# VisText: A Benchmark for Semantically Rich Chart Captioning

**Benny J. Tang**[*]
MIT CSAIL
benjtang@csail.mit.edu

**Angie Boggust**[*]
MIT CSAIL
aboggust@csail.mit.edu

**Arvind Satyanarayan**
MIT CSAIL
arvindsatya@mit.edu

## Abstract

Captions that describe or explain charts help improve recall and comprehension of the depicted data and provide a more accessible medium for people with visual disabilities. However, current approaches for automatically generating such captions struggle to articulate the perceptual or cognitive features that are the hallmark of charts (e.g., complex trends and patterns). In response, we introduce VisText: a dataset of 12,441 pairs of charts and captions that describe the charts' construction, report key statistics, and identify perceptual and cognitive phenomena. In VisText, a chart is available as three representations: a rasterized image, a backing data table, and a *scene graph* — a hierarchical representation of a chart's visual elements akin to a web page's Document Object Model (DOM). To evaluate the impact of VisText, we fine-tune state-of-the-art language models on our chart captioning task and apply prefix-tuning to produce captions that vary the semantic content they convey. Our models generate coherent, semantically rich captions and perform on par with state-of-the-art chart captioning models across machine translation and text generation metrics. Through qualitative analysis, we identify six broad categories of errors that our models make that can inform future work.

## 1 Introduction

Studies have shown that captions can improve the recall and comprehension of the data that charts depict (Hegarty and Just, 1993; Large et al., 1995). For instance, when a caption emphasizes visually prominent features of a chart, like a peak or a sharply declining trend, readers treat this information as the key takeaway (Kim et al., 2021). Moreover, for people with visual disabilities, captions (or equivalent descriptions such as alt text) are often the only means of accessing the presented data. However, as evidenced by numerous guidelines (Jung et al., 2021), producing high-quality chart captions is a non-trivial and laborious manual process. Thus, despite these advantages, charts are only rarely captioned in practice (Lundgard and Satyanarayan, 2022).

To bridge this gap, several research communities have begun to explore methods for automatically generating chart captions, including using templates and heuristics (Demir et al., 2008; Srinivasan et al., 2019), adapting image captioning techniques (Balaji et al., 2018; Chen et al., 2019a), or via data-to-text machine translation (Kantharaj et al., 2022; Obeid and Hoque, 2020). While promising, these approaches have largely produced captions that either describe a chart's construction (e.g., *"The graph is plot between 'Number of people' x-axis over 'Movie Genres' y-axis"* (Balaji et al., 2018)) or provide statistical summaries (e.g., *"Machinery and equipment was the most valuable commodity for Singapore in 2019"* (Kantharaj et al., 2022)). However, these captions do not articulate the perceptual and cognitive features that make charts a distinctive and compelling medium for communicating data (e.g., *"Prices of Big Tech corporations seem to fluctuate but nevertheless increase over time"* (Lundgard and Satyanarayan, 2022)). Indeed, as Lundgard and Satyanarayan (2022) find, both sighted and blind readers strongly prefer captions that express this type of content.

To automatically produce such semantically richer captions, we introduce VisText: a benchmark dataset of 12,441 pairs of charts and captions. VisText makes two key extensions over prior approaches. First, VisText offers three representations of charts: a rasterized image and backing data table, as in previous work; and a *scene graph*, a hierarchical representation akin to a web page's Document Object Model (DOM), that presents an attractive midpoint between the affordances of chart-as-image and chart-as-data-table. Second, for each chart, VisText provides a synthetically generated caption detailing its construction as well as

[*]Both authors contributed equally to this work.

a crowdsourced caption describing its statistical, perceptual, and cognitive features. These crowdsourced captions represent a substantial increase in data over prior comparable datasets (Mahinpei et al., 2022; Kantharaj et al., 2022).

To demonstrate the possible uses of the VisText dataset, we train three classes of models — text-based caption models, image-guided captioning models, and semantic prefix-tuning. Text-based captioning models fine-tune large language models for VisText's chart captioning task, revealing that both data table and scene graph representations can produce compelling and semantically rich captions. Following recent advancements in image-guided translation (Sulubacak et al., 2020), we leverage the additional visual affordances in chart images to develop image-guided chart captioning models. Finally, since users often have varying preferences about the type of semantic content in their captions (Lundgard and Satyanarayan, 2022), we apply semantic prefix-tuning to each of our models, enabling them to output customizable captions.

Our models generate coherent, semantically rich captions across the VisText charts. Evaluating against standard machine translation and text generation metrics reveals that our models consistently output captions that accurately describe the chart's construction, such as its chart type, title, and axis ranges. Through qualitative analysis of our model's captions, we find that our model competently outputs semantically rich captions that describe data trends and complex patterns. Further, we categorize six common captioning errors that can inform the future development of chart captioning models on the VisText dataset.

The VisText dataset and source code are available at: `https://github.com/mitvis/vistext`.

## 2 Related work

**Heuristic-Based Chart Captioning.** Automatically generating natural language descriptions of data tables dates back to Reiter and Dale (1997). Demir et al. (2008, 2010, 2012) survey this early work and describe the process of extracting insights from a chart by evaluating a list of propositions and composing selected propositions together to produce a natural language summary. More recently, data visualization researchers have explored heuristics that calculate summary statistics and templates to assemble natural language "data facts" (Srinivasan et al., 2019) or descriptions (Cui et al., 2019). While useful, these approaches yield standardized descriptions that lack the variation and linguistic construction that characterize semantically rich captions (Lundgard and Satyanarayan, 2022).

**Chart Captioning as Image Captioning.** With rapid advances of neural image captioning (Vinyals et al., 2015; Anderson et al., 2018), researchers have begun to adapt these methods for captioning charts. For instance, Balaji et al. (2018) develop a deep learning pipeline that ingests a PNG chart image, classifies the chart type, detects and classifyies textual content present in the chart, and uses this information to generate a textual description. Chen et al. (2019a,b, 2020) propose a simpler workflow using ResNet to encode the chart image and an LSTM with Attention to decode it into a natural language description. Both approaches share a pair of limitations. The captions they produce convey relatively simplistic information about the chart (e.g., title, axis labels, etc.) or articulate concepts in visual rather than data terms (e.g., *"Dark Magenta has the lowest value"*). While both approaches contribute associated datasets, their charts and captions are synthetically generated and may not represent real-world counterparts. SciCap (Hsu et al., 2021) addresses this limitation by scraping real-world charts from 290,000 arXiv papers; however, the baseline models trained on this dataset struggle to generate semantically rich captions.

**Chart Captioning as Text Translation.** Perhaps closest to our contribution is recent work modeling chart captioning as a data-to-text problem. For instance, Spreafico and Carenini (2020) train an encoder-decoder LSTM architecture to generate a natural language caption from time series data. Similarly, Obeid and Hoque (2020) and Kantharaj et al. (2022) explore how transformer architectures can translate tabular structures into captions. These data-to-text methods are more successful than chart-as-image captioning, yielding captions that better capture relevant information from the charts and have higher BLEU scores. Nevertheless, we observe two limitations with these data-to-text approaches that motivate our contribution. First, data-to-text methods are heavily reliant on access to a chart's data table. In practice, data tables are only rarely published alongside charts and methods that recover equivalent information via OCR experience a significant drop in performance (Kantharaj et al., 2022). Second, the associated datasets

do not contain sufficient training examples of captions that express semantically rich insights about the depicted data (i.e., the perceptual and cognitive phenoma that distinguish charts as a medium as distinct from data tables (Lundgard and Satyanarayan, 2022)). As a result, while the generated captions are compelling, they are largely limited to reporting statistics which sighted and blind readers prefer less than captions that convey complex trends and patterns (Lundgard and Satyanarayan, 2022).

## 3 The VisText Dataset

We designed the VisText dataset in response to two limitations existing datasets present for generating semantically rich chart captions. First, existing datasets represent charts as either rasterized images or as data tables. While useful, these representations trade off perceptual fidelity and chart semantics in mutually exclusive ways — images capture the perceptual and cognitive phenomena that are distinctive to charts (e.g., trends or outliers) but pixels cannot express the rich semantic relationships between chart elements (e.g., estimating plotted data values using axis labels). While the vice-versa is true (Lundgard and Satyanarayan, 2022), tables also present additional caveats. There is not always a one-to-one relationship between the semantics of a data table and chart (i.e., one data table may be the source for several distinctly different charts). Moreover, data tables are rarely published alongside charts; and, automatic data table extraction is error-prone due to the diversity of chart types and visual styles as well as the difficulty of reasoning about visual occlusion (Kantharaj et al., 2022; Luo et al., 2021; Jung et al., 2017)).

Second, if existing datasets provide captions that describe perceptual or cognitive features, these captions comprise only a small portion of the dataset. At best, LineCap (Mahinpei et al., 2022) offers 3,528 such captions for line charts only, while Chart-to-Text (Kantharaj et al., 2022) estimates that roughly 15% of the sentences in its captions across a variety of chart types express such content.

In contrast, VisText provides 12,441 crowdsourced English captions that articulate statistical, perceptual, and cognitive characteristics of bar, line, and area charts. In VisText, charts are available as not only data tables and rasterized images but also as *scene graphs*. Scene graphs are hierarchical representations that better preserve perceptual fidelity and chart semantics, are often the format for publishing web-based charts, and can be recovered from chart images (Poco and Heer, 2017).

### 3.1 Data Table Collection

The data tables found in VisText are sourced from the Statista dataset of the Chart-to-Text benchmark (Kantharaj et al., 2022). The tables were collected by crawling Statista.com in December 2020 and contain real-world data related to technology, trade, retail, and sports. We process these tables to make them amenable for chart generation, including stripping formatting symbols (e.g., $ and %), standardizing data strings, and identifying the measure type of each column (i.e., quantitative, categorical, or temporal). Data tables are discarded if they do not contain at least one quantitative field and one categorical or temporal field, or if other errors occur during the processing steps. We further down select to data tables containing between 2 to 20 columns and 10 to 500 rows. If a data table has over 500 rows, we randomly sample rows. In larger data tables, this step potentially affects how salient a trend is.

### 3.2 Chart Generation and Representation

Charts in the Chart-to-Text Statista dataset all feature the same layout and visual appearance. In contrast, we aim for richer visual diversity by generating charts using the Vega-Lite visualization library (Satyanarayan et al., 2016) via the Python Altair package (VanderPlas et al., 2018). To facilitate collecting high-quality captions, we focus on univariate charts: charts that depict one quantitative observation against a categorical or temporal variable. This focus is informed by recent work in the data visualization research community which has chosen single-series line charts as the target of study for natural language descriptions (Kim et al., 2021; Stokes et al., 2022). VisText also includes single-series bar and area charts as they typically exhibit similar perceptual features to line charts.

For each data table, we iterate through pairs of univariate fields. If the pair contains a temporal field, we randomly generate an area or line chart; if the pair contains a categorical field, we randomly generate a horizontal or vertical bar chart. For diversity in layout and visual appearance, we randomly rotate axis labels and apply one of fourteen themes provided by the Vega-Lite library. These themes mimic the visual style of common chart platforms or publishers (e.g., ggplot2 or the LA Times).

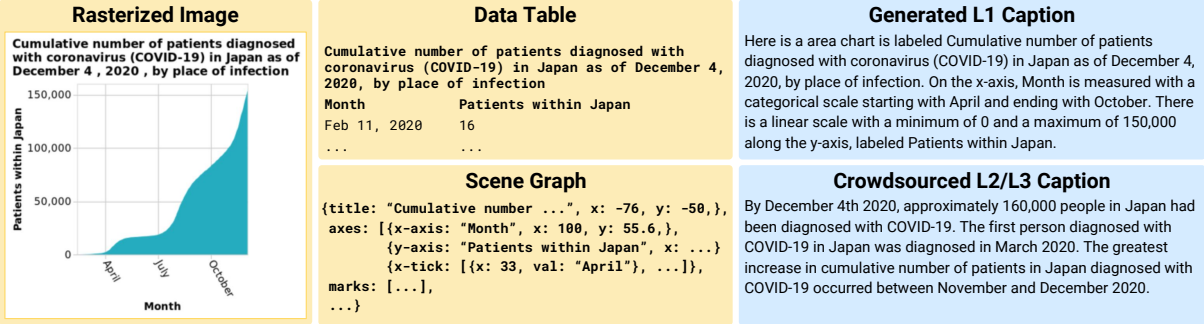



Figure 1: The VisText dataset consists of 12,441 charts represented as a rasterized image, data table, and scene graph. Before model training, each data table and scene graph is processed from its original form (shown) to a minimized and linearized text representation. Each chart is accompanied by a generated L1 caption describing the aspects of the chart's construction (e.g., chart type and axis labels) and a crowdsourced L2/L3 caption describing summary statistics and interesting trends (Lundgard and Satyanarayan, 2022).

In VisText, each chart is represented as a *rasterized image*, stored as an RGBA-encoded PNG file, as well as a *scene graph*. A scene graph is a textual representation of the rendered chart similar to a web page's Document Object Model (DOM). Scene graphs encode the position, value or content, and semantic role of all visual elements within a chart, including the individual marks (i.e., bars or points along the line), titles, axes gridlines, etc. Thus, scene graphs express the perceptual features of rasterized images in a more computationally-tractable form.

Scene graphs are a standard data structure for representing vector-based graphics — the most common format for publishing visualizations — and, thus, can be trivially recovered (e.g., by traversing the SVG text string). We extract the scene graph directly from the rendered chart using the Vega-Lite API. As most text generation models expect a linear set of input tokens, we flatten the scene graph via a depth-first traversal. To scale to large language models, we need to further reduce the size of the scene graph. Thus, we preserve the following elements which we hypothesize as being most critical for generating semantically rich captions: `title`, `title coordinates`, `axis labels`, `axis label coordinates`, `axis tick coordinates`, `mark coordinates`, and `mark sizes`. VisText includes both the original (hierarchical) and reduced (linearized) scene graphs.

