VIS 2025

Generalization of CNNs on Relational Reasoning with Bar Charts

Zhenxing Cui*, Lu Chen*, Yunhai Wang, Daniel Haehn, **Yong Wang**, Hanspeter Pfister

Shandong University · Zhejiang University · Renmin University of China · UMass Boston · Nanyang Technological University Singapore · Harvard

Background

- Convolutional Neural Networks (CNNs) are widely used for many visualization tasks

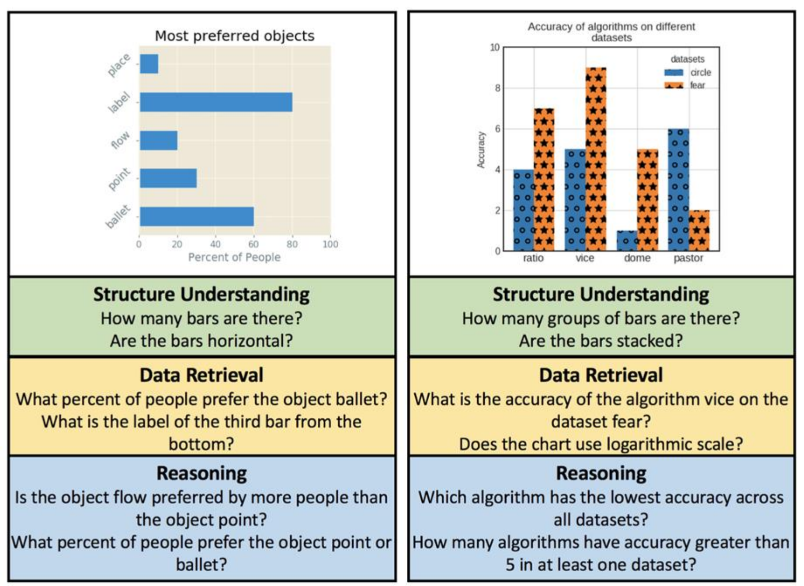
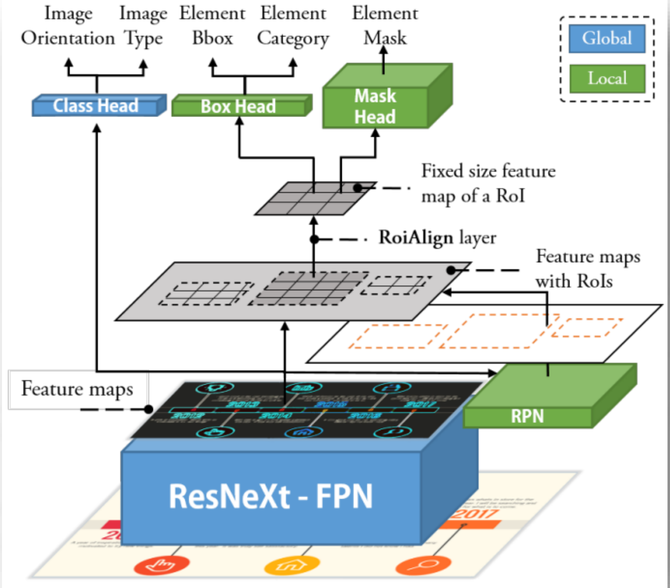


Chart Question Answering



Automated Chart Design

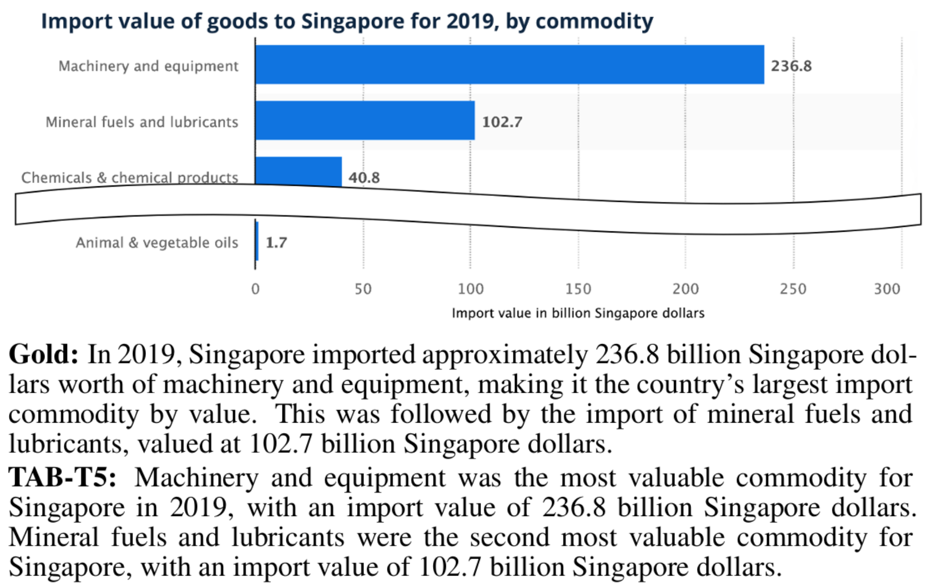
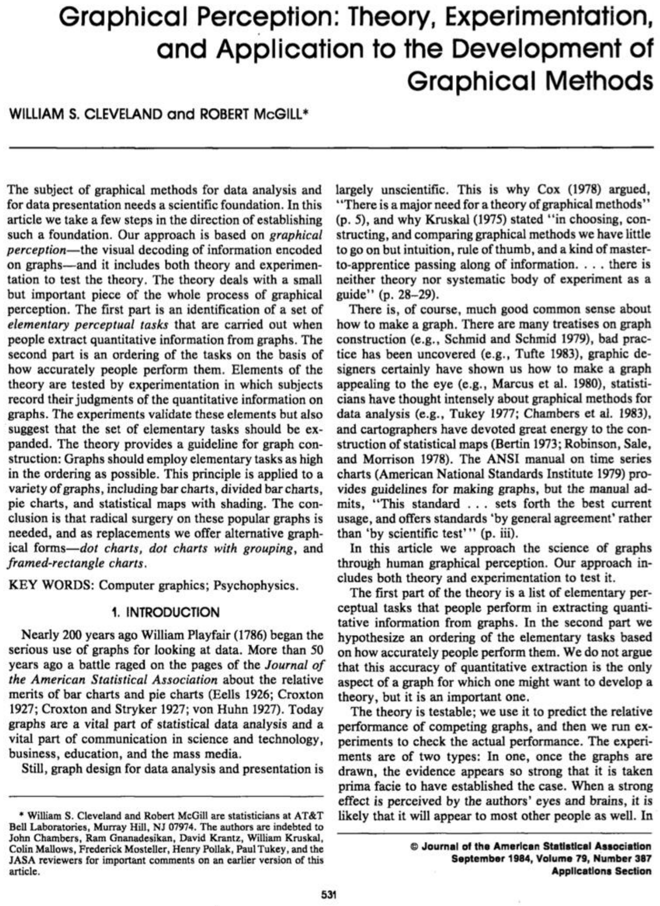


Chart Captioning

- But it remains underexplored how CNNs' graphical perception performance generalizes across visualization design variations

Background

- Graphical Perception: the ability to decode visually encoded quantities in visualizations



Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods

WILLIAM S. CLEVELAND and ROBERT McGILL*

The subject of graphical methods for data analysis and for data presentation needs a scientific foundation. In this article we take a few steps in the direction of establishing such a foundation. Our approach is based on *graphical perception*—the visual decoding of information encoded on graphs—and it includes both theory and experimentation to test the theory. The theory deals with a small but important piece of the whole process of graphical perception. The first part is an identification of a set of *elementary perceptual tasks* that are carried out when people extract quantitative information from graphs. The second part is an ordering of the tasks on the basis of how accurately people perform them. Elements of the theory are tested by experimentation in which subjects record their judgments of the quantitative information on graphs. The experiments validate these elements but also suggest that the set of elementary tasks should be expanded. The theory provides a guideline for graph construction: Graphs should employ elementary tasks as high in the ordering as possible. This principle is applied to a variety of graphs, including bar charts, divided bar charts, pie charts, and statistical maps with shading. The conclusion is that radical surgery on these popular graphs is needed, and as replacements we offer alternative graphical forms—*dot charts, dot charts with grouping*, and *framed-rectangle charts*.

KEY WORDS: Computer graphics; Psychophysics.

1. INTRODUCTION

Nearly 200 years ago William Playfair (1786) began the serious use of graphs for looking at data. More than 50 years ago a battle raged on the pages of the *Journal of the American Statistical Association* about the relative merits of bar charts and pie charts (Eells 1926; Croxton 1927; Croxton and Stryker 1927; von Huhn 1927). Today graphs are a vital part of statistical data analysis and a vital part of communication in science and technology, business, education, and the mass media.

Still, graph design for data analysis and presentation is largely unscientific. This is why Cox (1978) argued, "There is a major need for a theory of graphical methods" (p. 5), and why Kruskal (1975) stated "in choosing, constructing, and comparing graphical methods we have little to go on but intuition, rule of thumb, and a kind of master-to-apprentice passing along of information. . . . there is neither theory nor systematic body of experiment as a guide" (p. 28–29).

There is, of course, much good common sense about how to make a graph. There are many treatises on graph construction (e.g., Schmid and Schmid 1979), bad practice has been uncovered (e.g., Tufte 1983), graphic designers certainly have shown us how to make a graph appealing to the eye (e.g., Marcus et al. 1980), statisticians have thought intensely about graphical methods for data analysis (e.g., Tukey 1977; Chambers et al. 1983), and cartographers have devoted great energy to the construction of statistical maps (Bertin 1973; Robinson, Sale, and Morrison 1978). The ANSI manual on time series charts (American National Standards Institute 1979) provides guidelines for making graphs, but the manual admits, "This standard . . . sets forth the best current usage, and offers standards 'by general agreement' rather than 'by scientific test'" (p. iii).

