CHARTCAP: Mitigating Hallucination of Dense Chart Captioning

Junyoung Lim Jaewoo Ahn Gunhee Kim
Seoul National University
{junyoung.lim, jaewoo.ahn}@vision.snu.ac.kr, gunhee@snu.ac.kr
https://junyoung-00.github.io/ChartCap/

Abstract

Generating accurate, informative, and hallucination-free captions for charts remains challenging for vision language models, primarily due to the lack of large-scale, high-quality datasets of real-world charts. However, existing real-world chart datasets suffer from the inclusion of extraneous information that cannot be inferred from the chart and failure to sufficiently capture structural elements and key insights. Therefore, we introduce ChartCap, a large-scale dataset of 565K real-world chart images paired with type-specific, dense captions that exclude extraneous information and highlight both structural elements and key insights in detail. To build ChartCap, we design a four-stage pipeline that generates captions using only the discernible data from the chart and employ a cycle consistency-based human verification, which accelerates quality control without sacrificing accuracy. Additionally, we propose a novel metric, the Visual Consistency Score, which evaluates caption quality by measuring the similarity between the chart regenerated from a caption and the original chart, independent of reference captions. Extensive experiments confirms that models fine-tuned on ChartCap consistently generate more accurate and informative captions with reduced hallucinations, surpassing both open-source and proprietary models and even human-annotated captions.

1. Introduction

Charts are powerful tools for visualizing data distributions, trends, and patterns across various domains such as science, economics, and sociology. By presenting complex information in a concise and intuitive manner [23, 43], charts help readers gain meaningful insights for decision-making process. However, charts involve complex spatial relationships among various elements such as axes, labels, and legends, which can be interpreted differently depending on the chart type. Consequently, it is challenging not only for humans but also for vision language models (VLMs) to interpret complex charts [7, 21, 52].

Chart captioning is one of the core tasks to assess VLMs' ability to understand charts. Its goal is to generate natural descriptions of a chart image [20]. An ideal caption should (1) avoid inaccuracies about the chart [2, 17, 32], and (2) include a structural description of the chart components (e.g., title or legends) as well as key insights such as major statistics (e.g., maximum or minimum values) and perceptual patterns (e.g., data trends) [33, 54]. However, existing real-world chart datasets [14, 20, 25, 30, 50] suffer from two major issues: (1) they contain extraneous information in captions that cannot be inferred from the image, and (2) they fail to sufficiently capture the essential information specific to each chart type.

First, the datasets contain *extraneous information* within their captions, mainly because charts are usually embedded in source documents and their original captions are simply paired with chart images without verification. Captions are often written based not only on the chart itself but also on the surrounding context. As a result, these captions include the information that cannot be inferred from the chart image alone (but from text in the document together), as shown in Figure 1. This poses an ill-posed problem for expecting the model to predict information absent from the chart, ultimately leading to hallucination.

Second, real-world chart datasets lack sufficient structural description and key insights in text; they often omit critical information that the chart image conveys (see Figure 1). It is partly because authors do not specify some details in the captions, assuming that human readers can easily infer them from the figure [7, 35]. Such information varies depending on the chart type; for example, scatter plots highlight clusters and distributions, while line charts emphasize temporal trends and changes [9]. Hence, a *type-specific caption schema*, which specifies how to interpret critical information for each chart type, is required to enable models to generate informative captions.

To address these issues, we propose CHARTCAP to improve VLMs' captioning performance while mitigating hallucinations. CHARTCAP is a large-scale dataset of real-world chart images, containing 565K chart-caption pairs

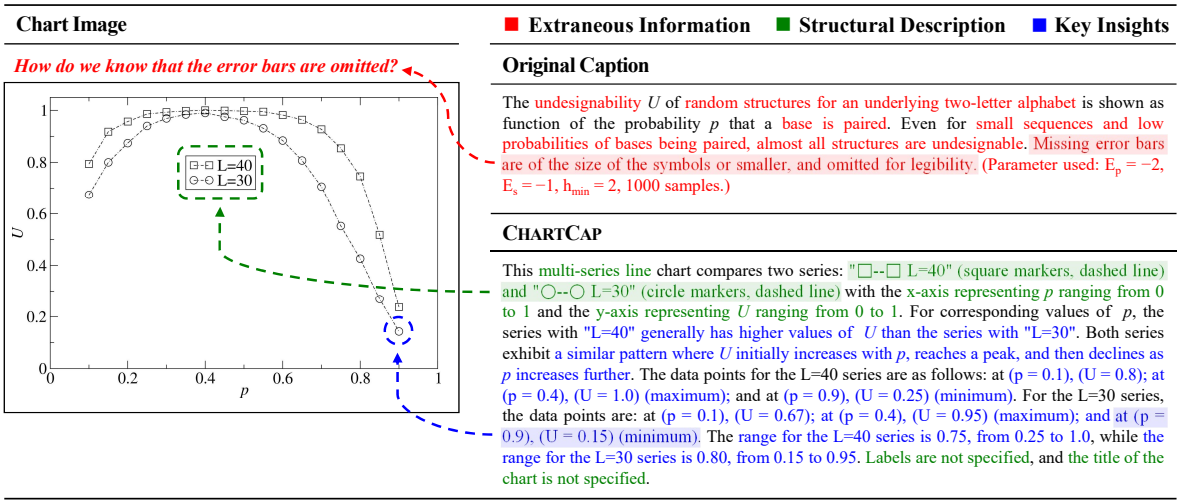

Figure 1. Comparison of the original caption and our CHARTCAP caption. The original caption includes extraneous information (in red), such as additional contextual details (e.g., missing error bars) and references to parameters (E_p, E_s, h_{min}), which cannot be inferred from the chart image. In contrast, CHARTCAP caption follows the line chart schema, relying on the information visible in the image. It includes a structural description (in green) and key insights (in blue). The chart is sourced from [6], collected by [25], and included in CHARTCAP.

that (1) exclude extraneous information not verifiable from the chart image and (2) provide structural description and key insights in a dense manner by following a type-specific caption schema. Drawing on research in the data visualization domain [24, 40], we define a caption schema that structures the core information to be included for each chart type. We then devise an automatic pipeline that generates captions using only the data inherent in each chart image, thereby minimizing the inclusion of extraneous information. Finally, we employ a cycle consistency-based [45, 64] human verification to ensure high-quality data pairs. Figure 1 illustrates a comparative example between CHARTCAP and the original captions.