### 3.3 Caption Generation and Collection

Our captioning process is guided by the framework developed by Lundgard and Satyanarayan (2022), which identifies four levels of semantic content: **L1** content enumerates aspects of the chart's construction (e.g., axis ranges); **L2** content reports summary statistics and relations (e.g., extrema); **L3** content synthesizes perceptual and cognitive phenomena (e.g., complex trends); and, **L4** content describes domain-specific insights (e.g., sociopolitical context). In subsequent studies, the authors find that while sighted readers typically prefer higher levels of semantic content, blind readers are split about the usefulness of L1 and L4 content. Thus, given these differing preferences, we define a single caption to express multiple levels of content separated across clauses or sentences. We only consider the first three levels of this model, and leave L4 content to future work. Following guidelines prescribed by the National Center for Accessible Media (NCAM), our captions begin with L1 content and then turn to L2 and L3 content (Gould et al., 2008).

We algorithmically generate L1 content and use a crowdsourced protocol to collect L2 and L3 content. This approach follows (Lundgard and Satyanarayan, 2022)'s computational considerations as well as results from Morash et al. (2015) who find that, even with instructions and guidelines, crowd workers do not describe a chart's structural elements sufficiently for blind readers. Thus, synthetically generating L1 content allows us to ensure that captions convey complete descriptions of the chart's structural elements. L1 content comprises 1 sentence conveying the chart type and title, and then 1 – 2 sentences describing the axes (including the titles, ranges, and scales). We use template randomization to generate a diverse range of L1 captions to mimic human variability and reduce the capacity of the model to overfit to a single L1 style. Three templates are defined for the first sentence and twenty-six template combinations for the subsequent sentences. During generation, we randomly select a pair of templates and fill in in-

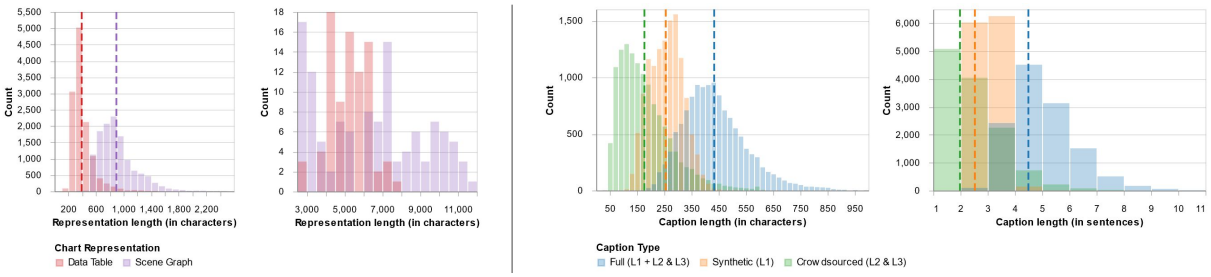



Figure 2: The VisText dataset contains data table and scene graph representations of each chart paired with L1 and L2/L3 captions. The distributions and means (dotted lines) of representations (left pair) and captions (right pair) are shown. As the distribution of chart representations has a long tail, we split it into two charts at 2,500 characters to better display the tail by re-scaling the y-axis of the second chart.

formation from the abstract chart specification. For additional diversity, we randomly drop scale information and swap template words with synonyms. Templates and synonym replacements are listed in Appendix E.2.

To crowdsource L2 and L3 content, we extend the protocol used by Lundgard and Satyanarayan (2022). After soliciting consent, we introduce the task: participants are presented with a chart image and corresponding L1 description; they are asked to write a description about the trends and patterns they observe without drawing on background knowledge or repeating L1 information. The introduction provides examples and explanations of valid and invalid responses. After acknowledging these examples, participants are asked to complete 5 random iterations of the task. To maximize the quality of our crowdsourced captions, we manually curated the charts and L1 descriptions used in the study. We discarded any charts that were challenging to read (e.g., colors were too similar, marks were not easily readable, etc.). Participants were recruited on the Prolific.co platform, took approximately 14 minutes to complete the study, and were compensated $3.25 ($14/hour). Additional details on our crowdsourcing process are in Appendix E.3.

We manually verified charts where participants failed an attention check and discarded invalid descriptions. Additionally, we manually inspected captions for personally identifiable information or offensive content. Using heuristics, we removed captions where respondents described charts as unclear or illegible and replaced newline characters with spaces. Although we attempted to fix incorrect spelling and casing errors using a similar heuristic-based approach, we observed that this process could improperly affect axis and chart names. As a result, these errors remain in our dataset.

### 3.4 Dataset Analysis

Figure 2 shows the distribution and means of the lengths of chart representations and captions. Synthetically generated L1 captions have roughly 1.5x more characters than crowdsourced L2/L3 captions ($\mu = 255$ vs. $\mu = 177$) but the average number of sentences are comparable (2.5 vs. 2). The VisText dataset consists of captions for 3,189 area charts, 6,238 bar charts, and 3,014 line charts — the roughly twice-as-many bar charts as area or line charts corresponds to the randomization of temporal fields during chart generation (Sec. 3.2). As some charts have multiple crowdsourced captions, we randomly split our dataset into training, validation, and test sets using the chart IDs to prevent data leakage across sets. This resulted in an approximate ratio of 80:10:10.

Finally, to understand the distribution of semantic content, we manually coded 2% (230) of crowdsourced captions. We followed a protocol inspired by Lundgard and Satyanarayan (2022) by breaking sentences down into independent statements and mapping these statements to their semantic content level. We marked statements as not categorizable if they did not map to the framework — for instance, if captions expressed commentary from crowd workers such as *"this chart is hard to read."* Our analysis revealed 11 L1 statements (2.4%), 180 L2 statements (39.7%), 253 L3 statments (55.7%), and 10 not categorizable statements (2.2%). While a handful express L1 content, the bulk of statements (95%) express L2 or L3 content, with approximately 1.4x L3 statements than L2.

## 4 Chart Captioning Models

To demonstrate the affordances of the VisText dataset, we train three classes of models. First, we fine-tune large language models to translate from textual chart representations to natural lan-

guage captions. These models evaluate the feasibility and impact of `scene-graph` models compared to prior `data-table` approaches (Kantharaj et al., 2022). Second, as VisText provides multiple chart representations, we adapt image-guided translation (Sulubacak et al., 2020; Cho et al., 2021) to develop two multimodal chart captioning models: `image-scene-graph` and `image-data-table`. Finally, since VisText offers captions at different semantic levels and prior work has shown significant differences in readers' preferences (Lundgard and Satyanarayan, 2022), we explore prefix-tuned models that selectively output L1, L2/L3, or L1+L2/L3 captions. Training details are in Appendix D.

### 4.1 Text-Based Chart Captioning

Informed by prior work (Kantharaj et al., 2022), we investigate text translation models for generating chart captions. In particular, Kantharaj et al. found that models that translate data tables to chart captions significantly outperform image captioning models. However, when data tables were not available, the authors found a significant drop in their models' ability to extract relevant information from the chart — an effect that was only slightly ameliorated by using OCR methods to extract text from chart images. In contrast, VisText's scene graphs can be more readily recovered from charts when data tables are not available — for instance, by processing the SVG format of web-based visualizations. Moreover, scene graphs offer a potentially richer source of information than data tables as they encode visual properties of the chart (e.g., coordinates and colors) and are less noisy than tokens recovered via OCR. Thus, to evaluate the feasibility and efficacy of scene graphs, we train a `scene-graph` text translation model and a baseline `data-table` model for comparison.

For each model, we fine-tune a pretrained ByT5 transformer model (Xue et al., 2022) on the VisText dataset. We choose ByT5 over T5 transformers (Raffel et al., 2020) because it uses a token-free, byte-encoding that eliminates the use of a tokenizer. As a result, it is robust to noisy inputs, minimizes the need for text preprocessing, and eliminates the out-of-dictionary problem. This allows our model to handle common typographical and chart reading errors in the crowdsourced L2 and L3 captions and increases generalizability to previously-unseen words that could be present in chart and axes titles.

### 4.2 Image-Guided Chart Captioning

Following recent advancements in image-guided machine translation (Sulubacak et al., 2020), we train image-guided captioning models using the VisText dataset. Images have improved text-based machine translation models by providing visual information complementary to natural language inputs. Similarly, chart images can contain visuals complementary to the textual specification. For instance, visual affordances that are important for perceiving a trend (e.g., gestalt relations, relative sizes/areas, etc.) may be obfuscated in the scene graph but better captured in the chart image.

We train three image-guided chart captioning models: `image`, `image-scene-graph`, and `image-data-table`. All models leverage the vision-language transformer model VL-T5 (Cho et al., 2021). VL-T5 is pretrained on image captioning and visual grounding tasks and was successfully applied to machine translation, making it suitable for chart captioning. We extract visual features for each VisText chart image using a Bottom-Up Feature Extractor (Anderson et al., 2018). To explore the value of images to chart captioning, our `image` model only takes in the image features, while `image-scene-graph` and `image-data-table` concatenate the image features with the chart's textual representations (scene graph or data table).

### 4.3 Semantic Prefix-Tuning

In real-world chart captioning settings, users want to vary the level of semantic content in their captions. For instance, while some blind users want verbose captions that describe the chart visuals, sighted users may only want captions that help them expose data trends (Lundgard and Satyanarayan, 2022). To develop models capable of such customization, we leverage prefix-tuning strategies alongside VisText's semantic caption breakdown. Prefix-tuning specifies a task alongside the input, permitting a single large language model to perform many different tasks. In our setting, we use prefix-tuning to specify the level of semantic content to include in the caption (Li and Liang, 2021).

We train each of our models with and without semantic prefix-tuning. With semantic prefix-tuning, we treat chart captioning as a multi-task fine-tuning problem, where the model is trained to generate the L1 and L2/L3 captions separately. In every epoch, the model sees each VisText chart twice, once with

the L1 prefix and caption and once with the L2/L3 prefix and caption.

## 5 Evaluation and Results

To evaluate the VisText dataset and our chart captioning models, we measure the readability and accuracy of generated captions and their similarity to the VisText target caption. We also qualitatively analyze the descriptiveness of generated L2/L3 captions and categorize common errors.

### 5.1 Quantitative Model Performance

We evaluate the results of our text-based and image-guided captioning models with and without prefix-tuning. We also compare to a current state-of-the-art chart captioning model that uses data table chart representations and a T5 generation model (Kantharaj et al., 2022). To measure the quality of output captions, we evaluate each model on machine translation and language generation metrics (Table 1).

**Chart images do not support captioning.** The `image` model performs the worst of all the chart captioning models. Its low perplexity and high error rates indicate it is highly confident in its inaccurate captions. While chart images contain the same information encoded in the chart's textual representations, it is presumably not adequately extracted by the model. Both the image model backbone (Cho et al., 2021) and the visual feature extractor (Anderson et al., 2018) are trained on natural images, making chart images out-of-distribution inputs that are likely to be poorly represented by these vision models. As the chart captioning task grows, model backbones, architectures, and feature extractors could be customized to chart images, which may improve image-based chart captioning.

**All models produce high quality L1 captions.** In our chart captioning setting, relation generation (Wiseman et al., 2017) measures how often the chart title, axis names, and axis scales in the input appear in the caption. Every model (except `image`) achieves a similarly-high relation generation score, indicating that every model can generate detailed L1 captions.

**Scene graphs perform as well as data tables.** Models trained on scene graph representations achieve similar performance across the evaluative metrics to models trained on data tables. As scene graphs can be more easily extracted from web-based charts images, they may be the preferred representation for future chart captioning models.

**Image-guiding does not improve captioning.** Our image-guided captioning models do not experience the significant increase in performance other image-guided translation tasks report. While in image-guided translation, images contain substantial additional information beyond the text, the image and textual representations in chart captioning often contain highly similar information. The small amount of additional information in images might benefit complex captioning tasks on multivariate charts or infographics; however, the current VisText captions rarely reference visual information not present in the scene graph or data table.

**Prefix-tuning is free.** Adding semantic prefix-tuning to our models does not significantly change their performance. Models trained with and without prefix-tuning are exposed to the same set of charts, so it is consistent that prefix-tuning would not impact the quality of output captions. Given prefix-tuned models are able to output L1, L2/L3, and L1+L2/L3 captions, prefix-tuning may be preferred if users require semantic customization.

### 5.2 Qualitative Caption Evaluation

To augment our quantitative evaluation, we qualitatively assess the descriptiveness and accuracy of the generated chart captions. Since L1 caption accuracy can be measured at scale via relation generation, we focus our evaluation on L2/L3 predictions.

Prior analysis tasked annotators with comparing the accuracy, coherence, and fluency of generated captions compared to a target caption (Kantharaj et al., 2022). Instead, our approach follows an inductive qualitative data analysis approach: iteratively analyzing captions in a "bottom-up" fashion to identify emergent patterns in how generated captions compare to the ground truth (Bingham and Witkowsky, 2021). We randomly sample 176 generated captions from the `scene-graph` model with prefix-tuning and break them into their independent L2 and L3 statements, resulting in 181 (48.27%) L2 statements and 194 (51.73%) L3 statements.