In this article we approach the science of graphs through human graphical perception. Our approach includes both theory and experimentation to test it.

The first part of the theory is a list of elementary perceptual tasks that people perform in extracting quantitative information from graphs. In the second part we hypothesize an ordering of the elementary tasks based on how accurately people perform them. We do not argue that this accuracy of quantitative extraction is the only aspect of a graph for which one might want to develop a theory, but it is an important one.

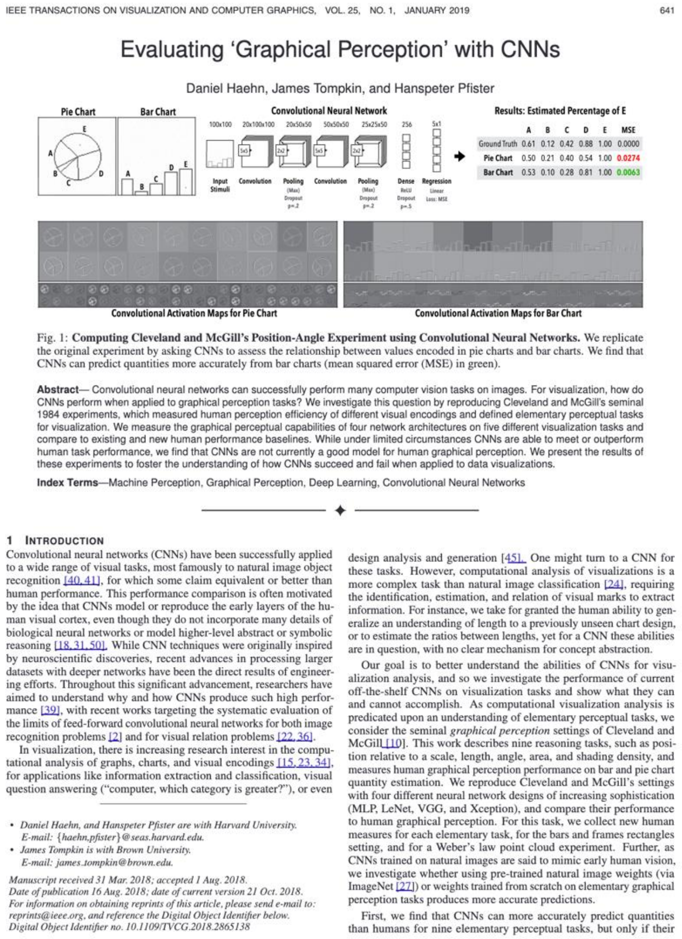
The theory is testable; we use it to predict the relative performance of competing graphs, and then we run experiments to check the actual performance. The experiments are of two types: In one, once the graphs are drawn, the evidence appears so strong that it is taken prima facie to have established the case. When a strong effect is perceived by the authors' eyes and brains, it is likely that it will appear to most other people as well. In

* William S. Cleveland and Robert McGill are statisticians at AT&T Bell Laboratories, Murray Hill, NJ 07974. The authors are indebted to John Chambers, Ram Gnanadesikan, David Krantz, William Kruskal, Colin Mallows, Frederick Mosteller, Henry Pollak, Paul Tukey, and the JASA reviewers for important comments on an earlier version of this article.

© Journal of the American Statistical Association
September 1984, Volume 79, Number 387
Applications Section

531

Cleveland & McGill, 1984



IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 25, NO. 1, JANUARY 2019 641

Evaluating 'Graphical Perception' with CNNs

Daniel Haehn, James Tompkin, and Hanspeter Pfister

Pie Chart
Bar Chart
Convolutional Neural Network
100x100 20x100x100 20x50x50 50x50x50 25x25x50 256 5x1
Input Stimuli
Convolution
Pooling (Max) Dropout p=.2
Convolution
Pooling (Max) Dropout p=.2
Dense ReLU Dropout p=.5
Regression Linear Loss: MSE

Results: Estimated Percentage of E

	A	B	C	D	E	MSE
Ground Truth	0.61	0.12	0.42	0.88	1.00	0.0000
Pie Chart	0.50	0.21	0.40	0.54	1.00	0.0274
Bar Chart	0.53	0.10	0.28	0.81	1.00	0.0063

Convolutional Activation Maps for Pie Chart
Convolutional Activation Maps for Bar Chart

Fig. 1: **Computing Cleveland and McGill's Position-Angle Experiment using Convolutional Neural Networks.** We replicate the original experiment by asking CNNs to assess the relationship between values encoded in pie charts and bar charts. We find that CNNs can predict quantities more accurately from bar charts (mean squared error (MSE) in green).

Abstract— Convolutional neural networks can successfully perform many computer vision tasks on images. For visualization, how do CNNs perform when applied to graphical perception tasks? We investigate this question by reproducing Cleveland and McGill's seminal 1984 experiments, which measured human perception efficiency of different visual encodings and defined elementary perceptual tasks for visualization. We measure the graphical perceptual capabilities of four network architectures on five different visualization tasks and compare to existing and new human performance baselines. While under limited circumstances CNNs are able to meet or outperform human task performance, we find that CNNs are not currently a good model for human graphical perception. We present the results of these experiments to foster the understanding of how CNNs succeed and fail when applied to data visualizations.

Index Terms—Machine Perception, Graphical Perception, Deep Learning, Convolutional Neural Networks

1 Introduction

Convolutional neural networks (CNNs) have been successfully applied to a wide range of visual tasks, most famously to natural image object recognition [40,41], for which some claim equivalent or better than human performance. This performance comparison is often motivated by the idea that CNNs model or reproduce the early layers of the human visual cortex, even though they do not incorporate many details of biological neural networks or model higher-level abstract or symbolic reasoning [18,31,50]. While CNN techniques were originally inspired by neuroscientific discoveries, recent advances in processing larger datasets with deeper networks have been the direct results of engineering efforts. Throughout this significant advancement, researchers have aimed to understand why and how CNNs produce such high performance [39], with recent works targeting the systematic evaluation of the limits of feed-forward convolutional neural networks for both image recognition problems [2] and for visual relation problems [22,36].

In visualization, there is increasing research interest in the computational analysis of graphs, charts, and visual encodings [15,23,34], for applications like information extraction and classification, visual question answering ("computer, which category is greater?"), or even design analysis and generation [45]. One might turn to a CNN for these tasks. However, computational analysis of visualizations is a more complex task than natural image classification [24], requiring the identification, estimation, and relation of visual marks to extract information. For instance, we take for granted the human ability to generalize an understanding of length to a previously unseen chart design, or to estimate the ratios between lengths, yet for a CNN these abilities are in question, with no clear mechanism for concept abstraction.

Our goal is to better understand the abilities of CNNs for visualization analysis, and so we investigate the performance of current off-the-shelf CNNs on visualization tasks and show what they can and cannot accomplish. As computational visualization analysis is predicated upon an understanding of elementary perceptual tasks, we consider the seminal *graphical perception* settings of Cleveland and McGill [10]. This work describes nine reasoning tasks, such as position relative to a scale, length, angle, area, and shading density, and measures human graphical perception performance on bar and pie chart quantity estimation. We reproduce Cleveland and McGill's settings with four different neural network designs of increasing sophistication (MLP, LeNet, VGG, and Xception), and compare their performance to human graphical perception. For this task, we collect new human measures for each elementary task, for the bars and frames rectangles setting, and for a Weber's law point cloud experiment. Further, as CNNs trained on natural images are said to mimic early human vision, we investigate whether using pre-trained natural image weights (via ImageNet [27]) or weights trained from scratch on elementary graphical perception tasks produces more accurate predictions.

First, we find that CNNs can more accurately predict quantities than humans for nine elementary perceptual tasks, but only if their

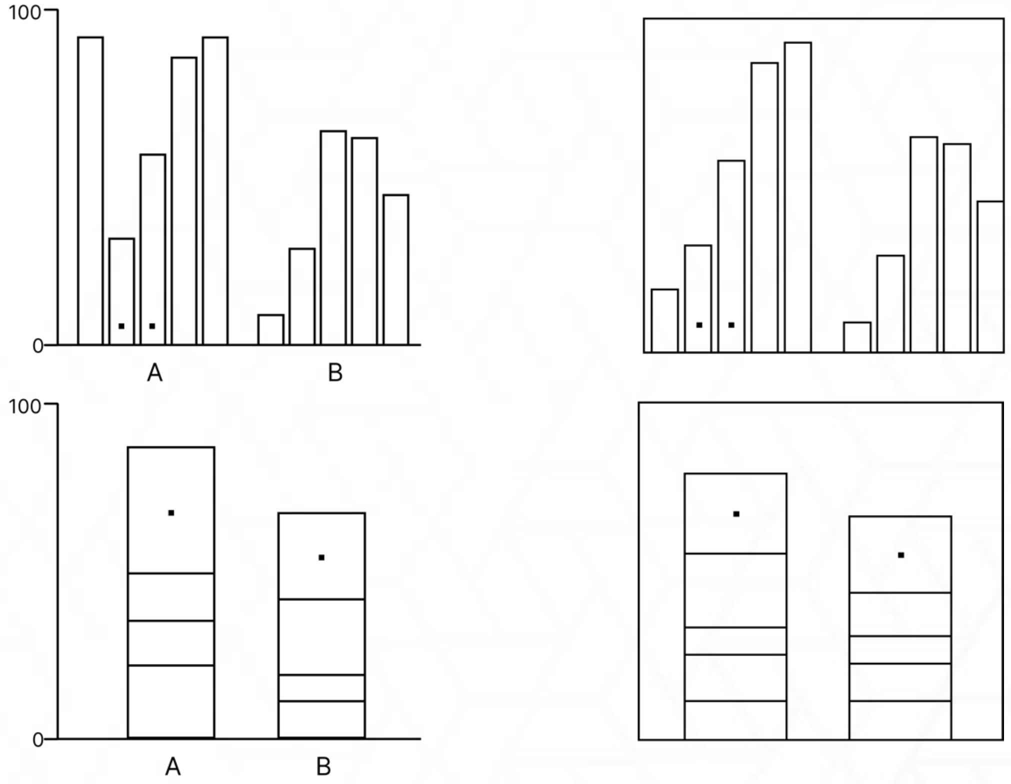
• *Daniel Haehn, and Hanspeter Pfister are with Harvard University. E-mail: {haehn,pfister}@seas.harvard.edu.*
• *James Tompkin is with Brown University. E-mail: james_tompkin@brown.edu.*

