

Assessing Graphical Perception of Image Embedding Models using Channel Effectiveness

Soohyun Lee^{*}
Seoul National University

Minsuk Chang[†]
Seoul National University

Seokhyeon Park[‡]
Seoul National University

Jinwook Seo[§]
Seoul National University

ABSTRACT

Recent advancements in vision models have greatly improved their ability to handle complex chart understanding tasks, like chart captioning and question answering. However, it remains challenging to assess how these models process charts. Existing benchmarks only roughly evaluate model performance without evaluating the underlying mechanisms, such as how models extract image embeddings. This limits our understanding of the model’s ability to perceive fundamental graphical components. To address this, we introduce a novel evaluation framework to assess the graphical perception of image embedding models. For chart comprehension, we examine two main aspects of channel effectiveness: accuracy and discriminability of various visual channels. Channel accuracy is assessed through the linearity of embeddings, measuring how well the perceived magnitude aligns with the size of the stimulus. Discriminability is evaluated based on the distances between embeddings, indicating their distinctness. Our experiments with the CLIP model show that it perceives channel accuracy differently from humans and shows unique discriminability in channels like *length*, *tilt*, and *curvature*. We aim to develop this work into a broader benchmark for reliable visual encoders, enhancing models for precise chart comprehension and human-like perception in future applications.

Index Terms: Graphical perception, channel effectiveness, image embeddings, clip

1 INTRODUCTION

Nowadays, emerging vision models are strongly influencing the domain of visualization, especially in handling charts. Image encoders are employed to automatically classify chart images [32, 3], explain charts [4, 34], or answer chart-based questions [16, 28]. This is due to the recent advancements in vision models’ ability to process visual data and perform diverse tasks (e.g., saliency prediction [24], image captioning [31, 36], or visual question answering [2]), often surpassing human-level performance [13, 17].

However, most existing benchmarks for chart understanding models focus on high-level tasks, such as task performance in question answering [28, 40] or image captioning [34] scenarios. These benchmarks can evaluate the model’s overall performance and utility but are too coarse to address how they perceive and interpret the fundamental graphical elements in charts at a perceptual level.

To address this gap, we introduce a novel evaluation framework for image embedding models based on the concept of ‘channel effectiveness,’ which considers two main aspects: accuracy (section 3, section 4) and discriminability (section 4). Our framework can measure how precisely vision models can interpret and discriminate the magnitude channel typically used in charts [27, 29]: *length*, *tilt*, *area*, *color luminance*, *color saturation*, and *curvature*.

First, we suggest using the linearity of image embeddings as a proxy for each channel’s accuracy while their magnitude increases. According to Steven’s power law, channel accuracy improves as the perceived magnitude increases linearly with the given stimuli. Then, we broadened our investigation of whether the order of measured linearity could be generalized across all combinations of controlled variables. Secondly, we suggest analyzing the distances between consecutive embeddings to evaluate each channel’s discriminability. We calculated the peaks from the smoothed distance graph to ascertain the number of distinguishable groups and how sensitive the model reacts to the magnitude of the channel.

We also present that our evaluation framework can measure the low-level performance of the model depending on the model’s goal:

- 1. Tasks requiring precise quantitative analysis** (e.g., determining exact values from a bar chart). In such cases, models should give precise answers from the graphical elements. Therefore, they must achieve higher channel accuracy and maintain low discriminability when the magnitude increases.
- 2. Tasks where models need to process charts as humans do** (e.g., interpreting trends from a line graph [6] or mimicking user studies [15]). In this scenario, the accuracy should align with the known perceptual effectiveness of human vision [20, 8] (should follow human perceptual ranking), and discriminability should mirror human ability [29, 35].

We applied our framework on CLIP [31], one of the state-of-the-art image embedding models pre-trained on a large-scale dataset of natural images. The result reveals that CLIP’s order of channel accuracy differs from human perception and each accuracy is much lower than being ideally linear. Furthermore, CLIP exhibits unique discriminability patterns that seem to follow human perception on certain channels, such as *length*, *tilt*, and *curvature*. We also found that CLIP’s perception conforms to Weber’s law [11], indicating that perceived changes in stimuli are proportionate to the magnitude of the initial stimuli. Comprehensively, we observed a tradeoff between accuracy and discriminability, where accuracy can be lowered when there exists a certain amount of discriminability.