Approximately half (241 / 512) of the L2 and L3 statements made in the generated captions are factually accurate. Moreover, many of the full sentences are written in a natural, human-like manner and generated captions frequently include both compound and complex sentences. On average, every generated caption has one L3 statement and zero to

| Input | PT | BLEU ↑ | Perplexity ↓ | RG ↑ | ROUGE-1 ↑ | ROUGE-2 ↑ | ROUGE-L ↑ | ROUGE-L SUM ↑ | WMD ↓ | TER ↓ |
|---|---|---|---|---|---|---|---|---|---|---|
| Kantharaj et al. (2022) | | 0.30 ± 1.27e−3 | 28.51 ± 1.02e−1 | 1.69 ± 8.13e−3 | 0.58 ± 8.67e−4 | 0.42 ± 1.73e−3 | 0.49 ± 9.33e−4 | 0.49 ± 9.67e−4 | 0.67 ± 2.43e−3 | 66.99 ± 4.88e−2 |
| Kantharaj et al. (2022) | ✓ | 0.30 ± 2.23e−3 | 31.15 ± 9.73e−1 | 1.69 ± 8.67e−4 | 0.59 ± 1.20e−3 | 0.43 ± 1.47e−3 | 0.49 ± 1.60e−3 | 0.49 ± 1.60e−3 | 0.67 ± 2.33e−3 | 66.97 ± 3.07e−1 |
| scene-graph | | 0.32 ± 4.07e−3 | 20.96 ± 3.09e+0 | 1.82 ± 2.67e−4 | 0.56 ± 1.42e−2 | 0.39 ± 1.62e−2 | 0.47 ± 1.04e−2 | 0.47 ± 1.04e−2 | 0.68 ± 8.33e−3 | 69.34 ± 2.31e+0 |
| data-table | | 0.32 ± 2.40e−3 | 20.65 ± 2.15e+0 | 1.69 ± 1.27e−3 | 0.56 ± 7.30e−3 | 0.39 ± 8.83e−3 | 0.47 ± 6.20e−3 | 0.47 ± 6.13e−3 | 0.68 ± 3.60e−3 | 70.21 ± 7.90e−1 |
| scene-graph | ✓ | 0.32 ± 2.13e−3 | 20.02 ± 2.25e+0 | 1.78 ± 4.25e−2 | 0.56 ± 6.70e−3 | 0.39 ± 6.23e−3 | 0.47 ± 6.37e−3 | 0.47 ± 6.40e−3 | 0.68 ± 1.23e−2 | 72.55 ± 1.75e+0 |
| data-table | ✓ | 0.32 ± 4.23e−3 | 24.23 ± 1.81e+0 | 1.73 ± 8.65e−2 | 0.57 ± 5.90e−3 | 0.40 ± 5.57e−3 | 0.48 ± 5.53e−3 | 0.48 ± 5.60e−3 | 0.67 ± 1.63e−3 | 70.29 ± 2.04e+0 |
| image | | 0.07 ± 1.07e−3 | 17.36 ± 9.46e−1 | 0.78 ± 1.04e−2 | 0.34 ± 5.87e−3 | 0.14 ± 3.60e−3 | 0.25 ± 4.03e−3 | 0.25 ± 4.07e−3 | 1.11 ± 7.10e−3 | 89.03 ± 9.12e−1 |
| image-scene-graph | | 0.30 ± 3.83e−3 | 28.15 ± 2.26e+0 | 1.82 ± 2.50e−3 | 0.59 ± 1.20e−3 | 0.43 ± 2.47e−3 | 0.49 ± 2.53e−3 | 0.49 ± 2.53e−3 | 0.66 ± 1.53e−3 | 67.45 ± 2.82e−1 |
| image-data-table | | 0.29 ± 1.20e−3 | 29.81 ± 2.62e−1 | 1.81 ± 1.20e−3 | 0.59 ± 5.67e−4 | 0.44 ± 1.03e−3 | 0.49 ± 2.17e−3 | 0.49 ± 2.23e−3 | 0.66 ± 6.33e−4 | 66.80 ± 2.77e−2 |
| image | ✓ | 0.07 ± 1.33e−3 | 24.08 ± 1.77e+0 | 0.58 ± 1.34e−2 | 0.33 ± 6.20e−3 | 0.13 ± 3.17e−3 | 0.23 ± 4.67e−3 | 0.23 ± 4.67e−3 | 1.11 ± 1.90e−3 | 100.04 ± 6.57e+0 |
| image-scene-graph | ✓ | 0.32 ± 9.90e−3 | 15.50 ± 4.45e−1 | 1.82 ± 2.67e−4 | 0.54 ± 8.23e−3 | 0.38 ± 4.93e−3 | 0.45 ± 3.63e−3 | 0.45 ± 3.57e−3 | 0.69 ± 7.13e−3 | 81.95 ± 4.53e+0 |
| image-data-table | ✓ | 0.32 ± 2.87e−3 | 17.29 ± 1.28e+0 | 1.81 ± 4.50e−3 | 0.54 ± 6.67e−3 | 0.38 ± 7.50e−3 | 0.45 ± 5.60e−3 | 0.45 ± 5.50e−3 | 0.68 ± 1.33e−3 | 80.21 ± 1.34e+0 |

Table 1: We compare our text-based models (`scene-graph` and `data-table`), our image-guided models (`image`, `image-scene-graph`, and `image-data-table`), and semantic prefix-tuning (PT) models to prior chart captioning models (Kantharaj et al., 2022). We evaluate each model using machine translation and text generation metrics, including BLEU (Papineni et al., 2002), Perplexity, Relation Generation (RG) (Wiseman et al., 2017), ROUGE (Lin, 2004), Word Mover's Distance (WMD) (Kusner et al., 2015), and Translational Error Rate (TER) (Snover et al., 2006). We report the mean and standard deviation of three independent models. Darker colors indicate better scores.

two L2 statements. Often this takes the form of a L3 general trend statement (e.g., *"The median annual family income in Canada has increased from 2000 to 2018"*) accompanied by an L2 minimum and maximum statement (*"The highest was in 2015 at 80k and the lowest was in 2000"*). For the remaining half of analyzed captions, we identified the following recurring types of errors:

**Identity Errors.** We identify 86 identity errors (22.93% of analyzed statements). An identity error occurs when an L2 or L3 statement incorrectly reports the independent variable for a given (often correctly identified) trend. For bar charts, this error means incorrectly reporting the categorical label associated with a bar (e.g., in Appendix Figure 5c: *"The most popular music activity is vinyl albums and vinyl singles"* should be *"The most popular music activity is tickets for festivals"*). For area and line charts, this error means incorrectly identifying the temporal point or range of the trend. With bar charts, in particular, we observed that the identities were often "off-by-one" (i.e., identifying a minimum or maximum value, but attributing it to the second-highest or second-lowest category).

**Value Errors.** A value error occurs when the quantitative data value of a statement is incorrect. Of the captions we analyzed, 3.20% (12) of statements contained a value error. For instance, as shown in Appendix Figure 4c, for the caption *"The total gate revenue from sporting events worldwide by region from 2006 to 2015 has increased from around 15 billion dollars to around 15 billion dollars"*, the value should be around 18 billion dollars. If it is ambiguous whether an error is an Identity or Value Error, we classify it as the former.

**Direction Errors.** A direction error occurs when the direction (which can be *increasing*, *decreasing*, or *stable*) of a trend in an L3 statement is incorrect. We uncovered 32 direction errors (8.53% of analyzed statements). For instance, in the caption *"The per capita consumption of sweet corn in the US has increased from 2000 to 2019"* (Appendix Figure 3c), the trend is actually decreased. In most direction errors, the identity (i.e., temporal range) is correct.

**Stability Errors.** A stability error occurs when the magnitude of a direction or the variance in a trend is incorrect. This can often refer to how much a trend is increasing or decreasing, such as *rapidly* or *slowly*, as well as whether it's a *steady* change or *highly-fluctuating* change. In Appendix Figure 4b, *"The comparable sales growth of Sam's Club in the United States from fiscal year 2006 to 2020 has been steadily decreasing from 2006 to 2020."* should read *"The comparable sales growth of Sam's Club in the United States from fiscal year 2006 to 2020 has been highly-fluctuatingly decreasing from 2006 to 2020."* 1.07% (4) of the statements we analyzed contained this error.

**Repetition.** Repetition is when a caption repeats a previously-generated claim, regardless of its correctness. 117 (31.2%) statements contained repetition, making it the most common error we encountered. For example, in Appendix Figure 4a, we see *"The average age at widow hood in the Netherlands has increased from 2008 to 2018. The average age at widow hood in the Netherlands has increased from 2008 to 2018."* Repetition is a known problem for text generation models with transformer backbones, like our chart captioning models (Fu et al., 2021).

**Nonsensical Errors.** If a L2 or L3 statement cannot be understood in context of the chart, or makes a fundamental mistake in interpretation, we label it as nonsensical error. We encountered 20 nonsensical errors in addition to the 395 statements we analyzed. For example, in Appendix Figure 5b, *"The most popular visitors was Harry Potter in 1999 and 2009."* misinterprets the chart. It might instead correctly read *"The destination with the most visitors after the TV/movie's release was New Zealand for The Lord of the Rings"*.

## 6 Discussion

We present VisText, a chart captioning dataset of 12,441 charts and semantically rich captions. The VisText charts are represented as a rasterized image, data table, and scene graph to provide diverse and complementary data modalities. Using VisText, we fine-tune large language models to generate natural language captions from textual chart representations and integrate image-guided chart captioning to leverage multimodal information. Utilizing the varied semantic content in VisText captions, we develop semantic prefix-tuned models that output semantically customized captions to meet diverse user needs. Evaluations reveal that our models output precise and semantically descriptive captions, performing on par with state-of-the-art chart captioning models (Kantharaj et al., 2022) across machine translation and text generation metrics.

Looking ahead, while accessibility remains a key domain that would benefit from automated chart captioning, and deploying automated chart captioning models into the field is an exciting prospect, we believe the most promising approach for future work lies in "mixed-initiative" (i.e., human + AI) chart authoring systems. In particular, as we describe in our Ethics Statement below, chart captioning models are currently prone to make a number of factual inaccuracies which can have severe harmful consequences. On the other hand, by integrating these models into chart authoring systems (e.g., Tableau, Charticulator, Data Illustrator, or Lyra), chart authors can intervene and make any necessary corrections. Indeed, such integration offers exciting opportunities to develop novel interactive methods for verifying generated captions. For instance, models like ours could generate an initial caption (or set of captions) based on the chart currently being authored; as the system has access to all three representations of the chart (the backing data table, chart image, and structured scene graph), it might automatically segment the caption into independent "data segments" and interactively link and map them to rows in the table or regions on the chart, akin to Kori (Latif et al., 2021).

## Limitations

**Computational Constraints.** Despite using modern GPUs, with large amounts of memory, we were forced to use the smallest-parameter variants of T5 and ByT5 as we encountered out-of-memory errors with the larger alternatives. More problematically, the quadratic relationship between sequence length and time/space complexity of transformer architectures (Vaswani et al., 2017), especially when using byte-level sequences (Xue et al., 2022), has had a significant impact on our model performance. In particular, to be computationally tractable, we were forced us to truncate our input and output sequences to, at most, 1,024 and 512 characters respectively (1,024 coming from the underlying ByT5 architecture (Xue et al., 2022)).

These character thresholds have likely had an outsized effect on `scene-graph` models. For instance, due to these character limits, we reduced scene graph sequences to only a minimal set of visual characteristics; VisText also includes the raw, unprocessed scene graphs which offer a richer source of information about the visual features that are important to how people decode charts (e.g., bounding boxes, color) but were unavailable to our models. Moreover, as Figure 2 shows, even with this reduced representation, the mean length of scene graph sequences is 948 characters (cf. 426 characters for data tables) with a wide distribution. Thus, despite `scene-graph` models achieving comparable performance to `data-table` models, the former saw a much smaller proportion of complete sequences as compared to the latter. This truncation step additionally negatively impacts charts with long titles or axis names — in such cases, we observed that the L2 or L3 caption would be altogether truncated before generation.

**Chart Types and the Visualization Design Space.** VisText is scoped to only univariate bar, area, and line charts. We chose to begin with these chart types informed by data visualization research that has focused on studying natural language descriptions of single-series line charts — a basic, but commonly occurring chart type that offers a compelling target of study as it most visibly surfaces any poten-

tial trends in the data (Kim et al., 2021; Stokes et al., 2022). Future work can now begin to consider more complex chart forms in a step-by-step manner. For instance, moving from univariate bar, area, and line charts to multivariate versions of these chart types (i.e., stacked bars and areas, grouped bars, and multi-series line charts). From there, work can also consider chart types that surface perceptual and cognitive phenomena in visually distinct ways (e.g., scatterplots, where trends appear as clusters of points; heatmaps, where color saturation often encodes a trend; or maps, where color or other layered elements such as symbols are used to represent data values). Finally, automated methods for captioning visualizations may eschew chart typologies altogether in favor of visualization grammars — by offering a more composable and combinatorial approach to the design space (Wilkinson, 2012), learning over visualization grammars may offer a more robust approach to captioning highly customized or unique visual forms.

For each future work direction, we anticipate scene graph representations to prove more fruitful than the data table. As the complexity of the visualization increases, its relationship to the data table only grows more ambiguous; the scene graph, on the other hand, directly encodes the visual form and thus remains faithful to it. As a result, to support such future work, VisText provides the raw specifications used to produce our charts (via the Vega-Lite visualization grammar (Satyanarayan et al., 2016)) as well as the raw, hierarchical scene graphs prior to our linearization and reduction step.