*Manuscript received 31 Mar. 2018; accepted 1 Aug. 2018.
Date of publication 16 Aug. 2018; date of current version 21 Oct. 2018.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TVCG.2018.2865138*

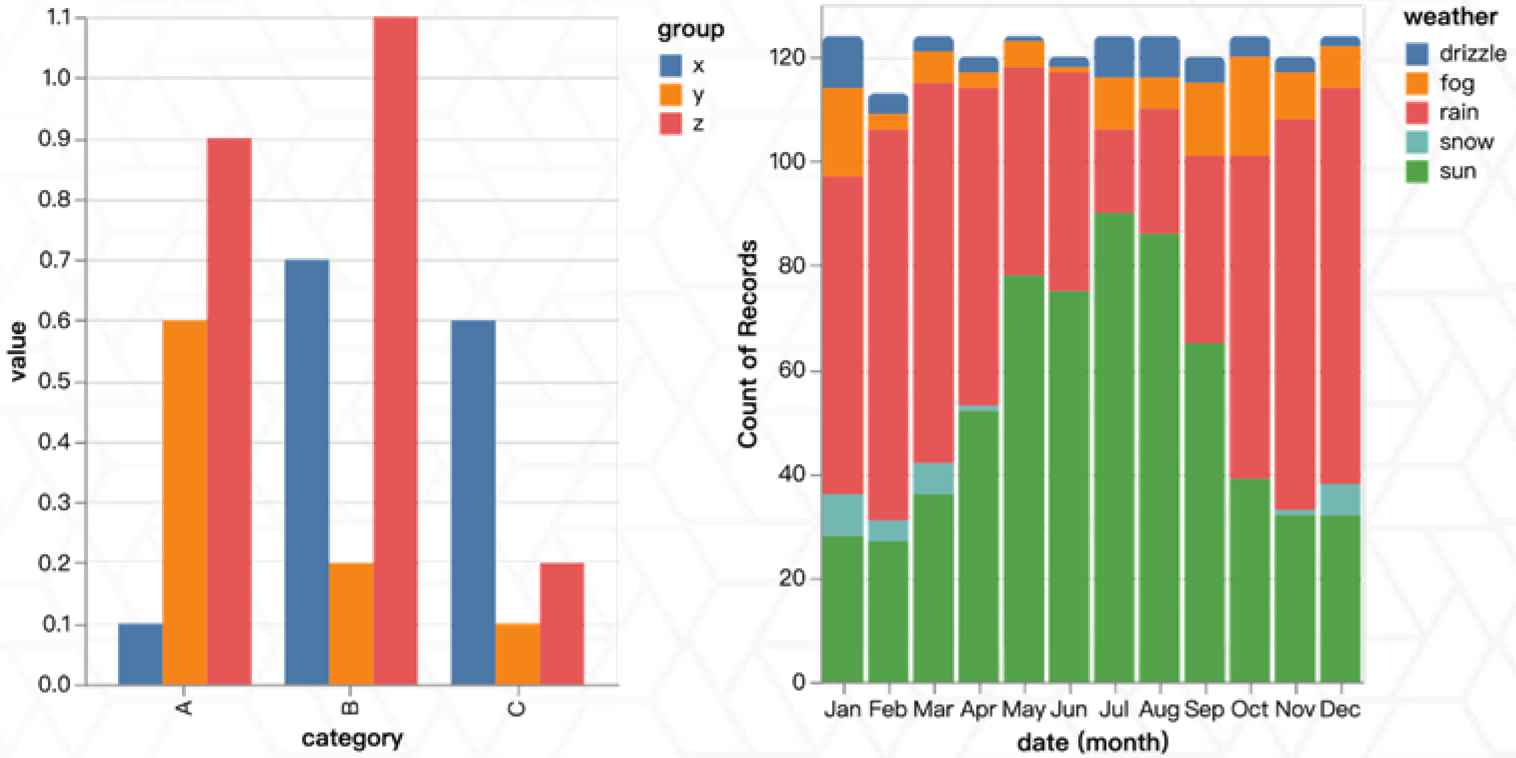
Haehn *et al.*, 2019

Background

- Graphical Perception: the ability to decode visually encoded quantities in visualizations