We present our framework for channel effectiveness as a foundational effort in establishing low-level benchmarks for chart comprehension. Furthermore, we suggest the visualization community explore additional low-level benchmarks, such as pre-attentive processing [19] or just-noticeable difference (JND) [11]. Our future initiatives include collecting crowdsourced data to validate our findings from discriminability and comparing them with results from various image embedding models to confirm the robustness and applicability of our benchmark.

2 RELATED WORK

2.1 Graphical Perception

Cleveland and McGill [8] tried to understand encodings in visualization through graphical perception. They measured humans’ graphical perception by defining 10 elementary perceptual tasks (e.g., position, length, angle, area, and volume), collectively known as channels, that people use to extract quantitative information from

^{*}e-mail: shlee@hcil.snu.ac.kr

[†]e-mail: jangsus1@snu.ac.kr

[‡]e-mail: shpark@hcil.snu.ac.kr

[§]e-mail: jseo@snu.ac.kr

graphs. They also performed pairwise experiments, such as comparing bar charts and pie charts to compare the difficulty between position and angle, and then ranked these channel effectiveness based on the accuracy of human perception.

Various research extended this in terms of participant scale [20], data type [22, 21], or task complexity [5, 39], leading to a broader and more solid understanding of graphical perception. Furthermore, the emergence of CNN models questioned researchers on whether the same findings also apply to the trained models [14].

Traditional methods primarily evaluate graphical perception by assessing channel effectiveness in terms of accuracy with human subjects or through models trained with predetermined answers. However, these approaches cannot be easily applied to computer vision models. Also, they typically miss examining other critical aspects of channel effectiveness [29] (e.g., discriminability [23], separability, popout, and grouping), which are crucial for understanding how visual information is perceived and processed.

Therefore, our paper introduces a method for understanding how unsupervised image training models perceive channels by analyzing the raw outputs of image embedding models. Beyond accuracy, our investigation includes discriminability, which refers to the ability to differentiate between similar visual elements, illustrating a broader perspective on the components of channel effectiveness.

2.2 Image Embeddings

Computer images are structured pixels of light and color intensities, which are not directly related to their intrinsic meaning. Therefore, various image encoding models were suggested to transform images into embeddings, a fixed-length vector representing their semantics [1, 28, 31, 25]. These image embedding models are often used in other models' backbone to classify, describe, interpret, or perform multimodal tasks [26, 9]. Also, various benchmarks were suggested to evaluate such models through QA tasks [40, 34]. However, we found that the image embedding itself has neither been investigated nor evaluated. Therefore, we present a methodology to measure graphical perception within these image embeddings and experiment with CLIP [31], a general state-of-the-art image embedding model used in various domains (medicine [37], fashion [7], or even user interfaces [30]).

3 LINEARITY AS CHANNEL ACCURACY

In pursuit of understanding how image embedding models like CLIP capture variations on different visual channels, we designed experiments to assess their sensitivity to changes in six different channels: *Length*, *Tilt*, *Area*, *Color Luminance*, *Color Saturation*, and *Curvature*. Previous studies have shown that the accuracy of human perception of these visual channels differs by order: length is most accurately perceived, followed by tilt, area, color luminance, color saturation, and curvature [8, 20]. We examined whether an image embedding model can produce embeddings with a strict order of accuracy.

3.1 Linearity with Fixed Control Variables

3.1.1 Experiment Design

In our experiment, we generated an image dataset of simple shapes, following the prior work [20]. Our goal was to eliminate any unintended bias from the background, ensuring a focus purely on the graphical perception of the elements. We first created images of a line segment with each channel applied with a certain magnitude on a white background. For each channel, we encoded the range of values in Figure 1 over 1000 steps. While testing on one channel, other channels are fixed to the controlled (default) value in Figure 1 (Length: 50%, Tilt: 0°, Area: None, Curvature: 0°, Luminance: 50%, Saturation: 100%). For example, in the luminance channel, the line segments with a Length of 50% and no Tilt or Curvature were rendered with varying degrees of brightness.