## Ethics Statement

**The Consequences of Incorrect Captions.** Weidinger et al. (2021) comprehensively survey the risks associated with the large language models (LLMs) that underlie our contribution. Of the six categories of risk they identify, harms stemming from models producing factually incorrect statements are not only most pertinent to our work, but are likely heighted as compared to general uses of LLMs given the context we are addressing: automatically captioning charts. In particular, people most often consume charts and visualizations in order to make data-driven decisions (Keim et al., 2008) — for instance, about whether to evacuate ahead of a hurricane (Padilla et al., 2018), or health & safety during the pandemic (Shneiderman, 2020). Moreover, recent results have shown that readers not only fixate for longer and are more likely to recall the textual content of and around visualizations (Borkin et al., 2015) but this textual content can strongly influence the takeaway message readers leave with even when it is at odds with the depicted data (Kong et al., 2018, 2019). Finally, these issues are exacerbated by the persuasive and rhetorical force of data and charts (Kennedy et al., 2016; Hullman and Diakopoulos, 2011), that often project a sense of authority and certainty (Correll, 2019). As a result, readers may not think to double check the accuracy of chart captions, and inaccurate statements that models may produce could lead to harmful downstream decisions.

To proceed ethically with this line of research, we believe that advances in data and modeling need to be closely followed by attention devoted to mitigating the risks of incorrect statements. At base, automatically generated captions should be identified as such at the forefront to raise readers' awareness about the potential for incorrect statements. And, interactive visual linking strategies (such as those explored by Kong and Agrawala (2012); Kim et al. (2018)) could be deployed to help readers manually verify the constituent statements of a caption against the chart. These strategies, however, place the burden of harm mitigation on readers. Thus, an alternate approach might never surface automatically generated captions to readers directly but instead use them as part of mixed-initiative systems for jointly authoring visualization and text, such as Kori (Latif et al., 2021). In such systems, automated chart captioning models would help to accelerate the authoring process — combatting the blank slate problem by providing an initial summary of the chart — and chart authors would make any necessary corrections prior to publication.

Besides these human-computer interaction (HCI) approaches for mitigating harm, an equally important direction for future work should leverage interpretability techniques to more deeply study what the models are learning. To what degree are chart captioning models stochastic parrots (Bender et al., 2021), and how much do they understand the information charts depict?

**Automated Captioning for Accessibility.** Although accessibility is a guiding motivation for the bulk of work in automated captioning (be it image captioning or, as in our case, chart captioning), studies find mixed reactions, at best, about these approaches among people with disabilities (PWDs).

For instance, accessibility educator and researcher Chancey Fleet described Facebook's automatic image descriptions as *"famously useless in the Blind community"* despite *"garner[ing] a ton of glowing reviews from mainstream outlets"* (Fleet, 2021; Hanley et al., 2021). This disconnect appears to stem from a more fundamental mismatch between what PWDs describe as their captioning needs, and what the research community — particularly through its automatic, quantitative evaluations — prioritizes (Jandrey et al., 2021). In particular, surveys with PWDs repeatedly surface the contextual nature of captions. Bennett et al. (2021) find that the context of use shapes the degree to which PWD are comfortable with captions describing people's race, gender, and disabilities — for instance, changing their preferences if they were in a white, cisgender, nondisabled, and professional company versus their own community. Similarly, Jung et al. (2022) find shifting preferences for the content image descriptions should convey across different photo activites — for example, when viewing or taking photos, participants wished for descriptions that conveyed spatial cues whereas when searching or reminiscing about photos, participants hoped for descriptions to connect to personal data or differentiating details.

In contrast, quantitative metrics of model performance compare generated captions to a single "ground truth" caption. This framing of success not only makes it difficult to develop contextually-varying caption generation but can actively penalize such investigations. For instance, with our work, we explored how prefix-tuning can be used to develop models that are responsive to users' preferences about semantic content. However, as described in Sec. 5.1, existing quantitative metrics of model performance (e.g., BLEU, ROUGE, WMD, and TER) show a drop in model performance despite our qualitative analysis indicating that these captions are indeed high quality.

Finally, our exploration of semantic prefix-tuning represents only a very preliminary step towards addressing the contextual captioning needs of PWDs. In particular, the semantic labels VisText assigns to captions were derived from prior work (Lundgard and Satyanarayan, 2022) that only explored natural language descriptions when *consuming presentations* of visualizations — one task from a broader palette (Brehmer and Munzner, 2013). Future work might instead extend the VisText dataset — and corresponding models — to consider captions for a broader range of tasks including consuming visualizations for scientific discovery, enjoyment or, producing, searching, or querying visualizations (Brehmer and Munzner, 2013).

## Acknowledgements

We thank Nicolás Kennedy and Alan Lundgard for their work developing an initial version of our crowdsourced study protocol. This research was sponsored by a Google Research Scholar Award, an NSF Award #1900991, the MLA@CSAIL initiative, and by the United States Air Force Research Laboratory under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

## References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086, Los Alamitos, CA, USA. IEEE Computer Society.

Abhijit Balaji, Thuvaarakkesh Ramanathan, and Venkateshwarlu Sonathi. 2018. Chart-text: A fully automated chart image descriptor. *ArXiv*, abs/1812.10636.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 610–623.

Cynthia L Bennett, Cole Gleason, Morgan Klaus Scheuerman, Jeffrey P Bigham, Anhong Guo, and Alexandra To. 2021. "it's complicated": Negotiating accessibility and (mis) representation in image descriptions of race, gender, and disability. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 1–19.

Andrea J Bingham and Patricia Witkowsky. 2021. Deductive and inductive approaches to qualitative data analysis. *Analyzing and interpreting qualitative data: After the interview*, pages 133–146.

Michelle A Borkin, Zoya Bylinskii, Nam Wook Kim, Constance May Bainbridge, Chelsea S Yeh, Daniel Borkin, Hanspeter Pfister, and Aude Oliva. 2015. Beyond memorability: Visualization recognition and recall. *IEEE Transactions on Visualization and Computer Graphics (VIS)*, 22(1):519–528.

Matthew Brehmer and Tamara Munzner. 2013. A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics (VIS)*, 19(12):2376–2385.

Charles Chen, Ruiyi Zhang, Sungchul Kim, Scott Cohen, Tong Yu, Ryan Rossi, and Razvan Bunescu. 2019a. Neural caption generation over figures. In *Adjunct Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the ACM International Symposium on Wearable Computers (UbiComp/ISWC)*, UbiComp/ISWC '19 Adjunct, page 482–485, New York, NY, USA. Association for Computing Machinery.

Charles Chen, Ruiyi Zhang, Eunyee Koh, Sungchul Kim, Scott Cohen, and Ryan Rossi. 2020. Figure captioning with relation maps for reasoning. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1526–1534.

Charles Chen, Ruiyi Zhang, Eunyee Koh, Sungchul Kim, Scott Cohen, Tong Yu, Ryan Rossi, and Razvan Bunescu. 2019b. Figure captioning with reasoning and sequence-level training. *arXiv preprint arXiv:1906.02850*.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1931–1942. PMLR.

Michael Correll. 2019. Ethical dimensions of visualization research. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 1–13.

Zhe Cui, Sriram Karthik Badam, M Adil Yalçin, and Niklas Elmqvist. 2019. Datasite: Proactive visual data exploration with computation of insight-based recommendations. *Information Visualization*, 18(2):251–267.

Seniz Demir, Sandra Carberry, and Kathleen F. McCoy. 2008. Generating textual summaries of bar charts. In *Proceedings of the International Natural Language Generation Conference (INLG)*, INLG '08, page 7–15, USA. Association for Computational Linguistics.

Seniz Demir, Sandra Carberry, and Kathleen F McCoy. 2012. Summarizing information graphics textually. *Computational Linguistics*, 38(3):527–574.

Seniz Demir, David Oliver, Edward Schwartz, Stephanie Elzer, Sandra Carberry, and Kathleen F McCoy. 2010. Interactive sight into information graphics. In *Proceedings of the International Cross Disciplinary Conference on Web Accessibility (W4A)*, pages 1–10.

Chancey Fleet. 2021. Things which garner a ton of glowing reviews from mainstream outlets without being of much use to disabled people. For instance, Facebook's auto image descriptions, much loved by sighted journos but famously useless in the Blind community. *Twitter*. https://twitter.com/ChanceyFleet/status/1349211417744961536.

Zihao Fu, Wai Lam, Anthony Man-Cho So, and Bei Shi. 2021. A theoretical analysis of the repetition problem in text generation. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*.

Bryan Gould, Trisha O'Connell, and Geoff Freed. 2008. Effective practices for description of science content within digital talking books: Guidelines for describing stem images.

Margot Hanley, Solon Barocas, Karen Levy, Shiri Azenkot, and Helen Nissenbaum. 2021. Computer vision and conflicting values: Describing people with automated alt text. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 543–554.

Mary Hegarty and Marcel-Adam Just. 1993. Constructing mental models of machines from text and diagrams. *Journal of memory and language*, 32(6):717–742.

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

Jessica Hullman and Nick Diakopoulos. 2011. Visualization rhetoric: Framing effects in narrative visualization. *IEEE Transactions on Visualization and Computer Graphics (VIS)*, 17(12):2231–2240.

Alessandra Helena Jandrey, Duncan Dubugras Alcoba Ruiz, and Milene Selbach Silveira. 2021. Image descriptions' limitations for people with visual impairments: Where are we and where are we going? In *Proceedings of the Brazilian Symposium on Human Factors in Computing Systems (IHC)*, IHC '21, New York, NY, USA. Association for Computing Machinery.

Crescentia Jung, Shubham Mehta, Atharva Kulkarni, Yuhang Zhao, and Yea-Seul Kim. 2021. Communicating visualizations without visuals: Investigation of visualization alternative text for people with visual impairments. *IEEE Transactions on Visualization and Computer Graphics (VIS)*, 28(1):1095–1105.

Daekyoung Jung, Wonjae Kim, Hyunjoo Song, Jeong-in Hwang, Bongshin Lee, Bohyoung Kim, and Jinwook Seo. 2017. Chartsense: Interactive data extraction from chart images. In *Proceedings of the Conference*

*on Human Factors in Computing Systems (CHI)*, CHI '17, page 6706–6717, New York, NY, USA. Association for Computing Machinery.

Ju Yeon Jung, Tom Steinberger, Junbeom Kim, and Mark S. Ackerman. 2022. *"So What? What's That to Do With Me?" Expectations of People With Visual Impairments for Image Descriptions in Their Personal Photo Activities*, page 1893–1906. Association for Computing Machinery, New York, NY, USA.

Shankar Kantharaj, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. 2022. Chart-to-text: A large-scale benchmark for chart summarization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4005–4023, Dublin, Ireland. Association for Computational Linguistics.

Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. 2008. Visual analytics: Definition, process, and challenges. In *Information visualization*, pages 154–175. Springer.

Helen Kennedy, Rosemary Lucy Hill, Giorgia Aiello, and William Allen. 2016. The work that visualisation conventions do. *Information, Communication & Society*, 19(6):715–735.

Dae Hyun Kim, Enamul Hoque, Juho Kim, and Maneesh Agrawala. 2018. Facilitating document reading by linking text and tables. In *Proceedings of the Annual ACM Symposium on User Interface Software and Technology (UIST)*, UIST '18, page 423–434, New York, NY, USA. Association for Computing Machinery.

Dae Hyun Kim, Vidya Setlur, and Maneesh Agrawala. 2021. Towards understanding how readers integrate charts and captions: A case study with line charts. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 1–11.

Ha-Kyung Kong, Zhicheng Liu, and Karrie Karahalios. 2018. Frames and slants in titles of visualizations on controversial topics. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 1–12.

Ha-Kyung Kong, Zhicheng Liu, and Karrie Karahalios. 2019. Trust and recall of information across varying degrees of title-visualization misalignment. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 1–13.

Nicholas Kong and Maneesh Agrawala. 2012. Graphical overlays: Using layered elements to aid chart reading. *IEEE Transactions on Visualization and Computer Graphics (VIS)*, 18(12):2631–2638.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research (PMLR)*, pages 957–966, Lille, France. PMLR.

Andrew Large, Jamshid Beheshti, Alain Breuleux, and Andre Renaud. 1995. Multimedia and comprehension: The relationship among text, animation, and captions. *Journal of the American society for information science*, 46(5):340–347.

Shahid Latif, Zheng Zhou, Yoon Kim, Fabian Beck, and Nam Wook Kim. 2021. Kori: Interactive synthesis of text and charts in data documents. *IEEE Transactions on Visualization and Computer Graphics (VIS)*, 28(1):184–194.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7871–7880, Online. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation.

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

Chin-Yew Lin and Franz Josef Och. 2004. ORANGE: a method for evaluating automatic evaluation metrics for machine translation. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 501–507, Geneva, Switzerland. COLING.

Alan Lundgard and Arvind Satyanarayan. 2022. Accessible visualization via natural language descriptions: A four-level model of semantic content. *IEEE Transactions on Visualization and Computer Graphics (VIS)*, 28(1):1073–1083.

Junyu Luo, Zekun Li, Jinpeng Wang, and Chin-Yew Lin. 2021. Chartocr: Data extraction from charts images via a deep hybrid framework. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. The Computer Vision Foundation.

Anita Mahinpei, Zona Kostic, and Chris Tanner. 2022. Linecap: Line charts for data visualization captioning models. In *IEEE Visualization and Visual Analytics (VIS)*, pages 35–39. IEEE.

Valerie S Morash, Yue-Ting Siu, Joshua A Miele, Lucia Hasty, and Steven Landau. 2015. Guiding novice web workers in making image descriptions using templates. *ACM Transactions on Accessible Computing (TACCESS)*, 7(4):1–21.

Jason Obeid and Enamul Hoque. 2020. Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model. In *Proceedings of the International Conference on Natural Language Generation (INLG)*, pages 138–147, Dublin, Ireland. Association for Computational Linguistics.

Lace M Padilla, Sarah H Creem-Regehr, Mary Hegarty, and Jeanine K Stefanucci. 2018. Decision making with visualizations: a cognitive framework across disciplines. *Cognitive research: principles and implications*, 3(1):1–25.