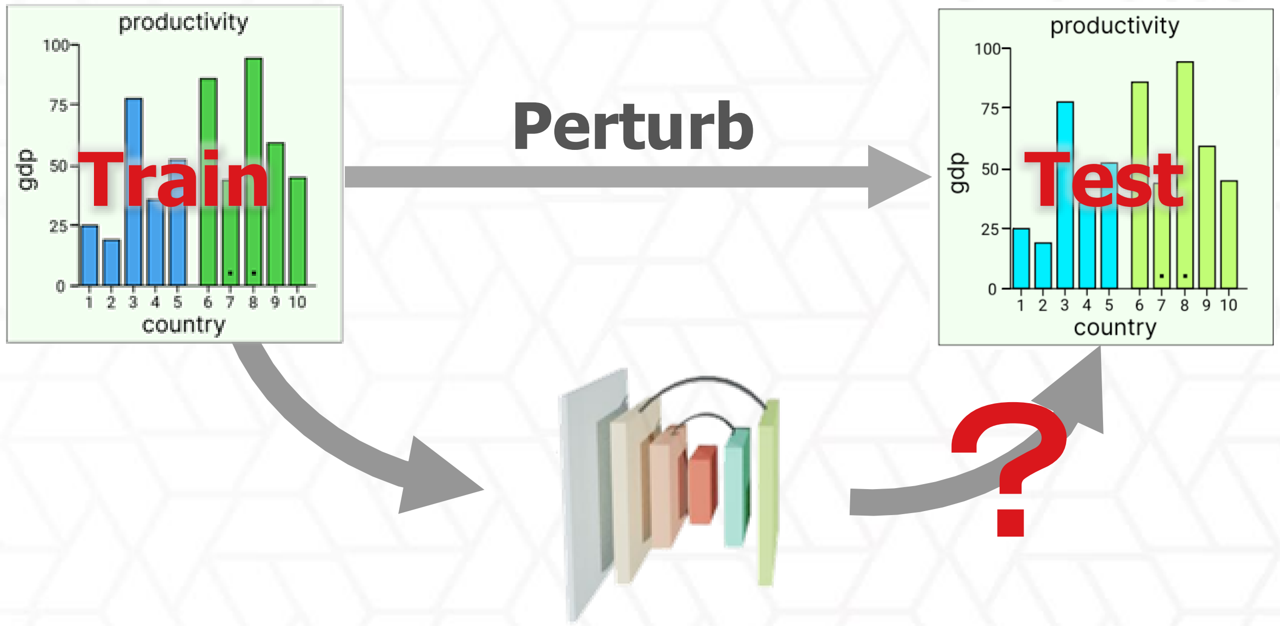
Oversimplified Charts



Standard Visualizations

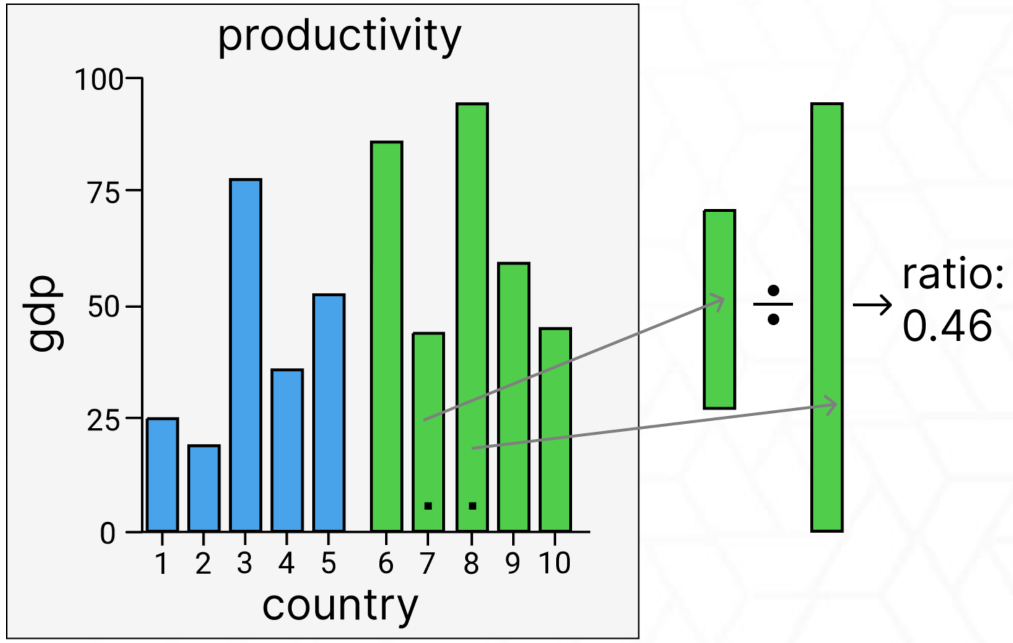
Research Questions

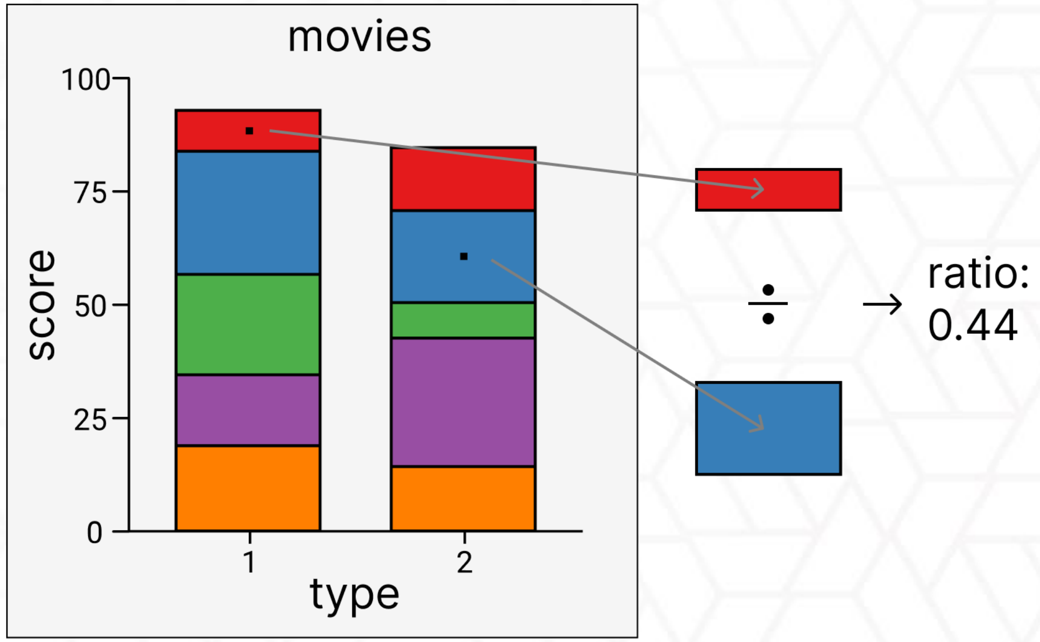
- How well do CNNs perform on standard visualizations with full design elements?
- How robust are CNNs to design perturbations such as color jitter?
- What are the differences between CNNs and humans in visual relational reasoning?



Relational Reasoning in Graphical Perception

- **Task**: Estimate the ratio of lengths (i.e., heights) between two target bars (targets indicated by black dots)





Benchmarking Representative CNNs

- Replicate Haehn *et al.*'s experiments [1] with systematically-tuned CNNs
- CNNs achieve very strong performance, better than previously-reported results

Architectures

- MLP
- AlexNet
- LeNet
- DenseNet
- VGG19
- ResNet152
- Xception126
- EfficientNet

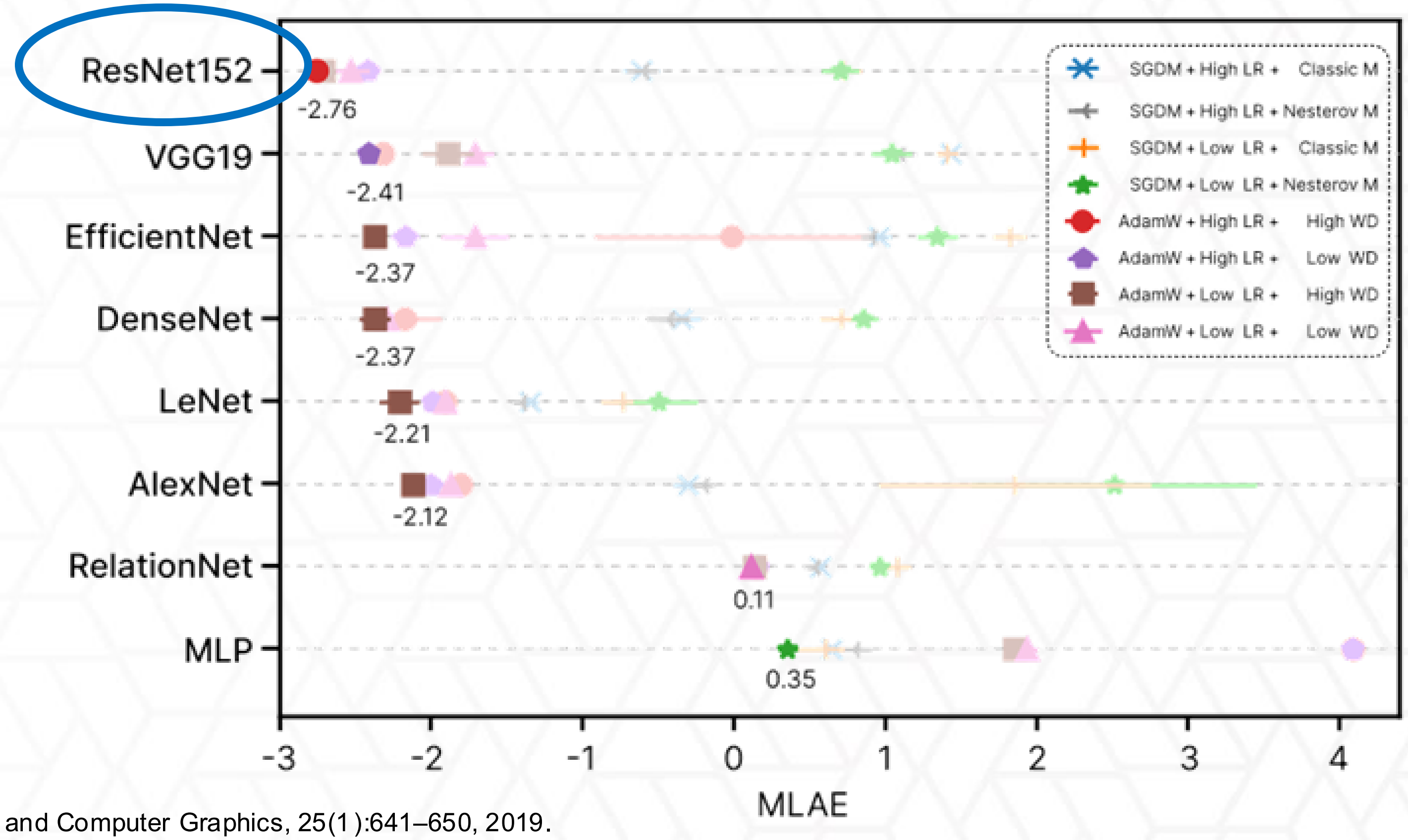
×

Optimizers

- AdamW
- SGDM

Hyper-parameters

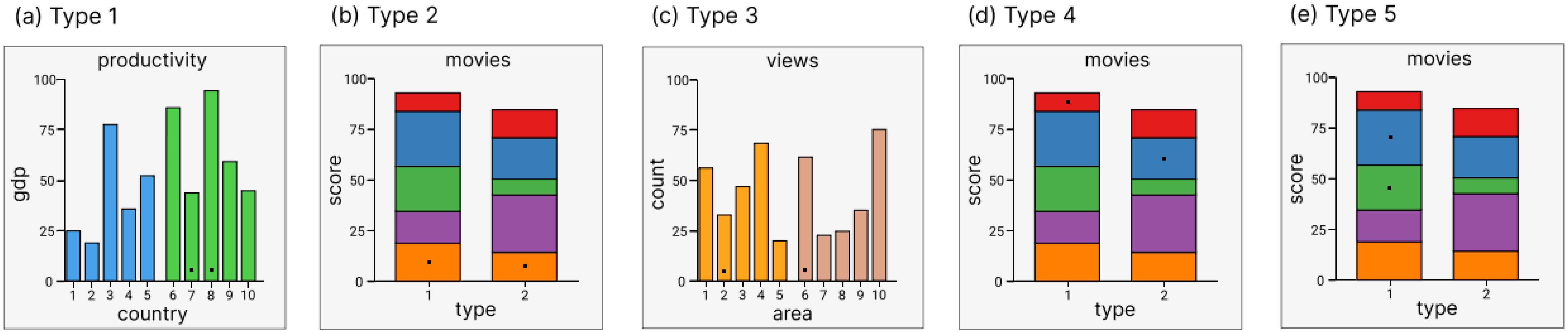
- Learning rate
- Momentum
- Weight decay



[1] D. Haehn, J. Tompkin, and H. Pfister. Evaluating 'graphical perception' with CNNs. IEEE Transactions on Visualization and Computer Graphics, 25(1):641–650, 2019.