Moreover, we propose a reference-free metric, the *Visual Consistency Score* (VCS), for evaluating chart captions. VCS exploits a recently powerful large language model (LLM) that translates a caption into Python code to generate a chart. Then, it compares the reconstructed chart to the ground-truth chart, overcoming the limitations of existing automated metrics, which struggle to capture the deep semantic quality of captions and are highly dependent on the quality of reference captions. In a head-to-head study, VCS demonstrated high agreement rate with human judgments, outperforming existing automatic metrics such as BERTScore [62].

Extensive experiments show that VLMs fine-tuned on CHARTCAP consistently generate more informative captions with fewer hallucinations, in terms of reference-based metrics, human evaluation, and the Visual Consistency Score, surpassing both open-source and proprietary models, including InternVL2.5 [8], Phi3.5-vision [1], ChartGemma [37], ChartInstruct-Llama2 [36] and Claude 3.5 Sonnet [4]. Moreover, the captions generated by CHARTCAP fine-tuned VLMs are more preferred to human-annotated captions from VisText [54] and Chart-to-Text [20] by human evaluators.

In summary, our main contributions are as follows:

1. We propose CHARTCAP, a large-scale 565K real-world chart caption dataset that is free from extraneous information and correctly conveys structural description and key insights via type-specific caption schema.
2. We propose the Visual Consistency Score (VCS), which evaluates the quality of chart captions by assessing deep semantic meaning without relying on reference captions.
3. Through extensive experiments, we show that VLMs trained on CHARTCAP generate high-quality, informative captions with fewer hallucinations.

2. Related Work

2.1. Datasets

For VLMs to generate accurate and informative chart captions, the training data should consist of real chart images, contain correct information, and capture the key insights conveyed by the chart. While synthetic datasets are scalable, models trained on synthetic data tend to exhibit limited robustness when applied to real-world charts [59, 61]. AutoChart [65], VisText [54], ChartLlama [12], and ChartSFT [39] are generated programmatically from raw data using visualization tools. However, ChartBench [59] shows that LLaVA [31] trained on synthetic ChartLlama [12] underperforms its pre-training baseline.

Dataset	Real-world charts	Free from extraneous info	Type-specific schema	Human annotation	Data pairs
ChartLlama [12]	✗	✓	✗	✗	11K
VisText [54]	✗	✓	✓	12K	12K
AutoChart [65]	✗	✓	✓	✗	24K
ChartGemma [37]	Δ	✓	✗	✗	62K
ChartSFT [39]	Δ	✗	✗	✗	1.0M
MMC [30]	Δ	✗	✗	✗	400K
ArxivCap [25]	✓	✗	✗	✗	3.9M*
ChartSumm [50]	✓	✗	✗	✗	84K
Chart-to-Text [20]	✓	✗	✗	8K	44K
SciCap [14]	✓	✗	✗	✗	134K
CHARTCAP	✓	✓	✓	**56K**	**565K**

Table 1. Comparison of CHARTCAP with public chart captioning datasets. Datasets marked with Δ include both real-world and synthetic charts. The asterisk (*) for ArxivCap indicates that it comprises both data-driven charts and non-data-driven ones such as conceptual diagrams or scientific illustrations. The Human annotation column means the number of chart-caption pairs annotated or verified by human. CHARTCAP encompasses 565K real-world chart-captions with human verification applied on the test set.

In contrast, ChartInstruct [36] collects real charts from 157 websites. However, it remains inaccessible to the public due to legal constraints. ChartSumm [50] and Chart-to-Text [20] collect chart-caption pairs from Statista, Pew, and Knoema, but are relatively small and lack informativeness [35]. ChartGemma [37] leverages Gemini 1.5 Flash to regenerate captions from chart images via zero-shot prompting. MMC-Instruct [30], SciCap [14], and ArxivCap [25] use scientific papers on arXiv, resulting in larger datasets, but model-generated captions are reported to be highly hallucinated [25].

On the other hand, our CHARTCAP leverages the visual diversity of real-world charts while excluding extraneous information and utilizing caption schema, thereby enabling VLMs to acquire more robust chart comprehension skills. More systematic comparison is presented in Table 1.

2.2. Automatic Evaluation Metrics

Various automatic evaluation approaches have been popularly used to measure the quality of generated captions, including BLEU [44], ROUGE [27], METEOR [5], and BERTScore [62], to name a few.

Despite their widespread use, these automatic evaluation metrics share common limitations. First, they fail to capture deeper linguistic or semantic nuances, captions are measured by aligning words or short phrases — even if the generated text contains factual errors or incoherent logic. Second, they are highly dependent on the quality of reference captions. Even if a generated caption accurately describes an image, it may be unfairly penalized if the reference caption is inaccurate or overly concise. Fundamentally, the true *ground-truth* (GT) in image captioning is the image itself, yet existing automatic metrics do not directly compare captions to the visual content of the image.

CLIPScore [13] utilizes CLIP [49] to directly compute the semantic similarity between an image and its caption. However, it primarily measures high-level semantic alignment [26] and cannot handle long captions, limiting its reliability as a comprehensive evaluation metric for tasks that require precise and detailed descriptions.

To address these challenges, we introduce a metric, **Visual Consistency Score**, which evaluates a generated caption by reconstructing the chart and computing the similarity between the reconstructed chart and the GT chart.

2.3. Hallucinations in VLMs

Hallucination in VLMs refers to the instances where the model generates text that does not align with the visual content [51]. One of the primary causes of hallucination is the misalignment between vision and language modalities [57]. To address this, Ciem [16] and Jiang et al. [18] employ contrastive learning with carefully crafted question-answer pairs that push misaligned representations away from correct ones. Liu et al. [29] propose containing both positive and negative instructions to strengthen model robustness. Sun et al. [53] and RLHF-V [60] refine the training process using human feedback to reward factual outputs, while HA-DPO [63], FDPO [11], and CLIP-DPO [41] leverage preference optimization by ranking and filtering generated responses.

However, they address object-centric tasks, leaving chart domain relatively unexplored. CHARTCAP tightly couples textual and visual cues in chart interpretation, enabling models trained on it to exhibit fewer hallucinations in more data-driven, abstract scenarios.

