Alt-Tag: Enhancing Alt Text Generation for Charts via Tag Guidance

Anonymous ACL submission

Abstract

Generating alternative (alt) text for scientific charts is essential for assisting screen reader users, most often blind persons and persons with visual impairments, in accessing chart content. Yet producing structured and informative alt text remains challenging for vision-language models (VLMs). In this work, we investigate the use of structural tags as intermediate representation to guide alt text generation for charts. We developed a structural tag schema for three chart types (bar, line, and pie chart). We performed alt text generation on two datasets across five tag injection strategies. We also extended the evaluation framework used in previous alt text generation research by examining lexical overlap, semantic similarity, image-text alignment, structural clarity, and information concentration in the output. Our results demonstrate that using tags as guidance can improve alt text generation performance, thereby also advocating the use of such structural tags in future human alt text description practice.

1 Introduction

Alternative text (alt text) of charts is critical for blind persons and persons with visual impairments who rely on screen reader software that transforms textual information into synthesized speech (Jung et al., 2021). Automatically generating alt text has matured into a task at the interface of computer vision and natural language processing due to recent advancements of vision-language models (VLMs) (Xu et al., 2024; Yan et al., 2025; Hu et al., 2026). As this task is closely related to chart summarization and image captioning, existing studies often rely on chart captions as ground truth data (Hsu et al., 2021; Singh et al., 2023; Liu et al., 2023; Masry et al., 2023; Han et al., 2023; Liu et al., 2024; Meng et al., 2024; Zhang et al., 2024; Huang et al., 2025). However, alt texts fundamentally differ from captions in that the former are meant to serve as a *substitutes* rather than as *complements* for visual content (Jung et al., 2021).

Previous research has investigated how to effectively present alt text information for scientific charts. While no single set of guidelines exists, Yan et al. (2025) find in their survey that most guidelines recommend organizing alt text in a hierarchical structure, beginning with information on basic visual elements and subsequently providing more detailed information on prominent data points or notable patterns. Several datasets have been developed based on such guidelines, such as VisText (Tang and Boggust, 2023), the HCI Dataset (Chintalapati et al., 2022), and Alt4Blind (Moured et al., 2024a). Following these guidelines, basic visual information is straightforward to describe, while information at more detailed levels is more challenging to present, as it requires richer contextual interpretation. Furthermore, the chart type may influence both the emphasized content and the descriptive strategies. Therefore, existing studies have not addressed this more detailed information of alt text for charts to a sufficient extent.

In this work, we extend current alt text generation guidelines by introducing a tag schema for detailed chart information for three common chart types: simple bar chart, simple line chart, and pie chart. In this context, simple refers to chart images having been created from data of a single series. We performed alt text generation with and without this schema, using different tag injection strategies, on two datasets. Additionally, we extended existing evaluation frameworks by incorporating additional metrics (Hsu et al., 2021; Singh et al., 2023; Liu et al., 2023; Masry et al., 2023; Han et al., 2023; Liu et al., 2024; Meng et al., 2024; Zhang et al., 2024; Huang et al., 2025). Our results show an improvement in alt text generation performance resulting from incorporating tags as guidance.

Our main contributions are as follows:

- We propose a structural tag schema across three chart types to guide detailed alt text generation for charts.
- We extend a popular existing dataset, VisText (Tang and Boggust, 2023), with the proposed tags and construct a new dataset, Chart-Alt, containing real chart images (as opposed to the synthetic images used in VisText) and tagged alt text.
- We extend existing evaluation frameworks for alt text generation by adding metrics that specifically assess structural clarity and information concentration.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 introduces the tag schema, the dataset annotation process, fine-tuning strategies, and evaluation metrics. Section 4 presents the results of the alt text generation task on two datasets with structural tag information. Section 5 discusses the findings, and Section 6 concludes by outlining limitations and future directions.