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

Jorge Poco and Jeffrey Heer. 2017. Reverse-engineering visualizations: Recovering visual encodings from chart images. *Computer Graphics Forum*, 36(3):353–363.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Conference on Machine Translation (WMT)*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1).

Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.

Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer. 2016. Vega-lite: A grammar of interactive graphics. *IEEE Transactions on Visualization and Computer Graphics (VIS)*, 23(1):341–350.

Ben Shneiderman. 2020. Data Visualization's Breakthrough Moment in the COVID-19 Crisis.

M Six Silberman, Bill Tomlinson, Rochelle LaPlante, Joel Ross, Lilly Irani, and Andrew Zaldivar. 2018. Responsible research with crowds: pay crowdworkers at least minimum wage. *Communications of the ACM*, 61(3):39–41.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Andrea Spreafico and Giuseppe Carenini. 2020. Neural data-driven captioning of time-series line charts. In *Proceedings of the International Conference on Advanced Visual Interfaces (AVI)*, AVI '20, New York, NY, USA. Association for Computing Machinery.

Arjun Srinivasan, Steven M. Drucker, Alex Endert, and John Stasko. 2019. Augmenting visualizations with interactive data facts to facilitate interpretation and communication. *IEEE Transactions on Visualization and Computer Graphics (VIS)*, 25(1):672–681.

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

Umut Sulubacak, Ozan Caglayan, Stig-Arne Grönroos, Aku Rouhe, Desmond Elliott, Lucia Specia, and Jörg Tiedemann. 2020. Multimodal machine translation through visuals and speech. *Machine Translation*, 34(2–3):97–147.

Jenny Tang, Eleanor Birrell, and Ada Lerner. 2022. Replication: How well do my results generalize now? the external validity of online privacy and security surveys. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 367–385.

Jacob VanderPlas, Brian Granger, Jeffrey Heer, Dominik Moritz, Kanit Wongsuphasawat, Arvind Satyanarayan, Eitan Lees, Ilia Timofeev, Ben Welsh, and Scott Sievert. 2018. Altair: interactive statistical visualizations for python. *Journal of Open Source Software*, 3(32):1057.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. 2015. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, Los Alamitos, CA, USA. IEEE Computer Society.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Leland Wilkinson. 2012. The grammar of graphics. In *Handbook of computational statistics*, pages 375–414. Springer.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

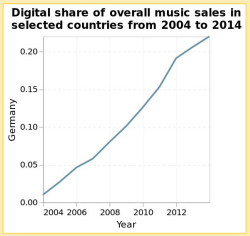
Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen,

Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics (ACL)*, 10:291–306.

# A Example Model Outputs

**Rasterized Image**




**Scene Graph**

```
title Digital share of overall music sales in selected countries from 2004 to 2014 x 100 y -34 x-axis x 100 y 21 Year y-axis x -32 y 100 Germany xtick x 0 val 2004 x 40 val 2006 x 80 val 2008 x 120 val 2010 x 160 val 2012 ytick y 200 val 0.00 y 155 val 0.05 y 109 val 0.10 y 64 val 0.15 y 18 val 0.20 marks line XY 0 190.909 desc XY 20.038 175.455 desc XY 40.022 158.182 desc XY 60.005 147.273 desc XY 79.989 127.273 desc XY 100.027 108.182 desc XY 120.011 85.455 desc XY 139.995 60.909 desc XY 159.9 ...
```

**Model Generated L1 Caption**

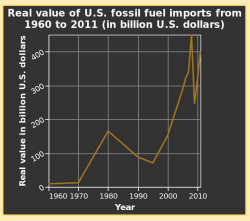
Digital share of overall music sales in selected countries from 2004 to 2014 is a line chart. The y-axis measures Germany while the x-axis measures Year.

**Model Generated L2/L3 Caption**

The digital share of music sales in Germany has increased from 2004 to 2014. The digital share of music sales in Germany has increased from 2004 to 2014.

(a) The `scene-graph` model's output L1 caption and L2/L3 caption for a VisText line chart of the *"Digital share of overall music sales in selected counties from 2004 to 2014"*. The model correctly identifies the chart's title and axis, and it accurately identifies the upward trend. However, it repeats this claim twice. See Section 5.2 for details on model repetition.

**Rasterized Image**




**Scene Graph**

```
title Real value of U.S. fossil fuel imports from 1960 to 2011 (in billion U.S. dollars) x 100 y -33.84 x-axis x 100 y 21 Year y-axis x -29.088 y 100 Real value in billion U.S. dollars xtick x 0 val 1960 x 39 val 1970 x 78 val 1980 x 118 val 1990 x 157 val 2000 x 196 val 2010 ytick y 200 val 0 y 156 val 100 y 111 val 200 y 67 val 300 y 22 val 400 marks line XY 0 196.04 desc XY 39.221 194.516 desc XY 78.43 126.778 desc XY 117.651 160.747 desc XY 137.256 168.604 desc XY 156.861 132.467 desc XY 176 ...
```

**Model Generated L1 Caption**

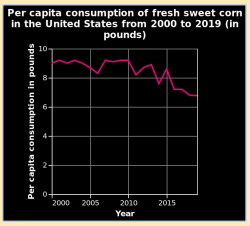
Here a line graph is called Real value of U.S. fossil fuel imports from 1960 to 2011 (in billion U.S. dollars). The y-axis measures Real value in billion U.S. dollars while the x-axis measures Year.

**Model Generated L2/L3 Caption**

The real value of US fossil fuel imports has increased from 1960 to 2011. The real value of US fossil fuel imports has increased from 1960 to 2011. The real value of US fossil fuel imports has increased from 1960 to 2011.

(b) The `scene-graph` model's output L1 caption and L2/L3 caption for a VisText line chart of the *"Real values of U.S. fossil fuel imports from 1960 to 2011 (in billion U.S. dollars)"*. The model correctly identifies the chart's title and axis, and it accurately identifies the upward trend. However, it repeats this claim twice. See Section 5.2 for details on model repetition.

**Rasterized Image**




**Scene Graph**

```
title Per capita consumption of fresh sweet corn in the United States from 2000 to 2019 (in pounds) x 100 y -54.5 x-axis x 100 y 21 Year y-axis x -23 y 100 Per capita consumption in pounds xtick x 0 val 2000 x 53 val 2005 x 105 val 2010 x 158 val 2015 ytick y 200 val 0 y 160 val 2 y 120 val 4 y 80 val 6 y 40 val 8 y 0 val 10 marks line XY 0 20.0 desc XY 10.548 16.0 desc XY 21.066 20.0 desc XY 31.585 16.0 desc XY 42.104 20.0 desc XY 52.651 26.0 desc XY 63.17 34.0 desc XY 73.689 16.0 desc XY 84.20 ...
```

**Model Generated L1 Caption**

This line graph is called Per capita consumption of fresh sweet corn in the United States from 2000 to 2019 (in pounds). The y-axis measures Per capita consumption in pounds while the x-axis measures Year.

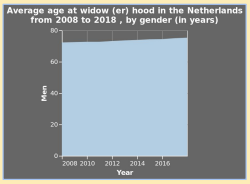
**Model Generated L2/L3 Caption**

The per capita consumption of sweet corn in the US has increased from 2000 to 2019. The per capita consumption of sweet corn in the US has increased from 2000 to 2019.

(c) The `scene-graph` model's output L1 caption and L2/L3 caption for a VisText line chart of the *"Per capita consumption of fresh sweet corn in the United States from 2000 to 2019 (in pounds)"*. The model correctly identifies the chart's title and axis. However, it makes a direction error and claims the increasing trend is actually decreasing. See Section 5.2 for details on direction errors.

Figure 3: `scene-graph` model captions of VisText line charts.

**Rasterized Image**




**Scene Graph**

```
title Average age at widow (er) hood in the Netherlands from 2008 to 2018 , by gender (in years) x 100 y -37 x-axis x 100 y 21 Year y-axis x -23 y 100 Men xtick x 0 val 2008 x 40 val 2010 x 80 val 2012 x 120 val 2014 x 160 val 2016 ytick y 200 val 0 y 150 val 20 y 100 val 40 y 50 val 60 y 0 val 80 marks area XY 0 19.25 H 180.75 desc XY 20.038 18.75 H 181.25 desc XY 40.022 18.25 H 181.75 desc XY 60.005 18.25 H 181.75 desc XY 79.989 17.0 H 183 desc XY 100.027 16.0 H 184 desc XY 120.011 15.25 H 184 ...
```

**Model Generated L1 Caption**

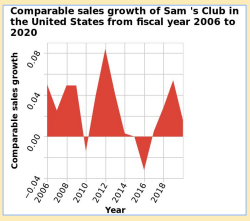
Average age at widow (er) hood in the Netherlands from 2008 to 2018, by gender (in years) is a area chart. The y-axis measures Men while the x-axis measures Year.

**Model Generated L2/L3 Caption**

The average age at widow hood in the Netherlands has increased from 2008 to 2018. The average age at widow hood in the Netherlands has increased from 2008 to 2018.

(a) The `scene-graph` model's output L1 caption and L2/L3 caption for a VisText area chart of the *"Average age at widow (er) hood in the Netherlands from 2008 to 2018, by gender (in years)"*. The model correctly identifies the chart's title and axis, and it accurately identifies the upward trend. However, it repeats this claim twice. See Section 5.2 for details on model repetition.

**Rasterized Image**




**Scene Graph**

```
title Comparable sales growth of Sam's Club in the United States from fiscal year 2006 to 2020 x -54.258 y -50 x-axis x 100 y 41.713 Year y-axis x -41.258 y 100 Comparable sales growth xtick x 0 val 2006 x 29 val 2008 x 57 val 2010 x 86 val 2012 x 114 val 2014 x 143 val 2016 x 171 val 2018 ytick y 200 val -0.04 y 169 val -0.02 y 138 val 0.00 y 108 val 0.02 y 77 val 0.04 y 46 val 0.06 y 15 val 0.08 marks area XY 0 61.538 H 76.923 desc XY 14.277 100 H 38.462 desc XY 28.555 63.077 H 75.385 desc XY ...
```

**Model Generated L1 Caption**

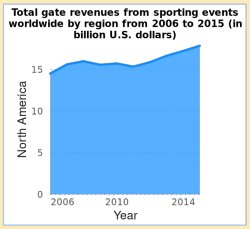
Here a area graph is called Comparable sales growth of Sam's Club in the United States from fiscal year 2006 to 2020. The y-axis measures Comparable sales growth while the x-axis measures Year.

**Model Generated L2/L3 Caption**

The comparable sales growth of Sam's Club in the United States from fiscal year 2006 to 2020 has been steadily decreasing from 2006 to 2020.

(b) The `scene-graph` model's output L1 caption and L2/L3 caption for a VisText are chart of the *"Comparable sales growth of Sam's Club in the United States from fiscal year 2006 to 2020"*. The model correctly identifies the chart's title and axis. However, it makes a stability error by stating the trend is steady when it is highly-fluctuating. See Section 5.2 for details on stability errors.

**Rasterized Image**




**Scene Graph**

```
title Total gate revenues from sporting events worldwide by region from 2006 to 2015 (in billion U.S. dollars) x 100 y -48.222 x-axis x 100 y 20 Year y-axis x -30 y 100 North America xtick x 0.012676199168441335 val 2006 x 44.430078085386874 val 2008 x 88.90832572761383 val 2010 x 133.32572761383227 val 2012 x 177.80397525605923 val 2014 ytick y 33.33333333333333 val 0 y 88.88888888888889 val 5 y 144.44444444444443 val 10 y 200 val 15 marks area XY 0 38.778 H 161.222 desc XY 22.209 ...
```

**Model Generated L1 Caption**

Total gate revenues from sporting events worldwide by region from 2006 to 2015 (in billion U.S. dollars) is a area chart. The y-axis measures North America while the x-axis measures Year.

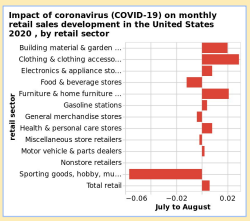
**Model Generated L2/L3 Caption**

The total gate revenue from sporting events worldwide by region from 2006 to 2015 has increased from around 15 billion dollars to around 15 billion dollars. The total gate revenue from sporting events worldwide by region from 2006 to 2015 has increased from around 15 billion dollars to around 15 billion dollars in 2015.

(c) The `scene-graph` model's output L1 caption and L2/L3 caption for a VisText area chart of the *"Total gate revenues from sporting events worldwide by region from 2006 to 2015 (in billion U.S. dollars)"*. The model correctly identifies the chart's title and axis. However, it makes a value error by claiming the revenue has increased to 15 billion dollars when it has actually increased to 18 billion dollars. See Section 5.2 for details on value errors.

Figure 4: `scene-graph` model captions of VisText area charts.

**Rasterized Image**




**Scene Graph**

```
title Impact of coronavirus (COVID-19) on monthly
retail sales development in the United States
2020 , by retail sector x -203 y -50 x-axis x 100
y 22 July to August y-axis x -190 y 130 retail
sector xtick x 0 val -0.06 x 44 val -0.04 x 89
val -0.02 x 133 val 0.00 x 178 val 0.02 ytick y
10 val Building material & garden equipment &
supplies dealers y 30 val Clothing & clothing
accessories stores y 50 val Electronics &
appliance stores y 70 val Food & beverage stores
y 90 val Furniture & home ...
```

**Model Generated L1 Caption**

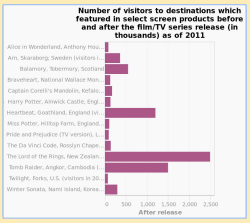
Impact of coronavirus (COVID-19) on monthly retail sales development in the United States 2020, by retail sector is a bar graph. The y-axis measures retail sector while the x-axis measures July to August.

