9):3826–3833, 2021. doi: 10.1109/TVCG.2021.3054916

[4] J. Hong, C. Seto, A. Fan, and R. Maciejewski. Do LLMs have visualization literacy? an evaluation on modified visualizations to test generalization in data interpretation. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–13, 2025. doi: 10.1109/TVCG.2025.3536358