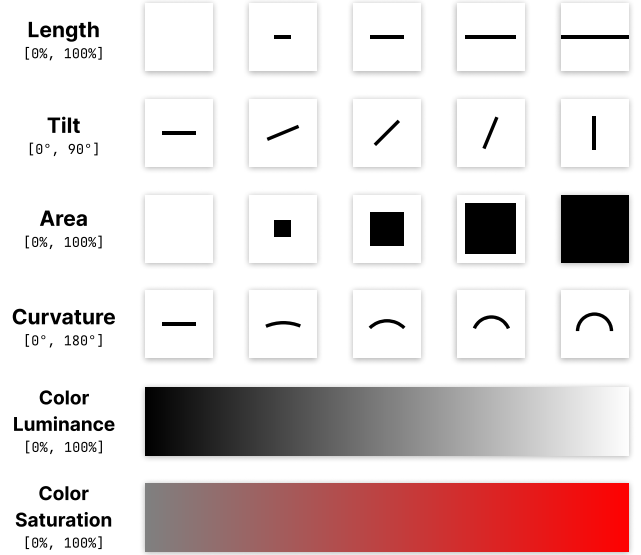


Figure 1: Examples of our variations for each channel. **Length** and **Area** of 100% means the line or square fills the screen. **Tilt** is adjusted from 0° to 90°. **Curvature** starts from a straight line to a semi-circular arc. Color Hue is fixed to 0 (red) when the **Luminance** and **Saturation** increase from 0% to 100%.

For each image, we extracted embeddings using three CLIP models [31] with different visual architecture; one ResNet [18] model (RN50x64) and two Vision Transformer [10] models (ViT-B/32 and ViT-L/14@336px). We then analyzed linearity using principal component analysis (PCA) [38], observing how well the first principal component could represent the distribution of embeddings. This allows us to measure the linearity of each channel's embedding space quantitatively. We also claim that linearity can be a good proxy for measuring channel accuracy, where the single-scale change in its stimuli directly appears in the embedding space. Also aligned with that, an accurate channel to humans means that human perception is proportional to stimuli [33].

3.1.2 Result

The linearity of each channel for each model is plotted in Figure 2. On the y-axis of the figure, the channels are arranged in the order that humans perceive more accurately, starting from the top. We can find that the order does not align with human perception for all models. Notably, model ViT-B/32 shows significantly lower linearity in channel tilt than other models. Additionally, for all models, *color luminance* shows the lowest linearity compared to other channels. This suggests that important data encoded as *color luminance* can potentially cause the CLIP model to recognize or misinterpret its precise value barely.

3.2 Linearity with Every Controlled Variables

We also investigated whether the findings can be generalized into circumstances where other channels are not fixed. This experiment reflects a more complex and realistic scenario similar to natural visual environments where data tends to be encoded in multiple channels simultaneously.

3.2.1 Experiment Design

We explored the linearity of channels by examining all possible combinations of channel variations, finding out whether a general order of channel effectiveness exists. In subsection 3.1, we experimented with varying channel magnitude with 1000 steps where

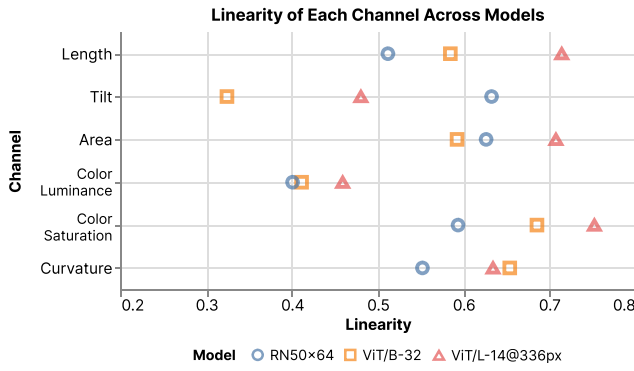


Figure 2: Linearity of various visual channels across different CLIP models. The Y-axis is the channel arranged in the order that humans perceive more accurately. Each channel’s linearity varies between models, which does not closely align with human perceptual accuracy. Also, the ViT-L/14@336px model usually shows better accuracy compared to other models.

other channels are fixed as controlled values in Figure 1. However, in this extended experiment, we reduced the number of steps to 20 while testing all combinations of other channels, resulting in 20^4 image variations per channel. We then measured every channel combination’s linearity and observed whether a general order of channel accuracy exists.