3. The CHARTCAP Dataset

Building a large-scale chart dataset with informative captions presents several challenges. First, it requires a clear definition of what information should be included in each caption. Second, an automated procedure with an appropriate schema is needed to generate high-quality captions at minimum cost. Therefore, we define a type-specific caption schema and a caption generation pipeline with four phases. Additionally, we facilitate efficient human verification using cycle consistency [45, 64], comparing the original chart image against a reconstructed image, enabling effective quality control of the test set.

The Chart Corpora. To assemble real-world charts, we collect 3.1 million chart images from ArxivCap [25], ChartSumm-Knoema [50], ChartCheck [2], and ChartQA-train [34] as a pool of data.

3.1. Defining the Caption Schema

We define a **type-specific caption schema** that outlines the structural description and key insights for each of nine chart

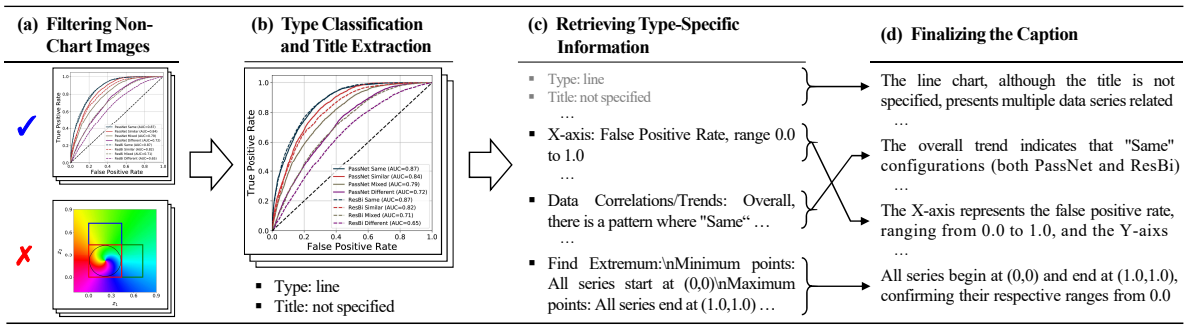

Figure 2. An example of the four-stage pipeline for our CHARTCAP: (a) filtering non-chart images, (b) classifying the chart type and extracting titles, (c) retrieving structural components and key insights, and (d) transforming the accumulated information into a coherent, sentence-level caption.

types, including *line, bar, pie, histogram, scatter, area, bubble, choropleth map, and treemap*, guided by the prior work in the field of data visualization. As a reference, *Visualization Analysis and Design* [40] provides a rigorous framework for designing visual representations.

To define the key insights for each chart type, we leverage the test blueprint from the *Visualization Literacy Assessment Test* (VLAT) [24], which identifies cognitive tasks for non-expert readers. Based on this framework, we minimize the ambiguity inherent in the criteria for crafting informative, high-quality captions [33, 46]. The complete schema is detailed in Appendix A.

3.2. Automated Dataset Generation Pipeline

We develop a four-stage pipeline, as depicted in Figure 2, to automate caption generation while balancing accuracy and computational cost via a combination of open-source and proprietary models. We report the accuracy of each stage by manual inspection on 100 randomly sampled instances.

Filtering Non-Chart Images. We first employ InternVL2.5-8B [8] to filter out non data-driven chart images (e.g., diagrams, schematics, illustrations). During this phase, multi-chart images are also removed, leaving us with 1.2 M images out of the initial set of 3.1 M. Manual inspection confirms 100% precision, implying that no false positives are retained.

Type Classification and Title Extraction. We use GPT-4o to obtain each chart's type and title. We filter out the charts that do not belong to the nine predefined types, leaving 577k chart images. If an explicit title is not detected, we assign the placeholder "not specified" to serve as a negative instruction, aiming to reduce hallucinations [29]. Manual evaluation shows an accuracy of 99%, with minor error due to ambiguous title placement within the chart.

Extracting Type-Specific Information. In accordance with our caption schema, we obtain structural components and key insights. We use GPT-4o for coarse-grained tasks such as identifying overall trends, while using Claude 3.5 Sonnet for more fine-grained tasks (e.g., locating exact max or min values). Preliminary experiments find that GPT-4o struggles to extract precise numerical values. Experiment details for this model selection are provided in Appendix C. If no information is extracted, it is labeled "not specified". Extracted information from the previous and current stages is accumulated in a semi-structured format as shown in Figure 2. Manual evaluation yields 94% accuracy, with some misinterpretation occurring in logarithmic-scale charts, scatter plots with no distinct correlations, and charts containing inset plots.

Finalizing the Caption. The semi-structured data is transformed into sentence-level captions. Given the relative simplicity of this stage, we use GPT-4o-mini to perform the transformation. Manual evaluation confirms that all transformations are accurate and preserve information.

3.3. Human Verification via Cycle Consistency

Despite the high performance of proprietary models, human verification remains indispensable for guaranteeing the quality of CHARTCAP. However, manually inspecting vast numbers of image-caption pairs is prohibitively time-consuming and expensive. To address this challenge, we introduce a cycle consistency-based human verification process, taking advantage of the millisecond-scale speed of human visual perception [3], as illustrated in Figure 3.

We generate Python code using Claude 3.5 Sonnet to recreate chart images from captions and then compare the reconstructed chart images with the originals. Applying human verification to 68K samples, we finalize a 56K test set. To validate the logical soundness of this verification process, we conduct qualitative and quantitative evaluation, detailed in Appendix D. Our findings are as follows.

1. Compared to direct image-caption comparison, our verification process is approximately 24 times faster while maintaining an F1 score of 95%.
2. Our verification process ensures both caption correctness and informativeness, making it well-suited for CHARTCAP's dense-captioning objectives.