GRAPE: A GRAphical PErception Dataset

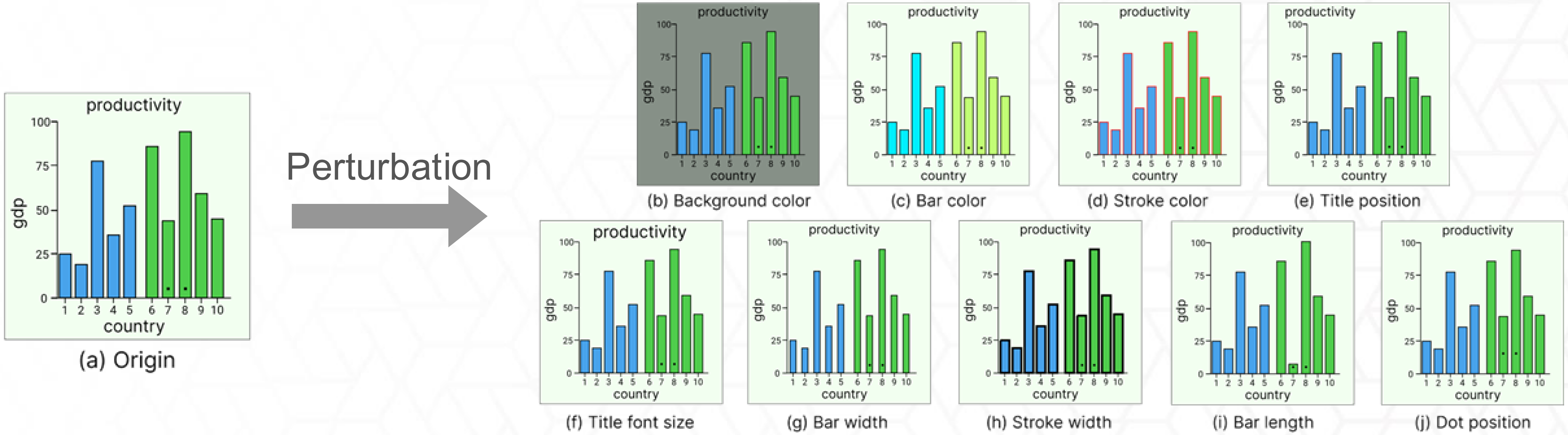
- Five bar-chart types synthesized programmatically with Vega-Lite
- 766K (500K for training & 266K for testing) standard visualizations
- Large and controllable dataset to manipulate design parameters



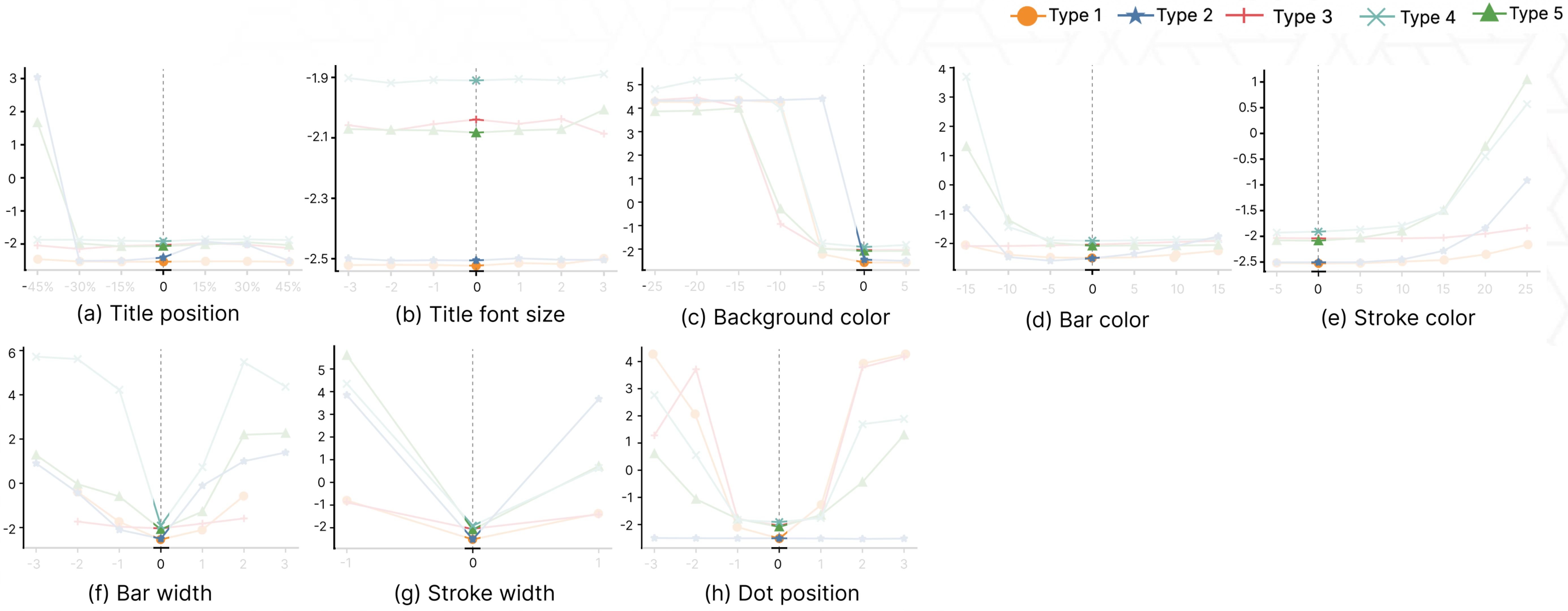
VIS 2025

Perturbation Setup

- 9 visual parameters: title position, title size, background color, bar color, stroke color, bar width, stroke width, bar length, dot position
- Independent and Identically Distributed (IID): test visualizations have similar encodings with the training samples
- Out-of-distribution (OOD): test and training visualizations are different

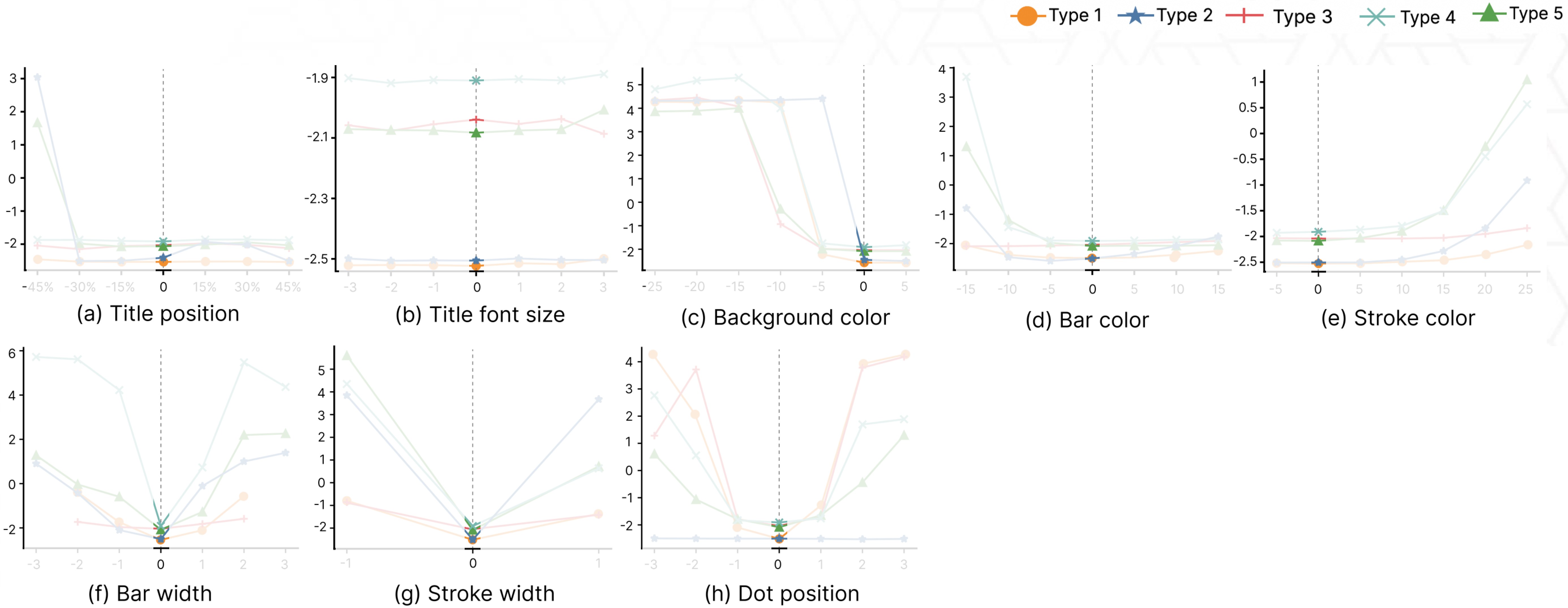


Results (IID): CNNs are Excellent



VIS2025

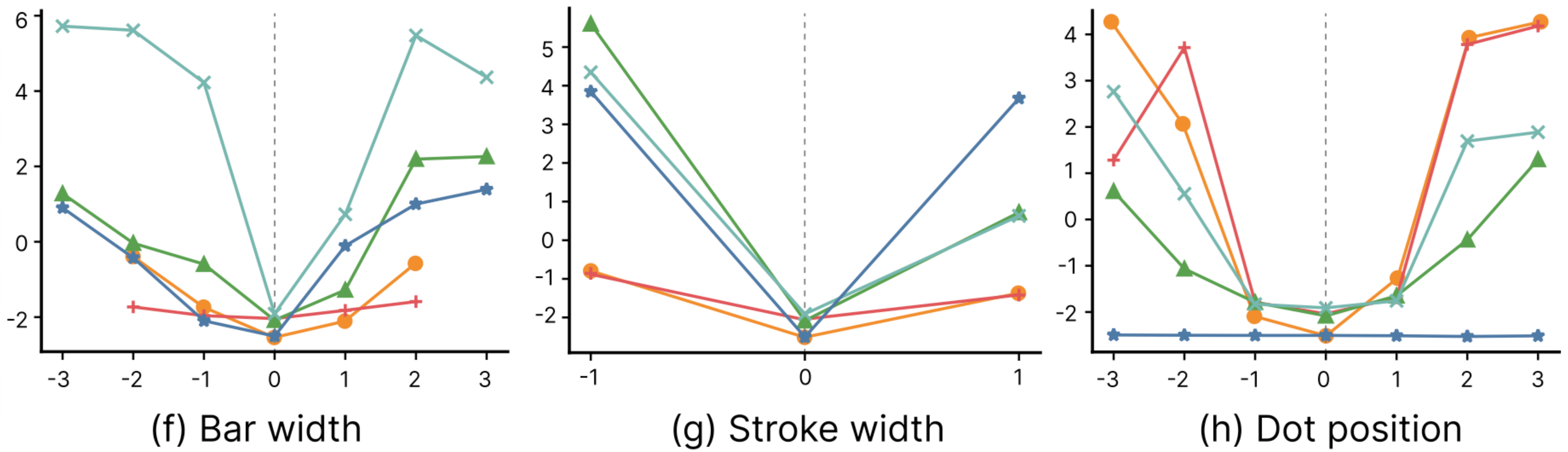
Results (OOD):
CNN Robustness Collapses on Visual Parameter Shifts