3.2.2 Result

The calculated linearity scores are plotted in Figure 3. We can easily observe a general order of linearity among the channels (*Color Saturation* > *Curvature* > *Length* > *Color Luminance* = *Tilt*), which is also similar to the result from subsection 3.1. Furthermore, we conclude that the CLIP model shows a significant difference in channel accuracy compared to human perception. This disproves the assumption that vision models would process the data similarly to humans.

3.3 Discussion

The previous two experiments reveal that CLIP’s channels’ effectiveness ranking differs from that of humans, and some of the channels show low linearity. Different channel rankings indicate that the model perceives images differently from humans, leading to a risk of inaccurate results when using CLIP embeddings to mimic human perceptions. Also, the low linearity channels can cause inaccurate answers to be produced when performing quantitative QA based on graphical elements through CLIP embedding. We suggest these results should be considered when using the image embedding model.

4 DISTANCE AS CHANNEL DISCRIMINABILITY

In this section, we explore the concept of discriminability (one measure of the channel’s effectiveness) within the context of image embedding models. Discriminability refers to the capacity to perceive distinct steps or changes within visual elements of an image. Conversely, the existence of such discriminability suggests that accuracy may be low. For instance, it has been observed that distinguishing more than six hues or more than six symbol shapes within a visual array can be challenging, suggesting a perceptual threshold for discriminability [27].

This threshold indicates the minimum difference required between two objects to be considered distinct. To assess this, we measured the Euclidean distance between image embeddings, as these distances can be interpreted as the model’s ability to differentiate between two images.

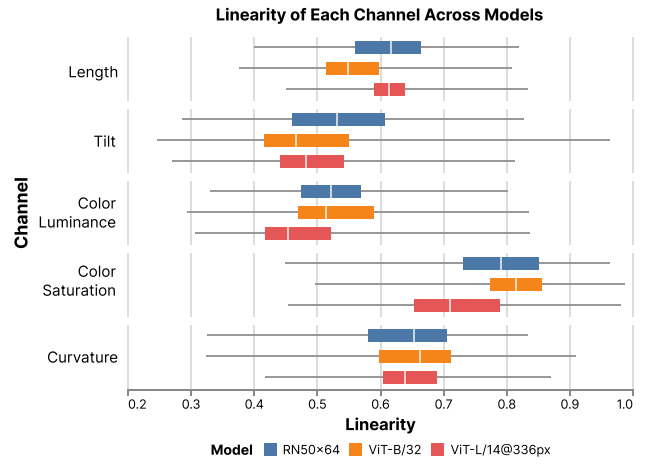


Figure 3: The box plot illustrates the linearity scores for each channel under every combination of controlled variables, showcasing general patterns and deviations. Since area cannot be applied together with length or curvature, we have generated combinations without area. Based on the plot, all models have similar overall rankings for the channels they perceive (*Color saturation* > *Curvature* > *Length* > *Tilt* = *Color luminance*). The whisker of this boxplot represents the min and max of the full data.

We measured the Euclidean distance between embeddings, calculated with the best-performing model (ViT-L/14@336px), of each consecutive pair in a series of 1000 images, each created by incrementally changing the value of one channel as described in subsection 3.1. The results of these measurements are smoothed with a Gaussian filter [12] ($\sigma = \lfloor \sqrt{1000} \rfloor = 32$) and plotted in Figure 4. The smoothed plot helps us observe the boundary between the groups the model perceives as different, thereby showing how much change in one channel is necessary before the model perceives the two images as different. We expect the number of distinct peaks can be a proxy for extracting separable groups throughout each channel.

4.1 Result & Discussion

The result from this experiment roughly explains how the model discriminates changes across different channels.