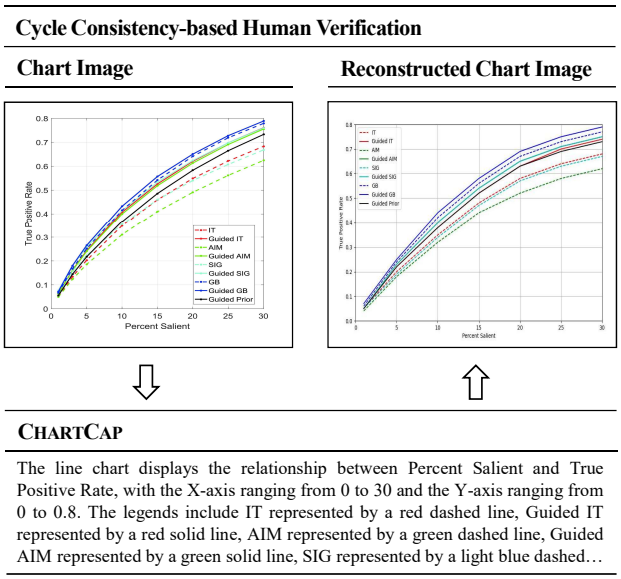

Figure 3. An illustration of the cycle consistency-based human verification for CHARTCAP. The original chart image (left) is compared with a reconstructed one (right) using a Python code from the caption (bottom). This process enables efficient human verification by assessing the accuracy and informativeness of the generated captions through visual consistency.

3.4. The Visual Consistency Score

How can we evaluate that a generated caption is faithful to its corresponding chart image? We believe that the best caption would correctly reproduce the chart, analogous to that a best generative model $P(x)$ is the one that can generate data x themselves.

This intuition leads us to propose a new captioning evaluation metric named the *Visual Consistency Score* (VCS), thanks to recent prominence of LLMs. Unlike natural images, charts have a unique characteristic: they can be deterministically generated from an intermediate modality—namely, code. Leveraging this property, we convert a given caption C_i into code G_i, subsequently producing a corresponding chart image $\hat{I}_i$. By measuring the similarity between this generated chart image $\hat{I}_i$ and the original chart image I_i, the VCS quantitatively evaluates the accuracy and informativeness of the caption C_i.

The VCS is computed by a two-stage procedure, code generation and image comparison. Given a caption C_i, an LLM is used to generate Matplotlib code G_i for recreating the chart. If G_i fails to execute, the code and runtime error message are supplied back to the LLM for debugging. This process is repeated until code execution succeeds, yielding a valid G_i, which is then executed to generate the chart image $\hat{I}_i$. The similarity between I_i and $\hat{I}_i$ is computed using a cosine similarity with a vision encoder. Finally, the VCS is the average similarity across all N samples:

$$\textbf{Visual Consistency Score} \;=\; \frac{1}{N}\sum_{i=1}^{N}\mathrm{Sim}(I_i,\hat{I}_i).$$

OCRScore. To evaluate how well textual elements are preserved, Optical Character Recognition (OCR) can be applied to both I_i and $\hat{I}_i$. Let $\mathcal{T}_i$ and $\hat{\mathcal{T}}_i$ be the sets of text strings extracted from I_i and $\hat{I}_i$, respectively. The OCRScore is as an F1 score based on precision (P) and recall (R):

$$P = \frac{\sum_{i=1}^{N}\left|\mathcal{T}_i \cap \hat{\mathcal{T}}_i\right|}{\sum_{i=1}^{N}\left|\hat{\mathcal{T}}_i\right|}, \qquad R = \frac{\sum_{i=1}^{N}\left|\mathcal{T}_i \cap \hat{\mathcal{T}}_i\right|}{\sum_{i=1}^{N}\left|\mathcal{T}_i\right|}$$

$$\textbf{OCRScore} = 2 \cdot \frac{P \times R}{P + R}.$$

Both VCS and OCRScore exhibit the highest agreement rates with human judgments among automated evaluation metrics such as BERTScore, demonstrating their practicality as reliable, scalable, and effective metrics for evaluating chart-caption quality. Detail of validation experiment is provided in Appendix E.

We use Claude 3.5 Sonnet for code generation, due to its superior performance in generating code [19]. For the vision encoder, we employ three variants of SigLIP2 [56], each at a resolution of 512, which achieves state-of-the-art performance across a variety of computer vision benchmarks [10, 28, 47, 55]. For OCR, we use PaddleOCR [42].

3.5. Dataset Analysis

Visual Consistency Score. We evaluate the Visual Consistency Score and OCRScore on 1K samples from each dataset. The results are presented in Table 2. CHARTCAP achieves the highest scores among all datasets, indicating that its captions are the most accurate to reconstruct the original chart information. The results indirectly reflect two key aspects: informativeness and the exclusion of extraneous information. Caption informativeness can be partially assessed by the average word count as CHARTCAP contains the longest captions with 231.1 words on average.

Human Evaluation. We conduct a head-to-head human evaluation by recruiting three annotators via Amazon Mechanical Turk (AMT), comparing 100 samples from CHARTCAP and ChartSumm [50] (the best dataset except ours). Each sample is evaluated based on informativeness, accuracy, fewer hallucinations, and overall preference. Details of human evaluation can be found in Appendix J. As illustrated in Figure 4, CHARTCAP consistently outperforms ChartSumm across all evaluated aspects, demonstrating higher overall quality recognized by human.

Dataset	Visual Consistency Score			OCRScore	Word Count
	Large	So400M	Base		
ArxivCap	0.7561	0.7421	0.7999	0.1781	43.7
ChartSumm	0.8940	0.9008	0.8898	0.2635	45.4
Chart-to-Text	0.6925	0.7089	0.7127	0.0951	62.2
SciCap	0.7861	0.8015	0.8457	0.1843	34.5
CHARTCAP	**0.8983**	**0.9089**	**0.9133**	**0.5424**	**231.1**

Table 2. Comparison of real-world chart datasets. The terms Large, So400M, and Base indicate three versions of the SigLIP2 encoder [56]: SigLIP2-{large, so400m, base}-512.

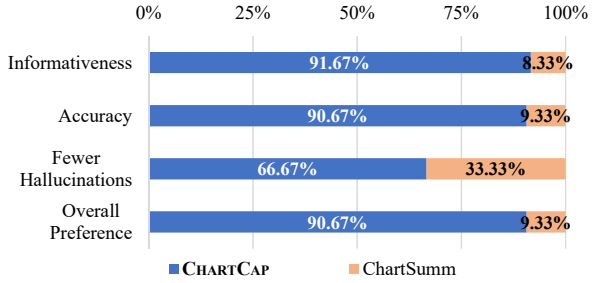

Figure 4. Results of the head-to-head human evaluation comparing CHARTCAP with ChartSumm [50].