VIS2025

Results (OOD):
CNN Robustness Collapses on Visual Parameter Shifts

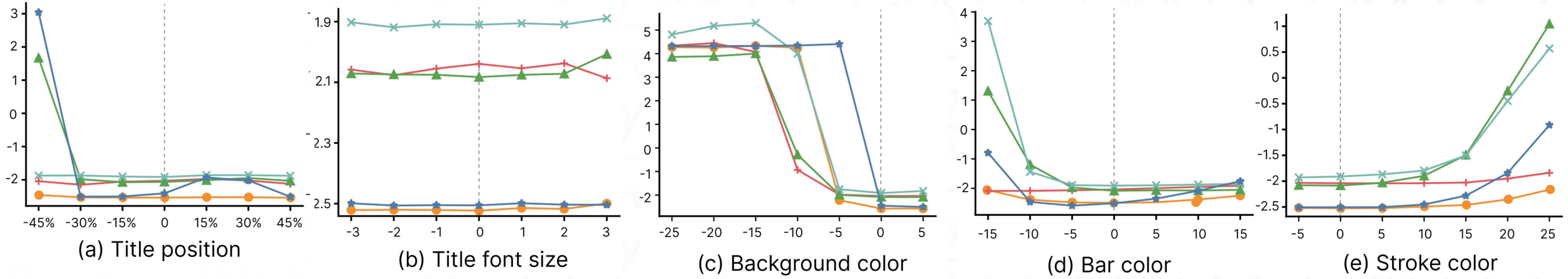
- Small changes in **bar width, stroke width**, and **dot position** → Big errors



VIS2025

Results (OOD): CNN Robustness Collapses on Visual Parameter Shifts

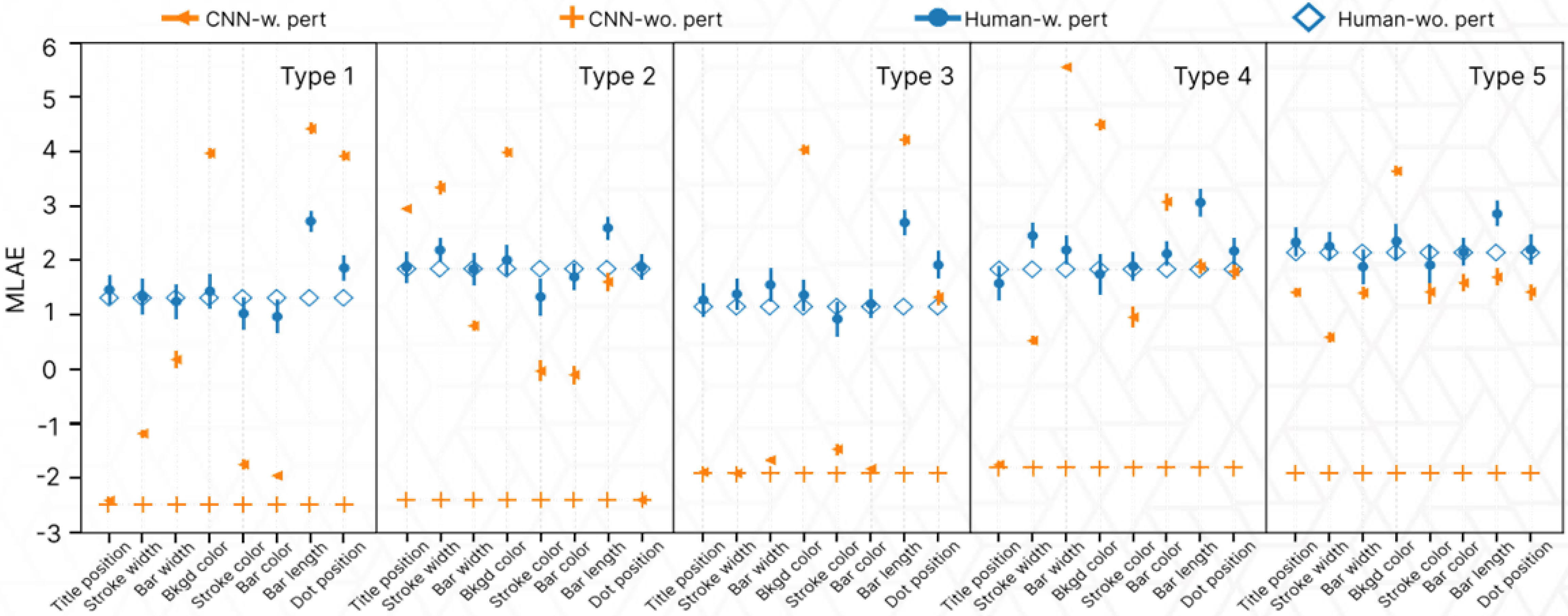
- Even "irrelevant" parameters (e.g., **title position**) affect graphic perception performance
- **Background luminance** drops sharply push up prediction error
- CNNs are relatively robust to changes of **title font size**, **bar and stroke colors**



VIS 2025

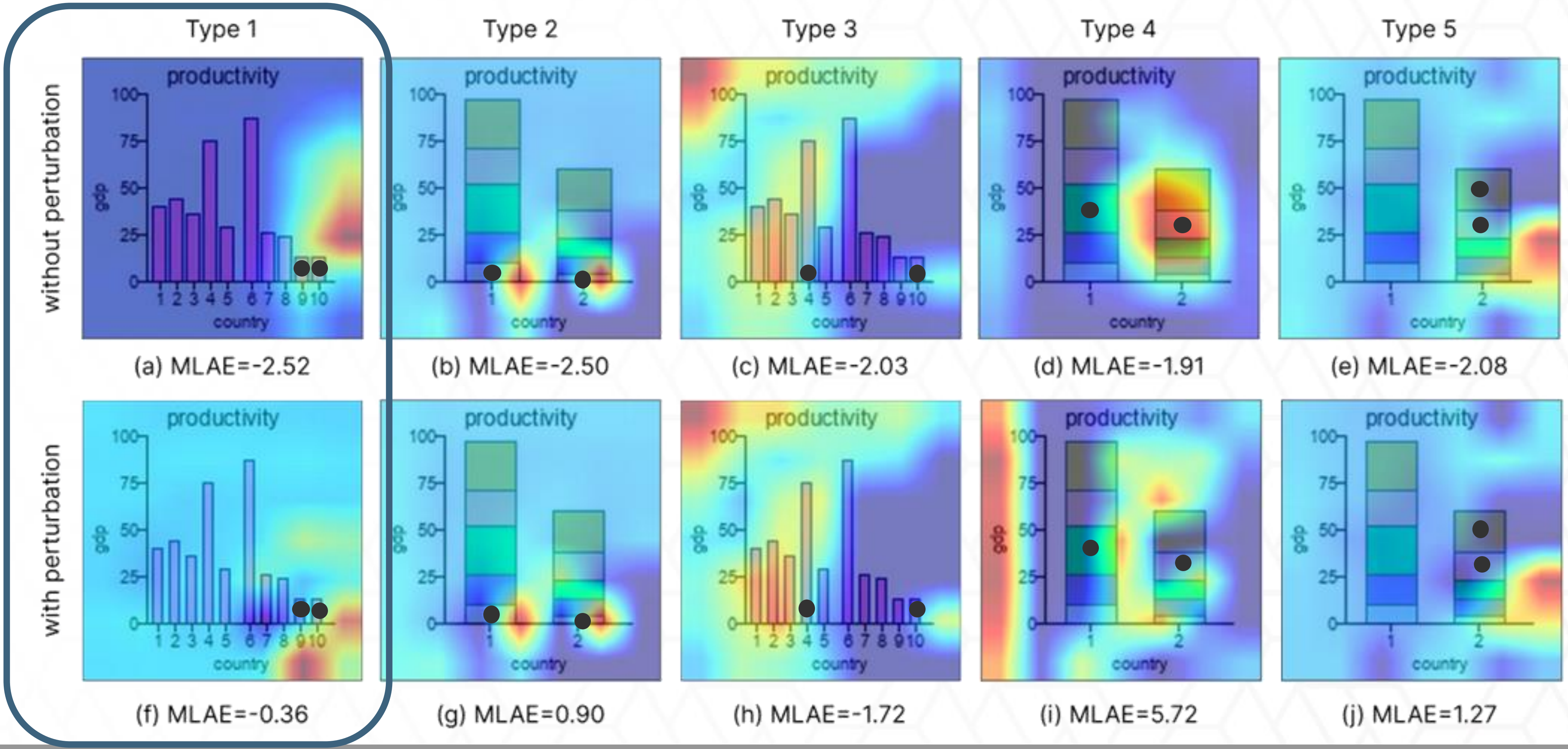
Humans vs CNNs

- Humans are worse than CNNs in IID conditions but more robust under perturbations (OOD)
- Interview feedback: participants focus on target bars; ignore nonessential styling