**Model Generated L2/L3 Caption**

The total retail sector has been impacted on clothing and clothing accessories sectors the most. The lowest impact was in sporting goods.

(a) The scene-graph model's output L1 caption and L2/L3 caption for a VisText bar chart of the *"Impact of coronavirus (COVID-19) on monthly retail sales development in the United States 2020, by retail sector"*. The model correctly identifies the chart's title and axis, and it correctly identifies the the most and least impacted sectors.

**Rasterized Image**




**Scene Graph**

```
title Number of visitors to destinations which
featured in select screen products before and
after the film/TV series release (in thousands)
as of 2011 x 100 y -62 x-axis x 100 y 26 After
release y-axis x -185 y 140 Title/location/years
xtick x 0 val 0 x 38 val 500 x 77 val 1,000 x 115
val 1,500 x 154 val 2,000 x 192 val 2,500 ytick y
10 val Alice in Wonderland, Anthony House,
England (visitors in 2008 and 2011) y 30 val Arn,
Skaraborg; Sweden (visitors in 1999 and 2009) y
50 val Balamory ...
```

**Model Generated L1 Caption**

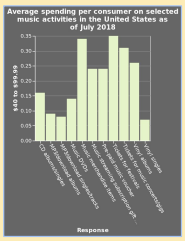
Here a bar graph is labeled Number of visitors to destinations which featured in select screen products before and after the film/TV series release (in thousands) as of 2011. The y-axis measures Title/location/years while the x-axis measures After release.

**Model Generated L2/L3 Caption**

The most popular visitors was Harry Potter in 1999 and 2009. The lowest visitors were Harry Potter in Alnwick Castle in England and Harry Potter in Alnwick Castle in England. The lowest visitors were Harry Potter in Alnwick Castle in England.

(b) The scene-graph model's output L1 caption and L2/L3 caption for a VisText are chart of the *"Number of visitors to destinations which featured in select screen products before and after the film/TV series release (in thousands) as of 2011"*. The model correctly identifies the chart's title and axis. However, it makes a nonsensical error by stating that most popular visitors was Harry Potter. See Section 5.2 for details on nonsensical errors.

**Rasterized Image**




**Scene Graph**

```
title Average spending per consumer on selected
music activities in the United States as of July
2018 x 110 y -52 x-axis x 110 y 165.054 Response
y-axis x -32 y 100 $40 to $99.99 xtick x 10 val
CD albums/singles x 30 val MP3/download albums x
50 val MP3/download singles/tracks x 70 val Music
DVDs x 90 val Music merchandise items x 110 val
Music streaming subscription gift cards x 130 val
Pre-paid music voucher x 150 val Tickets for
festivals x 170 val Tickets for music
concerts/gigs x 190 val  ...
```

**Model Generated L1 Caption**

Average spending per consumer on selected music activities in the United States as of July 2018 is a bar graph. The x-axis measures Response while the y-axis measures $40 to $99.99.

**Model Generated L2/L3 Caption**

The most popular music activity is vinyl albums and vinyl singles. The least popular music activity is vinyl albums.

(c) The scene-graph model's output L1 caption and L2/L3 caption for a VisText bar chart of the *"Average spending per consumer on selected music activities in the United States as of July 2018)"*. The model correctly identifies the chart's title and axis. However, it makes an identity error by incorrectly identifying the most popular music activity. See Section 5.2 for details on identity errors.

Figure 5: scene-graph model captions of VisText bar charts.

| Input | Level | BLEU ↑ | Perplexity ↓ | ROUGE-1 ↑ | ROUGE-2 ↑ | ROUGE-L ↑ | ROUGE-L SUM ↑ | WMD ↓ | TER ↓ |
|---|---|---|---|---|---|---|---|---|---|
| Kantharaj et al. (2022) | L1 | $0.43 \pm 1.33e{-3}$ | $67.58 \pm 2.56e{+0}$ | $0.73 \pm 3.00e{-4}$ | $0.61 \pm 1.00e{-3}$ | $0.63 \pm 1.33e{-3}$ | $0.63 \pm 1.40e{-3}$ | $0.58 \pm 2.67e{-3}$ | $53.48 \pm 2.19e{-2}$ |
| scene-graph | L1 | $0.43 \pm 4.67e{-3}$ | $61.01 \pm 3.41e{+0}$ | $0.74 \pm 2.70e{-3}$ | $0.61 \pm 6.80e{-3}$ | $0.63 \pm 5.67e{-4}$ | $0.63 \pm 5.33e{-4}$ | $0.57 \pm 1.72e{-2}$ | $53.05 \pm 2.60e{-1}$ |
| data-table | L1 | $0.42 \pm 2.73e{-3}$ | $70.68 \pm 4.33e{+0}$ | $0.74 \pm 5.40e{-3}$ | $0.62 \pm 6.50e{-3}$ | $0.63 \pm 2.70e{-3}$ | $0.63 \pm 2.97e{-3}$ | $0.57 \pm 1.33e{-2}$ | $53.27 \pm 4.48e{-1}$ |
| image | L1 | $0.09 \pm 2.20e{-3}$ | $70.14 \pm 4.24e{+0}$ | $0.40 \pm 2.50e{-3}$ | $0.20 \pm 3.47e{-3}$ | $0.30 \pm 2.50e{-3}$ | $0.30 \pm 2.47e{-3}$ | $1.09 \pm 8.33e{-3}$ | $83.69 \pm 3.25e{-1}$ |
| image-scene-graph | L1 | $0.43 \pm 2.35e{-3}$ | $55.66 \pm 7.58e{+0}$ | $0.74 \pm 1.65e{-3}$ | $0.62 \pm 1.10e{-3}$ | $0.63 \pm 1.50e{-3}$ | $0.63 \pm 1.50e{-3}$ | $0.58 \pm 6.05e{-3}$ | $53.22 \pm 1.27e{-1}$ |
| image-data-table | L1 | $0.43 \pm 1.53e{-3}$ | $64.57 \pm 1.05e{+0}$ | $0.74 \pm 1.07e{-3}$ | $0.62 \pm 1.43e{-3}$ | $0.63 \pm 2.20e{-3}$ | $0.63 \pm 2.27e{-3}$ | $0.57 \pm 4.47e{-3}$ | $53.47 \pm 6.81e{-2}$ |

(a) Model results using the L1 captions.

| Input | Level | BLEU ↑ | Perplexity ↓ | ROUGE-1 ↑ | ROUGE-2 ↑ | ROUGE-L ↑ | ROUGE-L SUM ↑ | WMD ↓ | TER ↓ |
|---|---|---|---|---|---|---|---|---|---|
| Kantharaj et al. (2022) | L2/L3 | $0.07 \pm 7.67e{-4}$ | $41.17 \pm 1.52e{+0}$ | $0.30 \pm 2.57e{-3}$ | $0.12 \pm 1.27e{-3}$ | $0.26 \pm 1.80e{-3}$ | $0.26 \pm 2.03e{-3}$ | $0.92 \pm 6.00e{-4}$ | $94.95 \pm 1.53e{+0}$ |
| scene-graph | L2/L3 | $0.07 \pm 8.07e{-3}$ | $18.81 \pm 3.74e{+0}$ | $0.28 \pm 1.65e{-2}$ | $0.11 \pm 9.43e{-3}$ | $0.25 \pm 1.02e{-2}$ | $0.24 \pm 1.02e{-2}$ | $0.92 \pm 8.90e{-3}$ | $120.62 \pm 6.72e{+0}$ |
| data-table | L2/L3 | $0.07 \pm 4.27e{-3}$ | $23.90 \pm 2.75e{+0}$ | $0.30 \pm 1.09e{-2}$ | $0.11 \pm 6.50e{-3}$ | $0.25 \pm 7.40e{-3}$ | $0.25 \pm 7.43e{-3}$ | $0.92 \pm 1.21e{-2}$ | $111.76 \pm 8.77e{+0}$ |
| image | L2/L3 | $0.02 \pm 2.73e{-3}$ | $7.64 \pm 7.19e{-1}$ | $0.17 \pm 1.22e{-2}$ | $0.03 \pm 3.83e{-3}$ | $0.14 \pm 1.02e{-2}$ | $0.14 \pm 1.02e{-2}$ | $1.19 \pm 7.40e{-3}$ | $148.95 \pm 1.79e{+1}$ |
| image-scene-graph | L2/L3 | $0.06 \pm 4.85e{-3}$ | $19.08 \pm 6.66e{-1}$ | $0.26 \pm 1.13e{-2}$ | $0.10 \pm 6.00e{-3}$ | $0.22 \pm 6.65e{-3}$ | $0.22 \pm 6.55e{-3}$ | $0.93 \pm 1.50e{-3}$ | $151.28 \pm 1.17e{+1}$ |
| image-data-table | L2/L3 | $0.06 \pm 5.13e{-3}$ | $19.02 \pm 1.79e{+0}$ | $0.27 \pm 5.00e{-3}$ | $0.11 \pm 3.23e{-3}$ | $0.23 \pm 3.47e{-3}$ | $0.23 \pm 3.53e{-3}$ | $0.92 \pm 2.20e{-3}$ | $144.20 \pm 6.11e{+0}$ |

(b) Model results using the L2/L3 captions.

Table 2: We separately evaluate our L1 and L2L3 captions on all the same metrics except for Relation Generation. In general, we observe that L1 captions perform better than the L2/L3 captions. Our models generate verbose L1 captions that are similar to the structure of our L1 templates, while the L2/L3 captions are human-generated and contain more variability. Darker colors indicate better scores.

# B Additional Evaluations

## B.1 Independent L1 and L2/L3 Caption Evaluation

To better understand how our models generate varying levels of semantic content, we separately evaluate our prefix-tuned models on L1 captioning and L2/L3 captioning tasks. Each prefix-tuned model can output an L1 or an L2/L3 caption for each chart. We evaluate these captions to their respective L1 or L2/L3 ground truth captions and report the results in Table 2.

Since we compute Relation Generation using only the L1 chart fields (e.g., chart title, axis scale, etc.), we do not report the results separately for L1 versus L2/L3 captioning. There is no direct Relation Generation analog for L2/L3 captions, since they are human-generated and do not follow a specific template. The Relation Generation for L1 captions is identical to the Relation Generation for L1/L2/L3 captions reported in Table 1.

## B.2 Evaluation Details

**Quantitative Model Performance Metrics.** We evaluate our models using NLP and machine translation metrics, including BLUE (Papineni et al., 2002; Lin and Och, 2004), Perplexity, Relation Generation (Wiseman et al., 2017), ROUGE (Lin, 2004), Word Mover's Distance (WMD), and Translation Edit Rate (TER) (Snover et al., 2006; Post, 2018). We implement Relation Generation per Wiseman et al. (2017), use the Gensim implementation of WMD, and use the Hugging Face implementation (Wolf et al., 2019) for the remaining metrics.

- *BLEU:* BLEU requires several gold standard references. In our evaluation setup, we use the test set caption as a single reference.
- *Perplexity:* We use a pretrained GPT-2 Medium model to compute Perplexity.
- *Relation Generation:* The fields we evaluate on are the chart title, axis names, and axis scales (if any).
- *Translation Edit Rate (TER):* Edits consist of deletions, additions, and substitutions, as present in SacreBLEU.

**Qualitative Caption Evaluation.** To produce our qualitative evaluation results (Sec. 5.2), we iteratively evaluated randomly sampled captions until there was no more marginal information about they types of errors to be gained from evaluating more captions. For each L2/L3 caption, we assess the number of

independent, mutually-exclusive L2 and L3 claims/statements that are being made. In comparison to evaluating at a sentence-level, this allows us to take a more nuanced approach that isn't limited by where the model has generated a full-stop. This approach allows us to more-accurately evaluate factual precision without overly-penalizing for a single mistake.