4. Experiments

We demonstrate the effectiveness of our CHARTCAP dataset: first, we show that VLMs fined-tuned on CHARTCAP attain strong dense captioning performance in terms of reference-based metrics, human evaluation, and the Visual Consistency Score. Second, we present that CHARTCAP-trained captioning models show compelling zero-shot captioning on two human-annotated benchmarks, VisText [54] and Chart-to-Text [20].

Base Models. We experiment with open-source, chart expert, and proprietary captioning models. For open-source models, we use InternVL2.5-78B [8], InternVL2.5-38B, InternVL2.5-26B, InternVL2.5-8B, and Phi3.5-vision-4B [1]. For chart expert models, we employ ChartGemma-2B [37] and ChartInstruct-Llama2-7B [36]. For proprietary models, we use the Claude 3.5 Sonnet [4], which not only achieves the best performance in our dataset but also reports the state-of-the-art performance on ChartQA [34] and CharXiv [58]. Additional baselines are provided in Appendix H.

Experiment Setup and Metrics. All models are prompted with the same instruction: *"Please provide a detailed caption for the chart."* along with the chart image as input. For metrics, we use SacreBLEU [48], ROUGE [27], METEOR [5], and BERTScore [62], with our Visual Consistency Score and OCRScore.

Training Settings. We perform supervised fine-tuning using LoRA fine-tuning [15] on InternVL2.5-8B and Phi3.5-vision-4B on the CHARTCAP training set (509K). We also fine-tune Phi3.5-vision-4B using 250K original captions from ArxivCap, ChartSumm-Knoema, and ChartCheck. Additionally, We fine-tune Phi-3.5-Vision-4B on the entire training set of ChartSumm. Fine-tuned models are denoted with the name of the training dataset (e.g., Phi3.5-Vision-4B$_{\text{CHARTCAP}}$).

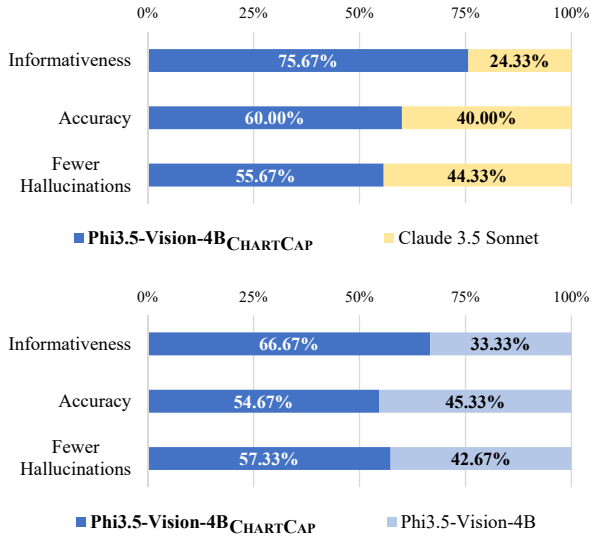

Figure 5. Results of human evaluation results comparing Phi3.5-Vision-4B$_{\text{CHARTCAP}}$ against Claude 3.5 Sonnet (top) and Phi3.5-Vision-4B (bottom) on the CHARTCAP test set.

4.1. Results on the CHARTCAP

Reference-based Metrics. Table 3 presents the results of the reference-based metrics on the CHARTCAP test set. Both InternVL2.5-8B$_{\text{CHARTCAP}}$ and Phi3.5-Vision-4B$_{\text{CHARTCAP}}$ achieve higher scores than all baseline models across all evaluation metrics. In contrast, Phi3.5-Vision-4B$_{\text{Original}}$ and Phi3.5-Vision-4B$_{\text{ChartSumm}}$ records significantly lower scores, even shows degradation of its base model. These results indicate that our fine-tuned models generate captions that align closely with the human-verified reference captions of CHARTCAP, which accurately capture the structural components and key insights of the charts sufficiently.

Human Evaluation. We conduct a human evaluation to assess caption accuracy, informativeness, and the extent of hallucination, as reference-based metrics do not measure absolute caption quality and struggle to effectively assess the degree of hallucination [22, 38]. For human evaluation, we select Phi3.5-Vision-4B$_{\text{CHARTCAP}}$ that shows the highest score on the reference-based metrics on CHARTCAP and compare head-to-head with one proprietary model (Claude 3.5 Sonnet) and one open-source model (Phi3.5-Vision-4B). We randomly sample 100 captions generated

Model	Reference-based Metrics				Visual Consistency Score			OCRScore
	sacreBLEU	ROUGE-L	METEOR	BERTScore	Large	So400M	Base	
Proprietary Model								
Claude 3.5 Sonnet	5.35	0.2265	0.2131	0.6606	0.8834	0.8771	0.8976	0.4868
Chart Expert Models								
ChartGemma-2B	0.73	0.1607	0.1082	0.5946	0.8314	0.8184	0.8565	0.2351
ChartIns-Llama2-7B	0.62	0.1144	0.0814	0.5157	0.6947	0.6759	0.7541	0.1830
Open-source Models								
InternVL2.5-78B	8.15	0.2510	0.2336	0.6642	0.8841	0.8766	0.8985	0.4677
InternVL2.5-38B	5.88	0.2331	0.2020	0.6551	0.8790	0.8700	0.8965	0.4300
InternVL2.5-26B	5.32	0.2350	0.1972	0.6546	0.8751	0.8674	0.8873	0.4144
InternVL2.5-8B	3.60	0.1770	0.1577	0.6139	0.8485	0.8372	0.8720	0.3456
InternVL2.5-8B$_{\text{CHARTCAP}}$	19.47	0.3393	0.3729	0.7238	0.8913	0.8828	0.9068	0.5089
Phi3.5-Vision-4B	8.41	0.2466	0.2501	0.6626	0.8433	0.8323	0.8696	0.4875
Phi3.5-Vision-4B$_{\text{Original}}$	0.09	0.0782	0.0384	0.5066	0.7782	0.7655	0.8137	0.1438
Phi3.5-Vision-4B$_{\text{ChartSumm}}$	1.31	0.1509	0.1322	0.6008	0.8002	0.7873	0.8207	0.2042
Phi3.5-Vision-4B$_{\text{CHARTCAP}}$	**23.82**	**0.3900**	**0.4084**	**0.7427**	**0.8933**	**0.8829**	**0.9092**	**0.5179**

Table 3. Results of reference-based metrics, Visual Consistency Scores, and OCRScore on the CHARTCAP test set.

by each model and recruit three crowd workers via AMT to select the better caption based on three aspects: (1) informativeness, (2) accuracy, and (3) fewer hallucinations. Further details on the human evaluation are provided in the Appendix J.