Length: First, when analyzing the discriminability of the length channel, we can see that when the length is short, the distance between adjacent embeddings becomes relatively high. Supported by Weber’s law [11], the model captures subtle changes well when the length is short, similar to humans. Interestingly, the distance graph for Length shows three or four distinct peaks with valleys between, indicating that the model recognizes length in four separate stages.

Tilt: The distances are noticeably higher for tilt at angles of 0° , 45° , and 90° . This pattern shows that the model primarily differentiates images based on whether the tilt is less than, equal to, or greater than 45° . In other words, we can assume that the model uses 45° as an important threshold for processing images.

Area: Upon analyzing the area, several small peaks were identified throughout the range. Unlike length, we cannot easily distinguish or split into distinct groups.

Color Luminance: We can observe a small peak near 0% and an extreme peak at 100%. This indicates that the model is particularly sensitive to changes in luminance at very low and very high values. It suggests that when it is very dark or very bright, it responds greatly to small changes in luminance, and in other areas, images are viewed relatively similarly.

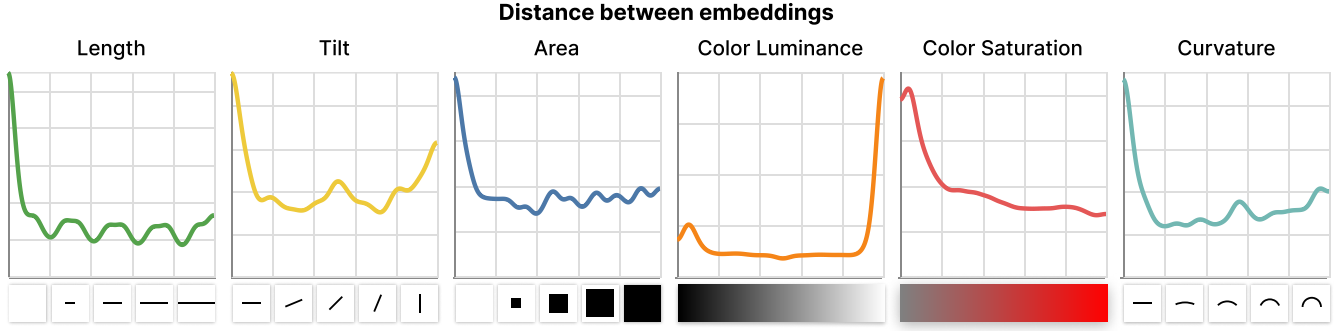


Figure 4: The smoothed plot of the Euclidean distances between image embeddings for incremental changes in each visual channel. Sample images below the chart are illustrations of stimuli variations in each channel. Peaks represent thresholds where the model perceives significant differences between images, indicating the discriminability of each channel. This visualization aids in identifying how many perceptual groups the model can distinguish in each channel.

Color Saturation: Distance for color saturation was high near the value of 0%. This peak indicates that the model is highly responsive to initial saturation increases in the low saturation state but becomes less sensitive once the change has been made.

Curvature: Curvature has very high distance values when approaching 0° , showing high sensitivity when there are almost no curves. Also, a peak around 90° appears to be a critical point similar to tilt, which becomes a perceptual boundary.

5 GENERAL DISCUSSION & FUTURE APPLICATION

We proposed an evaluation framework (section 3) that extensively investigated the channel effectiveness of the image embedding model and then applied it to the CLIP model. However, we summarize our limitations and suggest our corresponding future work.

5.1 Reliable Visual Encoders for Charts

The analysis from section 3 shows that generally trained vision models may interpret the encoded data unreliably, where linearity for channels except *color saturation* stays around 0.6. This can be problematic in chart question-answering models where accurate interpretation of visual cues is essential. Accurate interpretation requires an accurate perception of channels. Thus, accuracy for all channels should be linear (linearity score close to 1). Conversely, checking whether the model’s discriminability matches human perception can be crucial in chart captioning models. Chart captioning requires a holistic understanding of existing visual elements, where matching its discriminability with humans is essential. Therefore, in this case, the order and intensity should be similar to that of humans besides having higher channel accuracy. Therefore, we suggest that chart question-answering models should be trained to have higher channel accuracy, while chart captioning models should have similar accuracy as humans.