An example might take the form of "*The lowest value is X* **(claim 1)***, the highest value is Y* **(claim 2)***, and the second highest is Z* **(claim 3)***. Overall, it is increasing over time* **(claim 4)**." We observe that the first sentence is a compound sentence that consists of three independent clauses, each with a single factual L2 claim, while the second sentence is a single factual L3 claim. Let us assume that **claim 1** was factually incorrect. If we evaluate at a sentence-level, then the entire first sentence comprising of **claim 1**, **claim 2**, and **claim 3** would be incorrect. However, by breaking this caption into independent, mutually-exclusive claims, we can more precisely calculate the factual precision of our text generation.

| Experiment | Input | PT | BLEU ↑ | Perplexity ↓ | RG ↑ | ROUGE-1 ↑ | ROUGE-2 ↑ | ROUGE-L ↑ | ROUGE-L SUM ↑ | WMD ↓ | TER ↓ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Transformer Backbone | BART-base | | $0.27 \pm 5.03e{-3}$ | $43.06 \pm 5.76e{+0}$ | $1.69 \pm 9.80e{-3}$ | $0.59 \pm 9.67e{-4}$ | $0.43 \pm 4.73e{-3}$ | $0.48 \pm 2.07e{-3}$ | $0.48 \pm 2.00e{-3}$ | $0.65 \pm 1.47e{-3}$ | $67.08 \pm 1.50e{-1}$ |
| | T5-small | | $0.30 \pm 3.53e{-3}$ | $27.34 \pm 1.85e{+0}$ | $1.82 \pm 3.33e{-4}$ | $0.58 \pm 1.27e{-3}$ | $0.42 \pm 2.83e{-3}$ | $0.49 \pm 1.37e{-3}$ | $0.49 \pm 1.50e{-3}$ | $0.67 \pm 1.60e{-3}$ | $67.03 \pm 1.82e{-1}$ |
| | T5-small | ✓ | $0.30 \pm 5.40e{-3}$ | $29.52 \pm 1.35e{+0}$ | $1.15 \pm 7.70e{-2}$ | $0.52 \pm 1.06e{-2}$ | $0.35 \pm 8.77e{-3}$ | $0.43 \pm 7.83e{-3}$ | $0.43 \pm 7.90e{-3}$ | $0.71 \pm 1.22e{-2}$ | $76.44 \pm 1.28e{+0}$ |
| | Ours (ByT5-small) | | $0.32 \pm 4.07e{-3}$ | $20.96 \pm 3.09e{+0}$ | $1.82 \pm 2.67e{-4}$ | $0.56 \pm 1.42e{-2}$ | $0.39 \pm 1.62e{-2}$ | $0.47 \pm 1.04e{-2}$ | $0.47 \pm 1.04e{-2}$ | $0.68 \pm 8.33e{-3}$ | $69.34 \pm 2.31e{+0}$ |
| | Ours (ByT5-small) | ✓ | $0.32 \pm 2.13e{-3}$ | $20.02 \pm 2.25e{+0}$ | $1.78 \pm 4.25e{-2}$ | $0.56 \pm 6.70e{-3}$ | $0.39 \pm 6.23e{-3}$ | $0.47 \pm 6.37e{-3}$ | $0.47 \pm 6.40e{-3}$ | $0.68 \pm 1.23e{-2}$ | $72.55 \pm 1.75e{+0}$ |
| L1 Generation | new-seed | ✓ | $0.32 \pm 4.47e{-3}$ | $21.77 \pm 1.63e{+0}$ | $1.82 \pm 0.00e{+0}$ | $0.57 \pm 8.87e{-3}$ | $0.40 \pm 8.30e{-3}$ | $0.47 \pm 7.73e{-3}$ | $0.47 \pm 7.67e{-3}$ | $0.67 \pm 3.70e{-3}$ | $71.61 \pm 2.95e{+0}$ |
| | original-seed | ✓ | $0.32 \pm 2.13e{-3}$ | $20.02 \pm 2.25e{+0}$ | $1.78 \pm 4.25e{-2}$ | $0.56 \pm 6.70e{-3}$ | $0.39 \pm 6.23e{-3}$ | $0.47 \pm 6.37e{-3}$ | $0.47 \pm 6.40e{-3}$ | $0.68 \pm 1.23e{-2}$ | $72.55 \pm 1.75e{+0}$ |

(a) Ablation study results using the combined L1L2L3 captions.

| Experiment | Input | Level | BLEU ↑ | Perplexity ↓ | ROUGE-1 ↑ | ROUGE-2 ↑ | ROUGE-L ↑ | ROUGE-L SUM ↑ | WMD ↓ | TER ↓ |
|---|---|---|---|---|---|---|---|---|---|---|
| Transformer Backbone | T5-small | L1 | $0.42 \pm 7.87e{-3}$ | $73.01 \pm 5.20e{+0}$ | $0.64 \pm 1.64e{-2}$ | $0.52 \pm 1.44e{-2}$ | $0.56 \pm 1.10e{-2}$ | $0.56 \pm 1.09e{-2}$ | $0.65 \pm 1.06e{-2}$ | $62.76 \pm 1.10e{+0}$ |
| | Ours (ByT5-small) | L1 | $0.43 \pm 4.67e{-3}$ | $61.01 \pm 3.41e{+0}$ | $0.74 \pm 2.70e{-3}$ | $0.61 \pm 6.80e{-3}$ | $0.63 \pm 5.67e{-4}$ | $0.63 \pm 5.33e{-4}$ | $0.57 \pm 1.72e{-2}$ | $53.05 \pm 2.60e{-1}$ |
| L1 Generation | new-seed | L1 | $0.43 \pm 1.50e{-3}$ | $68.24 \pm 1.49e{+1}$ | $0.74 \pm 4.67e{-4}$ | $0.62 \pm 6.67e{-5}$ | $0.63 \pm 2.97e{-3}$ | $0.63 \pm 3.07e{-3}$ | $0.56 \pm 1.19e{-2}$ | $53.17 \pm 1.72e{-1}$ |
| | original-seed | L1 | $0.43 \pm 4.67e{-3}$ | $61.01 \pm 3.41e{+0}$ | $0.74 \pm 2.70e{-3}$ | $0.61 \pm 6.80e{-3}$ | $0.63 \pm 5.67e{-4}$ | $0.63 \pm 5.33e{-4}$ | $0.57 \pm 1.72e{-2}$ | $53.05 \pm 2.60e{-1}$ |

(b) Ablation study results using the L1 captions.

| Experiment | Input | Level | BLEU ↑ | Perplexity ↓ | ROUGE-1 ↑ | ROUGE-2 ↑ | ROUGE-L ↑ | ROUGE-L SUM ↑ | WMD ↓ | TER ↓ |
|---|---|---|---|---|---|---|---|---|---|---|
| Transformer Backbone | T5-small | L2/L3 | $0.06 \pm 2.67e{-3}$ | $35.81 \pm 4.13e{+0}$ | $0.25 \pm 6.43e{-3}$ | $0.09 \pm 3.43e{-3}$ | $0.22 \pm 5.73e{-3}$ | $0.22 \pm 5.60e{-3}$ | $0.99 \pm 8.70e{-3}$ | $113.33 \pm 2.94e{+0}$ |
| | Ours (ByT5-small) | L2/L3 | $0.07 \pm 8.07e{-3}$ | $18.81 \pm 3.74e{+0}$ | $0.28 \pm 1.65e{-2}$ | $0.11 \pm 9.43e{-3}$ | $0.25 \pm 1.02e{-2}$ | $0.244 \pm 1.02e{-2}$ | $0.92 \pm 8.90e{-3}$ | $120.62 \pm 6.72e{+0}$ |
| L1 Generation | new-seed | L2/L3 | $0.08 \pm 5.93e{-3}$ | $20.96 \pm 2.71e{+0}$ | $0.29 \pm 5.77e{-3}$ | $0.11 \pm 2.33e{-3}$ | $0.25 \pm 5.30e{-3}$ | $0.25 \pm 5.27e{-3}$ | $0.91 \pm 1.83e{-3}$ | $116.36 \pm 1.11e{+1}$ |
| | original-seed | L2/L3 | $0.07 \pm 8.07e{-3}$ | $18.81 \pm 3.74e{+0}$ | $0.28 \pm 1.65e{-2}$ | $0.11 \pm 9.43e{-3}$ | $0.25 \pm 1.02e{-2}$ | $0.244 \pm 1.02e{-2}$ | $0.92 \pm 8.90e{-3}$ | $120.62 \pm 6.72e{+0}$ |

(c) Ablation study results using the L2/L3 captions.

Table 3: We perform two ablation studies to measure the impact of our model architectures and L1 caption generation. Our Transformer Backbone ablation study compares our ByT5-small backbone to T5-small with and without prefix-tuning (PT) and BART-base. Our L1 Generation ablation study analyzes our stochastic L1 caption generation pipeline with different random seeds. We evaluate each model using machine translation and text generation metrics: BLEU, Perplexity, Relation Generation (RG), ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-L SUM, Word Mover's Distance (WMD), and Translational Error Rate (TER). Darker colors indicate better scores.

## C Ablation Studies

To evaluate our modeling and dataset design choices, we run ablation studies measuring the impact of our transformer model backbones and stochastic data generation pipeline. We report the results in Table 3.

**Transformer Backbone.** To understand the impact of our token-free, byte-to-byte architecture ByT5 model backbone, we explore other large language models. Specifically, we compare our 300M parameter ByT5-small model (Xue et al., 2022) with a 60M parameter T5-small (Raffel et al., 2020) and 140M parameter BART-base model (Lewis et al., 2020). We also apply prefix-tuning to the ByT5 and T5 models. We cannot apply prefix-uning to BART because BART does not support multi-task learning. Quantitatively, using ByT5 does not appear to significantly improve upon T5. However, we theorize that ByT5's token-free paradigm increases the input sequence length by compressing more input text into fewer input tokens.

**L1 Caption Generation.** Since we generate L1 captions stochastically, we evaluate whether our initial randomization impacted the model's results. We compare generate a second set of L1 captions using a different random seed. We see the results are nearly identical across all metrics, indicating our dataset captures a diverse set of L1 captions.

| Model | PT | Seeds | Epochs | Batch Size | Optim. | Adam $\beta 1$ | Adam $\beta 2$ | Adam $\epsilon$ | Weight Decay | LR |
|---|---|---|---|---|---|---|---|---|---|---|
| Kantharaj et al. (2022) | | 10, 20, 30, 40, 50 | 50 | 2 | AdamW | 0.9 | 0.999 | $1e-08$ | Linear | $5e-05$ |
| Kantharaj et al. (2022) | ✓ | 10, 20, 30, 40, 50 | 50 | 3 | AdamW | 0.9 | 0.999 | $1e-08$ | Linear | $5e-05$ |
| `scene-graph` | ✓ | 10, 20, 30, 40, 50 | 50 | 3 | AdamW | 0.9 | 0.999 | $1e-08$ | Linear | $5e-05$ |
| `data-table` | ✓ | 10, 20, 30, 40, 50 | 50 | 4 | AdamW | 0.9 | 0.999 | $1e-08$ | Linear | $5e-05$ |
| `scene-graph` | | 10, 20, 30, 40, 50 | 50 | 3 | AdamW | 0.9 | 0.999 | $1e-08$ | Linear | $5e-05$ |
| `data-table` | | 10, 20, 30, 40, 50 | 50 | 4 | AdamW | 0.9 | 0.999 | $1e-08$ | Linear | $5e-05$ |
| `image-scene-graph` | | 9555, 16710, 23578 | 50 | 4 | AdamW | 0.9 | 0.999 | $1e-06$ | 0.01 | $5e-05$ |
| `image-scene-graph` | ✓ | 1393, 16983, 23814 | 50 | 4 | AdamW | 0.9 | 0.999 | $1e-06$ | 0.01 | $5e-05$ |
| `image-data-table` | | 7504, 9586, 32579 | 50 | 4 | AdamW | 0.9 | 0.999 | $1e-06$ | 0.01 | $5e-05$ |
| `image-data-table` | ✓ | 4120, 7625, 19179 | 50 | 4 | AdamW | 0.9 | 0.999 | $1e-06$ | 0.01 | $5e-05$ |
| `image` | | 13423, 17963, 29028 | 50 | 32 | AdamW | 0.9 | 0.999 | $1e-06$ | 0.01 | $5e-05$ |
| `image` | ✓ | 4650, 7434, 15249 | 50 | 32 | AdamW | 0.9 | 0.999 | $1e-06$ | 0.01 | $5e-05$ |
| `BART-base scene-graph` | | 10, 20, 30, 40, 50 | 50 | 2 | AdamW | 0.9 | 0.999 | $1e-08$ | Linear | $5e-05$ |
| `T5-small scene-graph` | | 10, 20, 30, 40, 50 | 50 | 2 | AdamW | 0.9 | 0.999 | $1e-08$ | Linear | $5e-05$ |
| `T5-small scene-graph` | ✓ | 10, 20, 30, 40, 50 | 50 | 3 | AdamW | 0.9 | 0.999 | $1e-08$ | Linear | $5e-05$ |
| `new-seed scene-graph` | ✓ | 10, 20, 30, 40, 50 | 50 | 3 | AdamW | 0.9 | 0.999 | $1e-08$ | Linear | $5e-05$ |

Table 4: A summary of notable hyperparameters we used to train the baseline, text-based, image-guided, and ablation study models. For all parameters and code, see: https://github.com/mitvis/vistext.

# D Implementation Details

Code to train and evaluate our text-based and image-guided models is available at https://github.com/mitvis/vistext. Table 4 summarizes our model training parameters.

## D.1 Text-Based Chart Captioning

To train our text-based chart captioning models, we use the Huggingface implementation of ByT5 (Wolf et al., 2019). Due to hardware limitations, we use the ByT5-small model, which has 300M parameters. We fine-tune each model for 50 epochs, using Adam optimization with a learning rate of $5e{-}05$. To fit the input features into GPU memory, we truncate the input text (i.e., scene graph or data table) to 1024 tokens and the output caption to 512 tokens. We select the best model epochs based on the validation loss of the validation set. See Table 4 or the VisText GitHub repository[1] for each model's full training details and hyperparameters.

We train each model three times with and without prefix-tuning and report the mean and standard deviation in Table 1. We train each model on four NVidia V100 GPUs with 32GB of memory connected by an NVLink2 network. With prefix-tuning, training, evaluation, and inference took approximately 39 hours for the `scene-graph` model and 11 hours for the `data-table` models. Without prefix-tuning, training, evaluation, and inference took approximately 78 hours for the `scene-graph` model and 22 hours for the `data-table` models. We estimate that we trained each model between 30 to 45 times to achieve our final results.

## D.2 Image-Guided Chart Captioning

Our image-guided chart captioning models extend the VLT5 model (Cho et al., 2021), which is a multimodal extension of T5-base. We extract visual features from VisText's chart images using Bottom-Up Feature Extraction (Anderson et al., 2018) and 36 bounding boxes per image. After feature extraction, we fine-tune VLT5 on the VisText dataset for 50 epochs following the default VLT5 training protocol[2] (Cho et al., 2021). To fit the input features into GPU memory, we truncate the input text (i.e., scene graph or data table) to 1024 tokens and the output caption to 512 tokens. After 50 epochs, we select the epoch with the lowest validation loss as the best model. See Table 4 or the VisText GitHub repository[2] for each model's full training details and hyperparameters.