As shown in Figure 5, Phi3.5-Vision-4B$_{\text{CHARTCAP}}$ ranks consistently higher in all three human evaluation criteria. This consistency suggests that human judges feel Phi3.5-Vision-4B$_{\text{CHARTCAP}}$ generates more informative and accurate captions with fewer hallucinations compared to baseline models. Notably, despite having a smaller model size, Phi3.5-Vision-4B$_{\text{CHARTCAP}}$ surpasses the strong proprietary model, Claude 3.5 Sonnet, according to human judgments. This highlights fine-tuning on high-quality data could overshadow the model scale.

The Visual Consistency Score. Table 3 also presents the Visual Consistency Score and OCRScore for the CHARTCAP test set. Both InternVL2.5-8B$_{\text{CHARTCAP}}$ and Phi3.5-Vision-4B$_{\text{CHARTCAP}}$ exhibit higher Visual Consistency Score and OCRScore relative to all baselines, signifying that the captions they produce align more closely with the ground-truth chart structure and text elements. This stronger grounding in chart content further explains why they offer more accurate, informative, and low-hallucination captions than non-fine-tuned variants or other baselines.

4.2. Results on Other Human-Verified Benchmarks

We evaluate the zero-shot performance of previous captioning models on other human-verified benchmarks. We first test on the entire VisText test set, consisting of synthetic charts with human-authored captions following the caption schema from [33]. We also evaluate the models on the 1K PEW subset of Chart-to-Text, a real-world dataset whose subset has undergone human verification. The experiment setup is the same as the previous experiment.

Human Evaluation. We conduct a human evaluation on 100 samples from the VisText test set, comparing captions generated by Phi3.5-Vision-4B$_{\text{CHARTCAP}}$ with Claude 3.5 Sonnet and the human-authored ground-truth captions, under the same evaluation protocol.

As shown in Figure 6, Phi3.5-Vision-4B$_{\text{CHARTCAP}}$ outperforms both the ground-truth captions and Claude 3.5 Sonnet across all three evaluation aspects. Interestingly, human annotators judge that the model fine-tuned on CHARTCAP can generate better chart descriptions than human-authored ground-truth captions across all axes by a large margin.

The Visual Consistency Score. Table 4 – 5 present the Visual Consistency Score for the VisText test set, and the PEW subset of the Chart-to-Text dataset, respectively. As shown in both tables, InternVL2.5-8B$_{\text{CHARTCAP}}$ and Phi3.5-Vision-4B$_{\text{CHARTCAP}}$ achieve the highest Visual Consistency Scores and competitive OCRScores among all baseline models. Again, these two models surpass even the human-annotated ground-truth captions in accurately reconstructing the original chart images. In particular, for the VisText dataset, only the CHARTCAP-trained models outperform the human-authored ground-truth captions. The results also highlight the generalizability and effectiveness of captioning models trained with CHARTCAP.

Qualitative Examples. Figure 7 compares the captions and their reconstructed charts generated by Phi3.5-Vision-4B$_{\text{CHARTCAP}}$, human-authored ground-truth caption, and Claude 3.5 Sonnet for a chart from VisText. The caption generated by Phi3.5-Vision-4B$_{\text{CHARTCAP}}$ provide precise and detailed descriptions of both the chart's structural components and data. As a result, its reconstructed

Model	Visual Consistency Score		OCRScore
	Large	So400M	
Ground-truth Caption	0.9172	0.9151	0.3407
Claude 3.5 Sonnet	0.8970	0.9008	0.3286
InternVL2.5-8B	0.9093	0.9082	0.3172
InternVL2.5-8B$_{\text{CHARTCAP}}$	0.9401	0.9355	0.3360
Phi3.5-Vision-4B	0.8809	0.8814	**0.3826**
Phi3.5-Vision-4B$_{\text{CHARTCAP}}$	**0.9443**	**0.9382**	0.3414

Table 4. Visual Consistency Scores and OCRScore on the VisText test set.

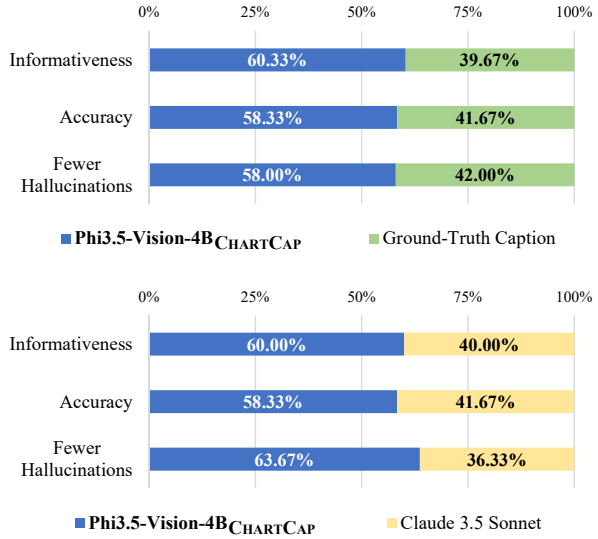

Figure 6. Human evaluation results comparing Phi3.5-Vision-4B$_{\text{CHARTCAP}}$ against ground-truth captions (top) and Claude 3.5 Sonnet (bottom) on the VisText test set.

chart closely resembles the original ground-truth, and consequently achieve the highest VCS and OCRScore. In contrast, the human-authored caption describes data trends in a simplified manner (e.g., merely stating that values increase), resulting in reconstructed charts that exhibit overly simplified data trends. Similarly, the caption generated by Claude 3.5 Sonnet describes the data trend without sufficient detail and incorrectly classifies the chart type, leading to a reconstructed chart that not only simplifies the trend but also displays an incorrect chart type.