5.2 Ambiguity in Peak Analysis for discriminability

In our analysis of discriminability in section 4, we conducted peak analysis to investigate the model’s effectiveness in differentiating channels. As discussed in the previous section (subsection 5.1), two-directional goals emerge: reporting high accuracy and matching human discriminability. To interpret charts accurately, the graph shapes for each channel in Figure 3 should remain constant, as consistent perception under consistent stimuli is required. In other words, no discriminability should be found. Conversely, if the goal is to align with human perception, the graph shapes should mirror the discriminability observed in humans.

However, several challenges arise with this approach in terms of reliability. First, hyperparameters for the Gaussian filter should be chosen carefully. Excessive smoothing and noise reduction might

lead to missing peaks in the original graph, potentially overlooking significant details. Additionally, when analyzing peaks in the graph, the threshold for distance between embeddings was set arbitrarily, leading to a subjective inspection of peaks. For instance, in the area under consideration in Figure 4, multiple peaks are evident and appear regular, yet the interpretation of each peak remains unclear.

Given that no studies have deeply investigated how similar these results are to human perception, our future work could involve detailed human studies. We believe that more thorough comparisons could be made to evaluate how closely the graph shapes resemble those humans perceive rather than focusing solely on peak analysis.

5.3 General Benchmark for Channel Effectiveness

Our current evaluation of channel effectiveness primarily focuses on the accuracy and discriminability of embeddings across six different magnitude channels. However, a comprehensive assessment of channel effectiveness should include additional metrics such as separability, popout, and grouping, which are crucial for understanding pre-attentive processing and just-noticeable difference (JND) in graphical perception.

To address these gaps, we propose developing a framework that measures model performance at a low level across these various metrics. Our framework can be extended as a standardized benchmark that evaluates these fundamental aspects of graphical perception and provides a platform for comparing different visual encoders. This benchmark would allow for a deeper understanding of how various models interpret and process graphical data, paving the way for advancements in chart comprehension technologies.

6 CONCLUSION

Our work introduces a novel framework for evaluating the graphical perception of image embedding models, focusing on the concept of channel effectiveness. Our comprehensive experiments using the CLIP model have revealed significant disparities in how vision models and humans perceive and interpret visual channels. We observed that the accuracy and discriminability of these channels differ markedly between the CLIP model and human perception, suggesting the careful use of image embedding models on perception-related tasks. Our findings highlight the potential for improving model reliability in tasks requiring human-like perception and precision, such as chart question answering and captioning. As a future work, we suggest extending our benchmark to assess other low-level channel effectiveness, enhancing the robustness and reliability of visual encoders across diverse applications. This approach not only promises reliable models in terms of graphical perception but also paves the way for future innovations in graphical data interpretation and machine learning in visualization.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2023R1A2C200520911)