We train each model three times with and without prefix-tuning and report the mean and standard deviation in Table 1. We train each model on four NVidia V100 GPUs with 1TB of memory. The `image` models take approximately 2 minutes per training epoch without prefix-tuning and approximately 3

[1] https://github.com/mitvis/vistext

[2] https://github.com/j-min/VL-T5

| Package | Version |
|---|---|
| bleurt | 0.0.2 |
| datasets | 2.10.1 |
| evaluate | 0.4.0 |
| gensim | 4.3.0 |
| h5py | 3.7.0 |
| nltk | 3.7 |
| numpy | 1.22.3 |
| pandas | 1.4.2 |
| pot | 0.9.0 |
| pytorch | 1.10.2 |
| pyyaml | 5.4.1 |
| sacrebleu | 2.0.0 |
| scipy | 1.7.3 |
| sentencepiece | 0.1.95 |
| tokenizers | 0.11.4 |
| torchvision | 0.13.1 |
| transformers | 4.24.0 |

Table 5: The package versions used for training and evaluating our VisText models. Further implementation details and code are available at: https://github.com/mitvis/vistext.

minutes per training epoch with prefix-tuning. The `image-scene-graph` and `image-data-table` models take approximately 10 minutes per training epoch without prefix-tuning and approximately 15 minutes per training epoch with prefix-tuning. We estimate that we trained each model between 5 to 10 times to achieved our final results.

### D.3 Ablation Models

We train our ablation models using the same parameters as our default models, only varying the parameter of interest. We train them on 16 virtual CPU cores on Xeon E5 hypervisors with 128GB of memory and PCI pass-through access to eight NVidia Titan XP GPUs with 12GB of memory.

### D.4 Notable Package Versions

Package versions are listed in Table 5.

## E Additional VisText Dataset Details

### E.1 Licensing

Our use of the raw Statista data from Kantharaj et al. (2022) is consistent with its intended use case. The data was licensed under the GNU General Public License v3.0. We release our data and code under GNU General Public License v3.0.

### E.2 L1 Caption Generation Process

The Level 1 captions are generated from a random process that chooses from 3 title templates and 6 axis templates. The title templates we use are:

- This is a `[chart-type]` titled `[chart-title]`
- This `[chart-type]` is titled `[chart-title]`
- `[chart-title]` is a `[chart-type]`

The axis templates we use for each axis are:

- On the `[axis]`, `[axis-label]` is plotted with a `[axis-scale]`
- `[axis-label]` is plotted with a `[axis-scale]` on the `[axis]`
- The `[axis]` plots `[axis-label]` with a `[axis-scale]`
- A `[axis-scale]` can be found on the `[axis]`, labeled `[axis-label]`
- This is a `[axis-scale]` on the `[axis]`, labeled `[axis-label]`

We additionally have one template for both axes:

- The `[axis1]` plots `[axis1-label]` with `[axis1-scale]` while the `[axis2]` plots `[axis2-label]` with `[axis2-scale]`

For each axis template, we randomly choose whether to include the axis scale. Furthermore, within each template, we further randomly swap words with synonyms. A list of words and their possible synonym substitutions are:

- **this:** *here, a*
- **chart:** *graph, diagram, plot*
- **titled:** *called, named, labeled*
- **on:** *along*
- **plotted**: *defined, measured, drawn, shown*
- **plots:** *measures, shows*
- **with:** *using, on, along, as*
- **found:** *seen*
- **labeled:** *marked*

### E.3 Crowdsourced Study Protocol

Figures 6–10 screenshot the introduction, eligibility and consent statements, instructions, and a task from our crowdsourced study. We recruited participants on the Prolific.co crowdsourcing platform, following conventions in the data visualization research community[3] and recent research results (Tang et al., 2022) that suggest Prolific yields higher quality results than Amazon Mechanical Turk. We conducted multiple pilot runs to calibrate the amount of time it would take participants to complete the study, and found that most participants were able to successfully do so within 14 minutes. Following Silberman et al. (2018), who advocate for paying workers at least minimum wage *at your location*, we choose to pay our participants $3.25 — a roughly $14/hour rate in line with the $14.25/hour minimum wage in Massachusetts at the time the study was conducted.

Our study was determined to be exempt by MIT's institutional review board (IRB). Participants had to explicitly provide their consent in order to proceed with the study — if participants did not consent, they were redirected back to the Prolific platform. The consent statement (Fig. 8) reminded participants of their rights (including that their participation is voluntary and consent could be revoked at any time), and encouraged participants to contact either the study PI or IRB board directly should they have any concerns. We constrained our participant pool (and eligibility requirements) to people living within the United States or United Kingdom who self-reported as being sighted with no vision or color impairments. We did not collect any additional demographic data from participants as we did not determine this to bias or otherwise affect the content we hoped to collect.

Each task (an example of which is shown in Fig. 10) included an attention check where participants were asked to correctly identify the chart type shown. If participants failed more than two attention checks, their submission was flagged for manual review — in practice, the bulk of participants who failed attention checks nevertheless produced valid captions and, thus, were paid fully. The task asked participants to complete a free response question to describe as completely as they could the trends and patterns observed, emphasizing that their response would be evaluated for correctness and completeness. Despite best practices suggesting a more structured, querying approach (called QID) can yield higher quality captions (Morash et al., 2015), we opted for our free-response approach as the benefits of QID (namely, in expressing the chart type, title, and axes units) would already be captured by our synthetically generated L1 captions. Moreover, in contrast to the templatized output produced by QID, we hoped that our free-response responses would yield more "natural" articulations of perceptual and cognitive trends, following the Lundgard and Satyanarayan (2022) framework.

[3] https://twitter.com/eagereyes/status/1187773534745088000

## Describing Data Visualizations

You are invited to participate in a research study about describing data visualizations to improve their accessibility to people who are blind or have low-vision. You have been asked to participate because we need to determine what language sighted readers use to describe charts, and the sorts of trends and patterns you identify when reading a chart. We will ask you to read data visualizations (such as a bar chart or line chart), and to also read a textual description of each visualization. Then, we will ask you to answer questions about each visualization, and to write your responses in English.

Full completion of this study consists of the following:

1. Reading through an introductory example task.
2. Completing up to 5 visualization description tasks, with 1 open-ended question per task (5 text input questions total).

The estimated time to complete this study is 14 minutes.

Figure 6: Introduction to the crowdsourcing study.

## Eligibility Statement

You are eligible for this study if and only if:

1. You are sighted and do not have color blindness or other color vision impairment.
2. You are comfortable reading and writing in English.
3. You are over 18 years of age.
4. You are comfortable with interpreting charts and graphics.
5. You have JavaScript enabled in your browser.

○ I have read the above information and attest that I AM eligible to participate in this study.

○ I have read the above information and attest that I AM NOT eligible to participate in this study.

Figure 7: Eligibility statement for the crowdsourcing study.

## Consent Statement

Please review the following information about this study.

- If you agree to participate in this study, we will ask you to complete a sequence of visualization description tasks.
- We expect this study to take 14 minutes, and you will receive $3.25 as compensation.
- We don't anticipate any risks from participating in this study to be greater than normal activity.
- There will be no direct benefits to you other than compensation.
- No personally identifying information will be revealed to anyone other than the researchers conducting this study.
- The records of this study will be kept private. In any sort of report we make public, we will not include any information that will make it possible to identify you.
- Research records will be kept on encrypted serves. Only the researchers will have access to the records.
- Participating in this study is completely voluntary. If you decline to participate, it will not affect your current or future relationship with the researchers.
- If you agree to participate, you are free to withdraw at any time. We will delete any data related to your participation upon your request.
- You may ask any questions about the research at any time. If you have questions about the research after you leave today you should contact [redacted].
- If you are not satisfied with response of research team, have more questions, or want to talk with someone about your rights as a research participant, you should contact the [redacted].

○ I have read the above information and AGREE to participate in this study.

○ I have read the above information and DO NOT AGREE to participate in this study.

Figure 8: Consent statement for the crowdsourcing study.

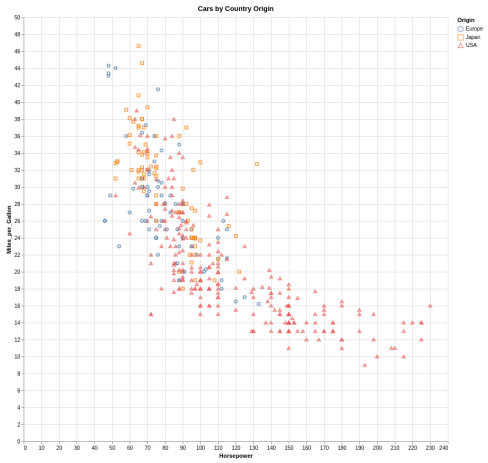
## Example Task

In this task, we first present a visualization, followed by a textual description of that visualization. Then, we ask you a question, and provide an example response.

This question the same as the ones you will be asked in the upcoming tasks. Note that the responses are open-ended text input. We ask that you try your best and respond in complete sentences.

Please confirm that you have read and understood each example question by clicking the button below it.

### Example Visualization




### Example Visualization Description

A scatterplot entitled “Cars by Country Origin” that plots cars’ Horsepower versus Miles_per_Gallon by Origin. Car Origin includes Europe, Japan, and USA, encoded by Color and Shape: Europe (blue, circle), Japan (orange, square), and USA (red, triangle). Horsepower is plotted on the horizontal x-axis from 0 to 250 with an increment of 50. Miles_per_Gallon is plotted on the vertical y-axis from 0 to 50 with an increment of 10.

### Example Question

Besides what was already said in the above description, what other conclusions can you draw from this visualization? **What trends or patterns can you observe?**

Please only articulate trends or patterns using information contained in the visualization. Please do not introduce background knowledge or assumptions about the data apart form what is represented in the visualization. Please state as many observations as you are able. **Please state each observation using one complete sentence. Your response will be evaluated for its correctness and completeness.**

### Example Responses

| |
|---|
| Japan produces the car with the greatest Miles_per_Gallon. USA produces the car with the fe Miles_per_Gallon. Most cars have around 40 to 120 Horsepower. Most of the cars can drive around 20 to 40 Miles_per_Gallon. |

**These sentences are good because they accurately reference information contained in the chart.**

| |
|---|
| Overall, as horsepower increases, miles per gallon decreases. USA cars have the highest horsepower. European and Japanese cars seem to have better miles per gallon. |

**These sentences are also good because they describe a trend in the data presented by the chart using only the information provided by the chart.**

### *Incorrect* Responses

Certain sentences should not be written. For example:

- *"The cars have between 0 Miles_per_Gallon and 50 Miles_per_Gallon."*
  - **While this is an accurate description of the chart, it should not be written because a similar description was already given in the above Example Description.**
- *"European cars are shown as blue circles, Japanese cars are orange squares, and American cars are read triangles."*
  - **This is a bad response because it focuses on the visual appearance of the chart rather than describing trends or patterns.**
- *"Compared to Europe and Japan, the USA does a poor job of manufacturing environmentally efficient cars. Japanese cars have the highest miles per gallon, and are therefore most environmentally friendly."*
  - **This is a bad response because it introduces background knowledge about environmental efficiency and friendliness that is not contained within the chart itself.**

○ I have read and understood the Example Question.

This concludes the Example Task. You will now be asked to complete up to 5 similar visualization description tasks.

○ I have read and understood the Example Task.

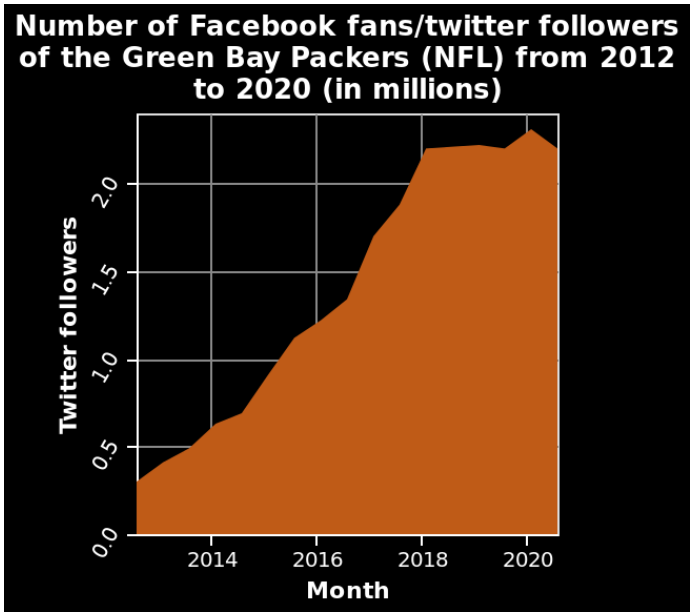
Figure 9: Instructions for the crowdsourcing study, with examples of correct and incorrect responses.

**Visualization Description Task (./imgs_47/multiColumn/data/3262_1.png)**

## Task

Directions: Please read the data visualization and the corresponding textual description. Then, please provide a response to the text-input question.

### Visualization




What type of visualization is shown above?

- ○ Scatter Plot
- ○ Line/Area Chart
- ○ Bar Chart

### Visualization Description

Number of Facebook fans/twitter followers of the Green Bay Packers (NFL) from 2012 to 2020 (in millions) is a area diagram. A linear scale from 0.00 to 2.00 can be found on the y-axis, marked Twitter followers. There is a linear scale from 2014 to 2020 on the x-axis, marked Month.

### Question

Besides what was already said in the above description, what other conclusions can you draw from this visualization? **What trends or patterns can you observe?**

Please only articulate trends or patterns using information contained in the visualization. Please do not introduce background knowledge or assumptions about the data apart form what is represented in the visualization. Please state as many observations as you are able. **Please state each observation using one complete sentence. Your response will be evaluated for its correctness and completeness.**

Figure 10: An example task from a specific run of the crowdsourcing study.