5. Conclusion

We introduced CHARTCAP, a large-scale dataset of 565K real-world chart images paired with type-specific captions that include both structural components and key insights in a dense manner while minimizing extraneous information. We constructed CHARTCAP via a four-phase caption generation pipeline with systematically devised caption-schema and cycle consistency-based human verification. We also proposed the Visual Consistency Score to assess caption

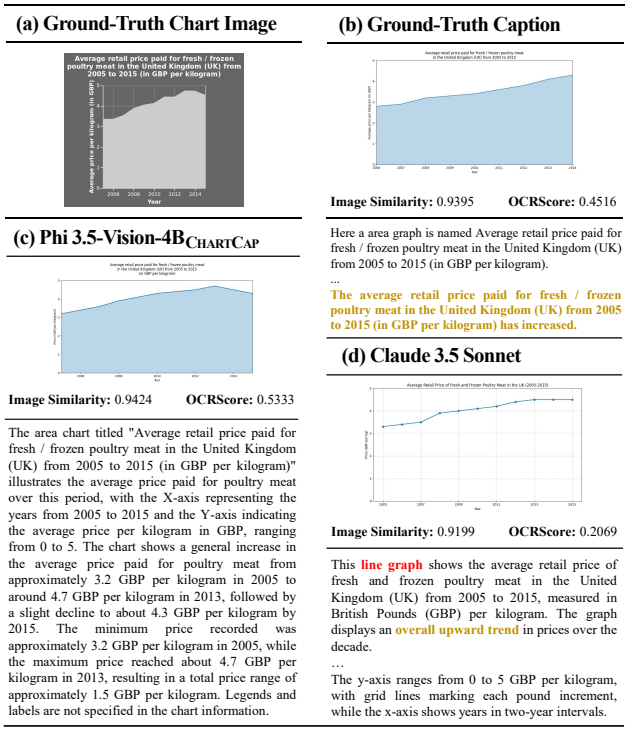

Figure 7. Qualitative examples from VisText, comparing (a) the ground-truth chart image with captions and their reconstructed charts from the captions of (b) human-authored ground-truth, (c) Phi3.5-Vision-4B$_{\text{CHARTCAP}}$, and (d) Claude 3.5 Sonnet.

Model	Visual Consistency Score		OCRScore
	Large	So400M	
Ground-truth Caption	0.6925	0.7089	0.0951
Claude 3.5 Sonnet	0.7495	0.7616	0.1603
InternVL2.5-8B	0.7362	0.7478	0.1272
InternVL2.5-8B$_{\text{CHARTCAP}}$	0.7946	0.8013	**0.1833**
Phi3.5-Vision-4B	0.7370	0.7490	0.1786
Phi3.5-Vision-4B$_{\text{CHARTCAP}}$	**0.7999**	**0.8075**	0.1789

Table 5. Visual Consistency Scores and OCRScore on the PEW subset of the Chart-to-Text.

quality by measuring the consistency between the original charts and the ones generated from captions. Models fine-tuned on CHARTCAP substantially enhance the quality of chart captions, even generates better caption than strong proprietary baseline and human annotated captions.

As a limitation, CHARTCAP utilizes a caption-schema built upon the blueprint of nine chart types defined by VLAT [24], which restricts the diversity of chart types covered. It is an interesting future work to expand the captioning schema and integrate it into the proposed pipeline, which could enable the creation of a more diverse large-scale dataset.

Acknowledgments

We thank the anonymous reviewers and Chaeyoung Lim for their valuable comments. This work is supported by the Samsung Electronics' University R&D program [Efficient fine-tuning of large multimodal models for domain-specific figure description], Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2019-II191082, SW StarLab, No. RS-2022-II220156, Fundamental research on continual meta-learning for quality enhancement of casual videos and their 3D metaverse transformation, and No. RS-2021-II211343, Artificial Intelligence Graduate School Program of Seoul National University), National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2023R1A2C2005573), and Seoul R&BD Program (VC230004) through the Seoul Business Agency (SBA) funded by The Seoul Metropolitan Government. Gunhee Kim is the corresponding author.

References

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *ArXiv preprint*, abs/2404.14219, 2024. 2, 6

[2] Mubashara Akhtar, Nikesh Subedi, Vivek Gupta, Sahar Tahmasebi, Oana Cocarascu, and Elena Simperl. Chartcheck: Explainable fact-checking over real-world chart images. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13921–13937, 2024. 1, 3

[3] Kaoru Amano, Naokazu Goda, Shin'ya Nishida, Yoshimichi Ejima, Tsunehiro Takeda, and Yoshio Ohtani. Estimation of the timing of human visual perception from magnetoencephalography. *Journal of Neuroscience*, 26(15):3981–3991, 2006. 4

[4] Anthropic. Introducing Claude 3.5 Sonnet, 2024. 2, 6

[5] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, 2005. Association for Computational Linguistics. 3, 6

[6] Bernd Burghardt and Alexander K. Hartmann. Rna secondary structure design. *Physical Review E*, 75(2), 2007. 2

[7] Sandra Carberry, Stephanie Elzer, and Seniz Demir. Information graphics: an untapped resource for digital libraries. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 581–588, 2006. 1

[8] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *ArXiv preprint*, abs/2412.05271, 2024. 2, 4, 6

[9] William S Cleveland. *The elements of graphing data*. Wadsworth Publ. Co., 1985. 1

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009. 5

[11] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18135–18143. AAAI Press, 2024. 3

[12] Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. Chartllama: A multimodal llm for chart understanding and generation. *ArXiv preprint*, abs/2311.16483, 2023. 2, 3

[13] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 3

[14] Ting-Yao Hsu, C Lee Giles, and Ting-Hao Huang. SciCap: Generating captions for scientific figures. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3258–3264, Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 1, 3

[15] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 6

[16] Hongyu Hu, Jiyuan Zhang, Minyi Zhao, and Zhenbang Sun. Ciem: Contrastive instruction evaluation method for better instruction tuning. *ArXiv preprint*, abs/2309.02301, 2023. 3

[17] Kung-Hsiang Huang, Mingyang Zhou, Hou Pong Chan, Yi R Fung, Zhenhailong Wang, Lingyu Zhang, Shih-Fu Chang, and Heng Ji. Do lvlms understand charts? analyzing and correcting factual errors in chart captioning. *ArXiv preprint*, abs/2312.10160, 2023. 1

[18] Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046, 2024. 3

[19] Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-

bench: Can language models resolve real-world github issues? *ArXiv preprint*, abs/2310.06770, 2023. 5