REFERENCES

- [1] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, vol. 35, pp. 23716–23736, 2022. 2
- [2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015. 1
- [3] F. Bajić and J. Job. Chart classification using siamese cnn. *Journal of Imaging*, 7(11), 2021. 1
- [4] A. Balaji, T. Ramanathan, and V. Sonathi. Chart-text: A fully automated chart image descriptor. *arXiv preprint arXiv:1812.10636*, 2018. 1
- [5] C. X. Bearfield, C. Stokes, A. Lovett, and S. Franconeri. What does the chart say? grouping cues guide viewer comparisons and conclusions in bar charts. *IEEE Transactions on Visualization and Computer Graphics*, 30(8):5097–5110, 2024. 2
- [6] C. X. Bearfield, L. van Weelden, A. Waytz, and S. Franconeri. Same data, diverging perspectives: The power of visualizations to elicit competing interpretations. *IEEE Transactions on Visualization and Computer Graphics*, 30(6):2995–3007, 2024. 1
- [7] P. J. Chia, G. Attanasio, F. Bianchi, S. Terragni, A. R. Magalhães, D. Goncalves, C. Greco, and J. Tagliabue. Contrastive language and vision learning of general fashion concepts. *Scientific Reports*, 12(1):18958, 2022. 2
- [8] 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, 1984. 1, 2
- [9] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, and S. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024. 2
- [10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2
- [11] G. T. Fechner. *Elements of Psychophysics*, vol. 1. Holt, Rinehart and Winston, United States of America, 1966. Original work published 1860. 1, 3
- [12] C. F. Gauss. *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*, vol. 7. FA Perthes, 1877. 3
- [13] R. Geirhos, D. H. Janssen, H. H. Schütt, J. Rauber, M. Bethge, and F. A. Wichmann. Comparing deep neural networks against humans: object recognition when the signal gets weaker. *arXiv preprint arXiv:1706.06969*, 2017. 1
- [14] D. Haehn, J. Tompkin, and H. Pfister. Evaluating ‘graphical perception’ with cnns. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):641–650, 2019. 2
- [15] P. Hämäläinen, M. Tavast, and A. Kunnari. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2023. 1
- [16] Y. Han, C. Zhang, X. Chen, X. Yang, Z. Wang, G. Yu, B. Fu, and H. Zhang. Chartllama: A multimodal llm for chart understanding and generation. *arXiv preprint arXiv:2311.16483*, 2023. 1
- [17] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015. 1
- [18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. 2
- [19] C. Healey and J. Enns. Attention and visual memory in visualization and computer graphics. *IEEE Transactions on Visualization and Computer Graphics*, 18(7):1170–1188, 2012. 1
- [20] J. Heer and M. Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 203–212, 2010. 1, 2
- [21] J. Heer, N. Kong, and M. Agrawala. Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 1303–1312, 2009. 2
- [22] W. Javed, B. McDonnell, and N. Elmqvist. Graphical perception of multiple time series. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):927–934, 2010. 2
- [23] H. Jeon, G. J. Quadri, H. Lee, P. Rosen, D. A. Szafir, and J. Seo. Clams: a cluster ambiguity measure for estimating perceptual variability in visual clustering. *IEEE Transactions on Visualization and Computer Graphics*, 2023. 2
- [24] A. Kroner, M. Senden, K. Driessens, and R. Goebel. Contextual encoder-decoder network for visual saliency prediction. *Neural Networks*, 129:261–270, 2020. 1
- [25] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu. Image-text embedding learning via visual and textual semantic reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):641–656, 2023. 2
- [26] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2
- [27] J. Mackinlay. Automating the design of graphical presentations of relational information. *ACM Trans. Graph.*, 5(2):110–141, apr 1986. 1, 3
- [28] A. Masry, D. Long, J. Q. Tan, S. Joty, and E. Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279. Association for Computational Linguistics, Dublin, Ireland, May 2022. 1, 2
- [29] T. Munzner. *Visualization analysis and design*. CRC press, 2014. 1, 2
- [30] S. Park, W. Kim, Y.-H. Kim, and J. Seo. Computational approaches for app-to-app retrieval and design consistency check. *arXiv e-prints*, pp. arXiv-2309, 2023. 2
- [31] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021. 1, 2
- [32] M. Savva, N. Kong, A. Chhajta, L. Fei-Fei, M. Agrawala, and J. Heer. Revision: Automated classification, analysis and redesign of chart images. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pp. 393–402, 2011. 1
- [33] S. S. Stevens. On the psychophysical law. *Psychological review*, 64(3):153, 1957. 2
- [34] B. J. Tang, A. Boggust, and A. Satyanarayan. Vistext: A benchmark for semantically rich chart captioning. *arXiv preprint arXiv:2307.05356*, 2023. 1, 2
- [35] R. Veras and C. Collins. Discriminability tests for visualization effectiveness and scalability. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):749–758, 2020. 1
- [36] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2015. 1
- [37] Z. Wang, Z. Wu, D. Agarwal, and J. Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022. 2
- [38] S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987. 2
- [39] C. Xiong, C. Stokes, Y.-S. Kim, and S. Franconeri. Seeing what you believe or believing what you see? belief biases correlation estimation. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):493–503, 2023. 2
- [40] Z. Xu, S. Du, Y. Qi, C. Xu, C. Yuan, and J. Guo. Chartbench: A benchmark for complex visual reasoning in charts. *arXiv preprint arXiv:2312.15915*, 2023. 1, 2