Why Do CNNs Fail OOD? Attention on the Wrong Pixels

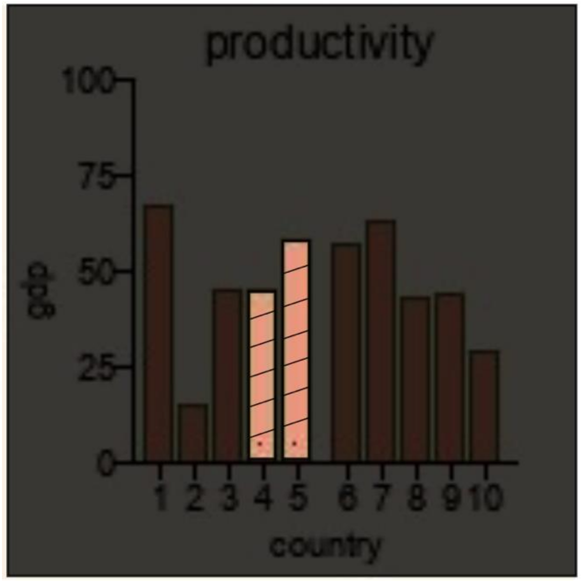
- Grad-CAM saliency maps rarely highlight target bars
- Model attention shifts noticeably under minor perturbations of bar width

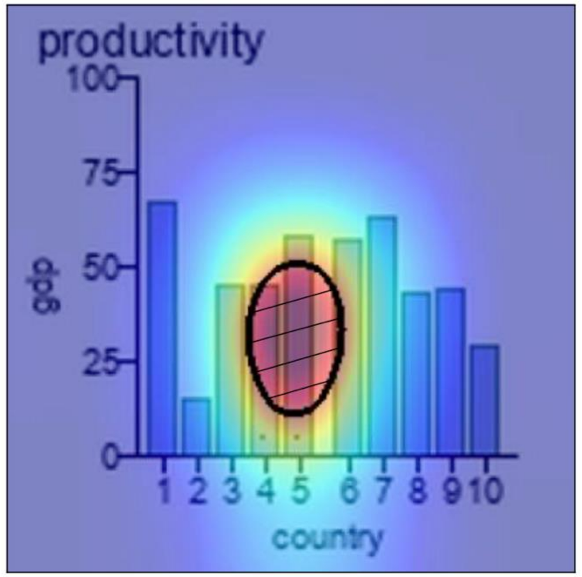


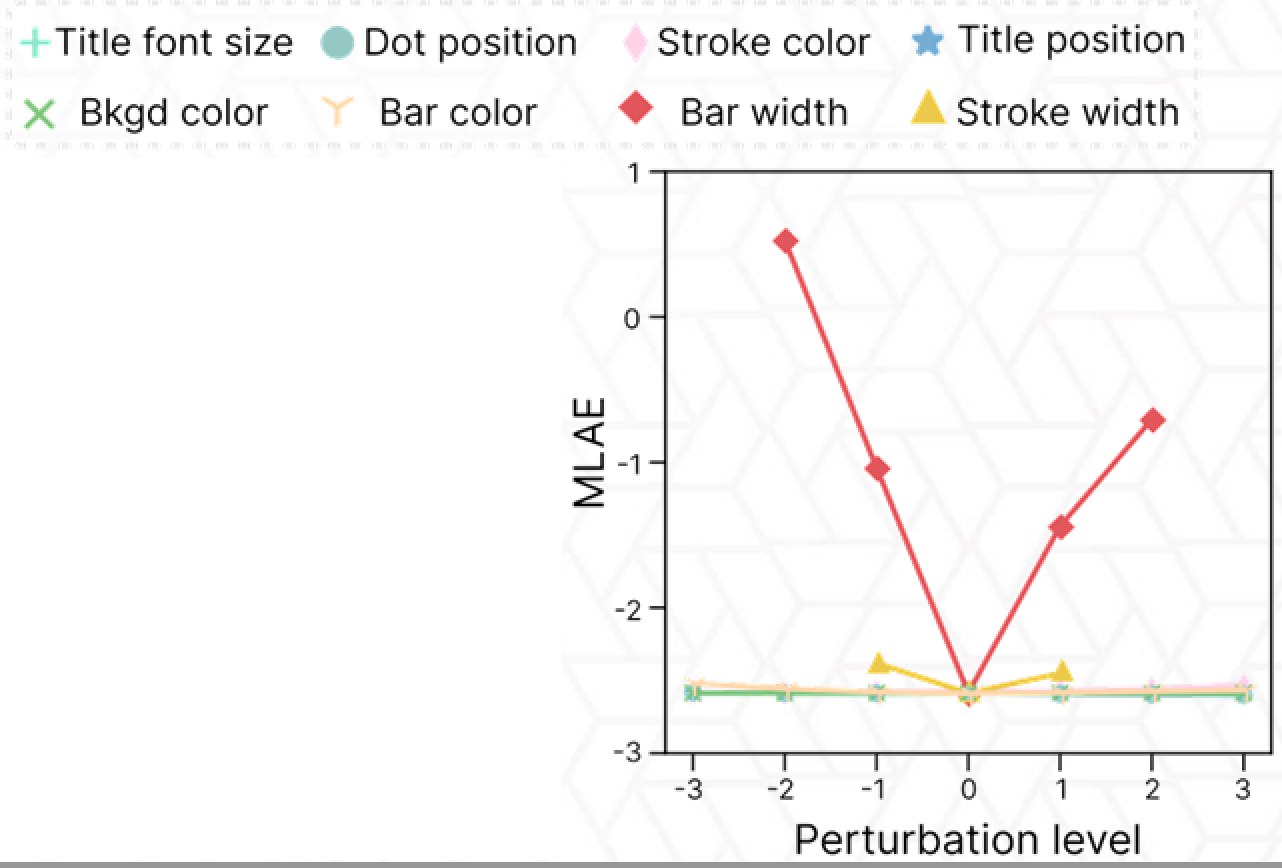
VIS 2025

Can We Fix It? Segmentation Masks

- Add a target-bar mask channel (RGB-α) → Better target localization, some robustness gains
- Yet, still sensitive to shape-related visual parameters like bar and stroke width

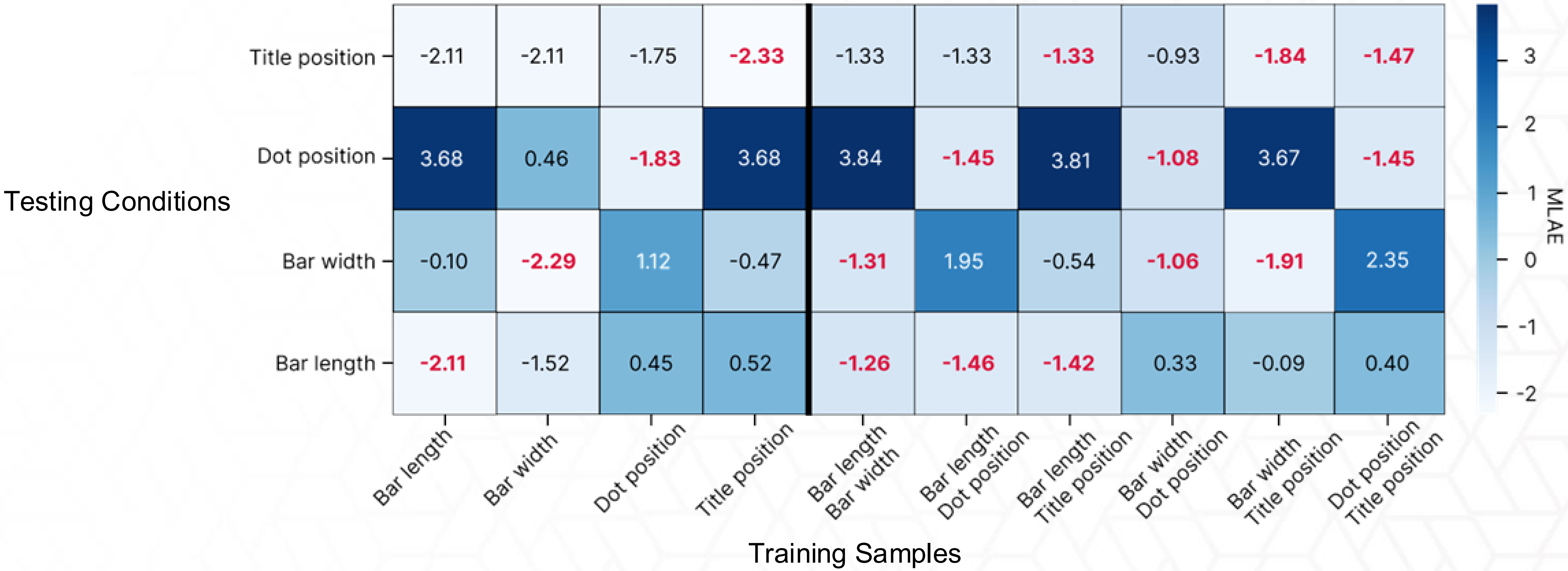






Does Data Augmentation Solve It?

- Performance improves on those perturbations seen at train time
- Generalization to unseen perturbations remains weak



VIS2025

Take-Home Messages

- Small and unseen shifts of visual parameters break CNNs’ graphical perception performance
- CNNs still fail to see charts like humans
- Simple target masks and dataset augmentation aren’t enough for enhancing the generalization of CNNs

Call for More Research on Benchmarking AI4Vis!