[20] Shankar Kantharaj, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. Chart-to-text: A large-scale benchmark for chart summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4005–4023, Dublin, Ireland, 2022. Association for Computational Linguistics. 1, 2, 3, 6

[21] Dae Hyun Kim, Enamul Hoque, and Maneesh Agrawala. Answering questions about charts and generating visual explanations. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, pages 1–13. ACM, 2020. 1

[22] Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online, 2020. Association for Computational Linguistics. 6

[23] Jill H Larkin and Herbert A Simon. Why a diagram is (sometimes) worth ten thousand words. *Cognitive science*, 11(1): 65–100, 1987. 1

[24] Sukwon Lee, Sung-Hee Kim, and Bum Chul Kwon. Vlat: Development of a visualization literacy assessment test. *IEEE transactions on visualization and computer graphics*, 23(1):551–560, 2016. 2, 4, 8

[25] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *ArXiv preprint*, abs/2403.00231, 2024. 1, 2, 3

[26] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore, 2023. Association for Computational Linguistics. 3

[27] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. 3, 6

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 5

[29] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2023. 3, 4

[30] Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. MMC: Advancing multimodal chart understanding with large-scale instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1287–1310, Mexico City, Mexico, 2024. Association for Computational Linguistics. 1, 3

[31] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 2

[32] Leo Yu-Ho Lo, Ayush Gupta, Kento Shigyo, Aoyu Wu, Enrico Bertini, and Huamin Qu. Misinformed by visualization: What do we learn from misinformative visualizations? In *Computer Graphics Forum*, pages 515–525. Wiley Online Library, 2022. 1

[33] Alan Lundgard and Arvind Satyanarayan. Accessible visualization via natural language descriptions: A four-level model of semantic content. *IEEE transactions on visualization and computer graphics*, 28(1):1073–1083, 2021. 1, 4, 7

[34] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland, 2022. Association for Computational Linguistics. 3, 6

[35] Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. UniChart: A universal vision-language pretrained model for chart comprehension and reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14662–14684, Singapore, 2023. Association for Computational Linguistics. 1, 3

[36] Ahmed Masry, Mehrad Shahmohammadi, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. Chartinstruct: Instruction tuning for chart comprehension and reasoning. *ArXiv preprint*, abs/2403.09028, 2024. 2, 3, 6

[37] Ahmed Masry, Megh Thakkar, Aayush Bajaj, Aaryaman Kartha, Enamul Hoque, and Shafiq Joty. Chartgemma: Visual instruction-tuning for chart reasoning in the wild. *ArXiv preprint*, abs/2407.04172, 2024. 2, 3, 6

[38] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online, 2020. Association for Computational Linguistics. 6

[39] Fanqing Meng, Wenqi Shao, Quanfeng Lu, Peng Gao, Kaipeng Zhang, Yu Qiao, and Ping Luo. Chartassisstant: A universal chart multimodal language model via chart-to-table pre-training and multitask instruction tuning. *ArXiv preprint*, abs/2401.02384, 2024. 2, 3

[40] Tamara Munzner. *Visualization analysis and design*. CRC press, 2014. 2, 4

[41] Yassine Ouali, Adrian Bulat, Brais Martinez, and Georgios Tzimiropoulos. Clip-dpo: Vision-language models as a source of preference for fixing hallucinations in lvlms. In *European Conference on Computer Vision*, pages 395–413. Springer, 2025. 3

[42] PaddleOCR. PaddleOCR Documentation, 2024. 5

[43] Anshul Vikram Pandey, Anjali Manivannan, Oded Nov, Margaret Satterthwaite, and Enrico Bertini. The persuasive power of data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2211–2220, 2014. 1

[44] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics. 3

[45] Fatemeh Pesaran Zadeh, Juyeon Kim, Jin-Hwa Kim, and Gunhee Kim. Text2Chart31: Instruction tuning for chart generation with automatic feedback. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11459–11480, Miami, Florida, USA, 2024. Association for Computational Linguistics. 2, 3

[46] Steven T Piantadosi, Harry Tily, and Edward Gibson. The communicative function of ambiguity in language. *Cognition*, 122(3):280–291, 2012. 4

[47] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2641–2649. IEEE Computer Society, 2015. 5

[48] Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, 2018. Association for Computational Linguistics. 6

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 3

[50] Raian Rahman, Rizvi Hasan, Abdullah Al Farhad, Md Tahmid Rahman Laskar, Md Hamjajul Ashmafee, and Abu Raihan Mostofa Kamal. Chartsumm: A comprehensive benchmark for automatic chart summarization of long and short summaries. *ArXiv preprint*, abs/2304.13620, 2023. 1, 3, 5, 6

[51] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium, 2018. Association for Computational Linguistics. 3

[52] Chase Stokes, Vidya Setlur, Bridget Cogley, Arvind Satyanarayan, and Marti A Hearst. Striking a balance: Reader takeaways and preferences when integrating text and charts. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):1233–1243, 2022. 1

[53] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *ArXiv preprint*, abs/2309.14525, 2023. 3

[54] Benny Tang, Angie Boggust, and Arvind Satyanarayan. VisText: A benchmark for semantically rich chart captioning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7268–7298, Toronto, Canada, 2023. Association for Computational Linguistics. 1, 2, 3, 6

[55] Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. 5

[56] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 5, 6

[57] Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *ArXiv preprint*, abs/2403.18715, 2024. 3

[58] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *ArXiv preprint*, abs/2406.18521, 2024. 6

[59] Zhengzhuo Xu, Sinan Du, Yiyan Qi, Chengjin Xu, Chun Yuan, and Jian Guo. Chartbench: A benchmark for complex visual reasoning in charts. *ArXiv preprint*, abs/2312.15915, 2023. 2

[60] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816, 2024. 3

[61] Xingchen Zeng, Haichuan Lin, Yilin Ye, and Wei Zeng. Advancing multimodal large language models in chart question answering with visualization-referenced instruction tuning. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 2

[62] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 2, 3, 6

[63] Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *ArXiv preprint*, abs/2311.16839, 2023. 3

[64] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-

consistent adversarial networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2242–2251. IEEE Computer Society, 2017. 2, 3

[65] Jiawen Zhu, Jinye Ran, Roy Ka-Wei Lee, Zhi Li, and Kenny Choo. AutoChart: A dataset for chart-to-text generation task. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1636–1644, Held Online, 2021. INCOMA Ltd. 2, 3