Evaluating ‘Graphical Perception’ with Multimodal LLMs

Rami Huu Nguyen* University of Massachusetts Boston
Kenichi Maeda† University of Massachusetts Boston
Mahsa Geshvadi‡ University of Massachusetts Boston
Daniel Haehn§ University of Massachusetts Boston

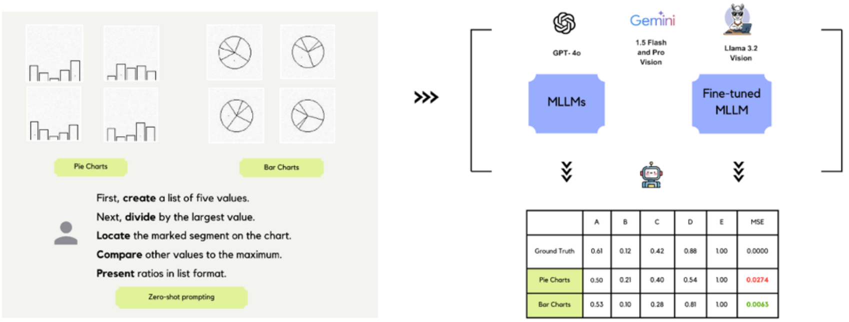


Figure 1: **Computing Cleveland and McGill’s Position-Angle Experiment using Multimodal Large Language Models**. We replicate the original experiment by asking MLLMs to interpret values in pie and bar charts using zero-shot prompting, where models follow instructions without prior examples. Results highlight that MLLMs predict values more accurately from bar charts (mean squared error (MSE) in green).

Abstract

Multimodal Large Language Models (MLLMs) have remarkably progressed in analyzing and understanding images. Despite these advancements, accurately regressing values in charts remains an underexplored area for MLLMs. For visualization, **how do MLLMs perform when applied to graphical perception tasks?** Our paper investigates this question by reproducing Cleveland and McGill’s seminal 1984 experiment and comparing it against human task performance. Our study primarily evaluates fine-tuned and pretrained models and zero-shot prompting to determine if they closely match human graphical perception. Our findings highlight that MLLMs outperform human task performance in some cases but not in others. We highlight the results of all experiments to foster an understanding of where MLLMs succeed and fail when applied to data visualization.

Index Terms: Multimodal Large Language Models, Graphical Perception, Machine Perception, Deep Learning

1 Introduction

Nowadays, data visualization has become increasingly important in our lives [21, 11]. There has been a rising research focus on computational techniques for studying charts, and graphs. [2, 21, 26], which are applied in several applications, including data extraction, classification, visual Q&A (e.g., “computer, which section is greater?”), and design evaluation or synthesis. MLLMs have made significant progress in analyzing and understanding images [9, 22, 1, 20, 14, 27, 8]. Although MLLMs perform well in understanding charts, they struggle in generalization and face difficulties accurately answering chart-related questions [1, 11]. This requires the MLLMs to understand both language and information derived in charts and apply reasoning skills to provide correct answers [12, 10]. Most current MLLMs are pre-trained vision and knowledge, which means those models are trained before with general knowledge, and they might struggle with new application [13, 15, 17, 16], which potentially lead to incorrect visual understanding. Understanding images (computer vision) poses unique challenges as compared to understanding language [16] [19]. Language often relies on structured syntax and grammar, while chart data depends on spatial relationships, patterns, and context [18]. Hence, analyzing chart data might be more challenging for the MLLMs. What’s more, the limitation of MLLMs also persists: MLLMs find it difficult to recognize small objects or tiny details in pictures [24, 22, 23]. Additionally, MLLMs currently have difficulty pinpointing the important details in the images that are unclear or absent in the images [25]. Also, humans use senses such as sight and language to understand the world and recognize new objects based on their knowledge.[17, 27]. Zero-shot prompting follows similar principles as human abilities, with its main purpose being to improve MLLMs using the zero-shot prompts to make them perform better without the need for additional training. Cleveland and McGill introduced the concept of graphical perception, explaining how humans visually interpret information from graphs [3, 4]. Cleveland and McGill defined elementary perceptual tasks as mental-visual processes and ranked how complex those tasks are

*e-mail: rami@mpsych.org
†e-mail: kenichi.maeda001@umb.edu
‡e-mail: mahsa.geshvadi001@umb.edu
§e-mail: daniel.haehn@umb.edu

The Perils of Chart Deception: How Misleading Visualizations Affect Vision-Language Models

Ridwan Mahbub* York University
Mohammed Saidul Islam* York University
Md Tahmid Rahman Laskar* York University
Mizanur Rahman* York University
Mir Tafseer Nayeem† University of Alberta
Enamul Hoque* York University

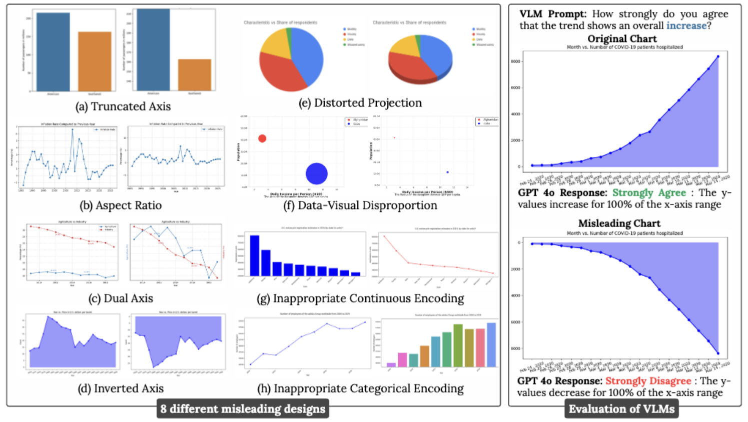


Figure 1: Overview of the eight misleading chart designs studied: (a) Truncated Axis, (b) Aspect Ratio Distortion, (c) Dual Axis, (d) Inverted Axis, (e) Distorted Projection, (f) Data-Visual Disproportion, (g) Inappropriate Continuous Encoding, and (h) Inappropriate Categorical Encoding. Each pair shows the original (left) and misleading (right) chart. The rightmost section displays GPT-4o outputs for both; in the misleading case, the inverted y-axis leads GPT-4o to wrongly infer a decreasing trend.

Abstract

Information visualizations are powerful tools that help users quickly identify patterns, trends, and outliers, facilitating informed decision-making. However, when visualizations incorporate deceptive design elements—such as truncated or inverted axes, unjustified 3D effects, or violations of best practices—they can mislead viewers and distort understanding, spreading misinformation. While some deceptive tactics are obvious, others subtly manipulate perception while maintaining a façade of legitimacy. As Vision-Language Models (VLMs) are increasingly used to interpret visualizations, especially by non-expert users, it is critical to understand how susceptible these models are to deceptive visual designs. In this study, we conduct an in-depth evaluation of VLMs’ ability to interpret misleading visualizations. By analyzing over 16,000 responses from ten different models across eight distinct types of misleading chart designs, we demonstrate that most VLMs are deceived by them. This leads to altered interpretations of charts, despite the underlying data remaining the same. Our findings highlight the need for robust safeguards in VLMs against visual misinformation.

Index Terms: Misleading Visualizations, Large Language Models, Vision Language Models, Taxonomy, Evaluation

1 Introduction

Visualizations are powerful tools for transforming complex data into accessible narratives, helping diverse audiences uncover patterns, trends, and anomalies. Across domains, from journalism and public policy to healthcare, finance, and social media, visualizations drive data storytelling and inform high-stakes decisions [35]. However, this communicative power can be a double-edged sword. Subtle manipulations such as truncated axes, skewed aspect ratios, and gratuitous 3D embellishments can produce misleading visualizations that distort perception without altering the underlying data [17, 6, 25]. These practices don’t fabricate facts; instead, they subtly alter their visual representation to amplify or minimize perceived differences, influencing narratives and decisions [17, 6, 25]. Figure 1 illustrates *eight* such design tactics studied in this work. For instance, in Figure 1(b), an inflation trend is made to look significantly flatter by changing the aspect ratio—an alteration that may go unnoticed but can fundamentally mislead the viewer. Such

*e-mail: {rmahbub, saidulis, tahmid20, mizanurr, enamulh}@yorku.ca
†e-mail: mnayeem@ualberta.ca

Visual Informatics
Available online 29 September 2025, 100285
In Press, Journal Pre-proof What's this?

Research article

How well will LLMs perform for graph layout tasks?

Yilun Fan [a c], Xianglei Lyu [a], Lei Wang [b], Ying Zhao [a], Fangfang Zhou [a], Yong Wang [c]

Show more

Add to Mendeley Share Cite

https://doi.org/10.1016/j.visinf.2025.100285 Get rights and content

Highlights

- We propose a systematic evaluation framework for examining LLMs’ capabilities in graph layout tasks, which covers graph data understanding, layout generation, and layout evaluation.
- We conduct a large-scale comparative study across different graph types, sizes, formats, prompting modes, and layout constraints, using three mainstream LLMs.
- We provide an in-depth analysis of LLMs’ strengths and limitations for graph layout tasks, and report the major findings in terms of their potential and capability boundaries for graph layouts.

Latest AI Models

Different Visualization Tasks

VIS2025

Generalization of CNNs on Relational Reasoning with Bar Charts

Zhenxing Cui*, Lu Chen*, Yunhai Wang, Daniel Haehn, **Yong Wang**, Hanspeter Pfister

Project Page: https://www.yunhaiwang.net/tvcg2024/CNN_Generalization
Code & Data: https://github.com/Ideas-Laboratory/Graphical-Perception

VIS2025