2 Related work

2.1 Alt text generation guidelines

Previous works have investigated the composition of alt text for charts for screen reader users. Zong et al. (2022) and Jung et al. (2021) suggested that alt text should follow a hierarchical structure progressing from general content to specific details. Lundgard and Satyanarayan (2021) proposed a four-level (L1–L4) theoretical semantic model, which has become the most widely adopted model in chart accessibility research. In this model, the L1 level describes the basic visual elements of a chart, the L2 and L3 levels capture statistics and complex patterns, and the L4 level focuses on the domain or social background of the chart content (see an example in Table 6 in Appendix A). The first three levels have been shown to be the most crucial for blind persons and persons with visual impairments when accessing chart content, with L4 level considered largely optional (Lundgard and Satyanarayan, 2021).

While generating L1-level content is comparatively straightforward, the above studies provide limited guidance on how to logically structure L2- and L3-level content, specifically regarding *which information to present and in what order*. As a result, generating L2- and L3-level alt text remains challenging. For example, results reported from the VisText dataset (Tang and Boggust, 2023) showed that L1-level alt text generation achieved a BLEU-4 score (see Section 2.3 for an overview of common evaluation metrics) of 0.43, whereas L2L3-level (combined in the VisText dataset) generation achieved only 0.07 BLEU-4. Similarly, Moured et al. (2024b) reported this generation gap in their work, where L1-level generation achieved 37.6% accuracy, whereas L2L3-level generation achieved only 5.6% accuracy.

2.2 Structural representation in the chart summarization task

Intermediate representations have been explored as a potential way to provide additional structured information for chart *summarization* task. Choi and Jo (2022) introduced recipe templates that provide high-level guidance for organizing summaries into components such as overview, description, comparison, and trend. Similarly, Tang and Boggust (2023) explored the use of scene graphs as an alternative structured representation of visual input, demonstrating improvements over approaches without scene graph representations. To our knowledge, no prior approach has investigated providing an intermediate structural representation for generating L2L3-level information of alt text, which together with the L1 level constitutes a complete alt text.

2.3 Evaluation metrics

Yan et al. (2025) in their survey of alt text generation found that the most commonly used evaluation metrics are BLEU-4 (Papineni et al., 2002) and ROUGE_L (Lin, 2004). Other metrics, such as Translation Edit Rate (TER) (Snover et al., 2006) and Word Mover's Distance (WMD) (Kusner et al., 2015), have also been applied (Tang and Boggust, 2023). Together, they make up the class of reference-based metrics.

In contrast, reference-free metrics directly capture the relation between the input image and a generated alt text. For example, CLIPScore (Hessel et al., 2021) measures semantic similarity between an image and its caption by calculating the cosine similarity between the image and text embeddings in a shared space learned by the CLIP-ViT-B-32 model [1]. ChartVE (Huang et al., 2024) is designed

[1] https://huggingface.co/sentence-transformers/clip-ViT-B-32

specifically for chart images and is built by fine-tuning UniChart VLM (Masry et al., 2023) to measure factual consistency between a chart image and its corresponding description.

Several studies additionally employ human evaluation, which comes at the expense of time and cost (Obeid and Hoque, 2020; Masry et al., 2023; Liu et al., 2024). The recent rise of employing LLM/VLM as judges has emerged as an alternative (Li et al., 2025; Pu et al., 2025; Szymanski et al., 2025; Laskar et al., 2025). Among these previous works, Laskar et al. (2025) demonstrated that small and open-source VLMs can serve as reliable evaluators for the chart *comprehension* task. To our knowledge, our work is the first to leverage the LLM-/VLM-as-a-judge paradigm for the alt text generation task.

3 Tag-guided alt text generation

This section presents the methods underlying the alt text generation experiments reported on in this paper. The procedure for creating the structural tag schema is explained in Section 3.1. Section 3.2 describes the annotation and creation of two datasets. Section 3.3 presents the process of fine-tuning a chart-specific VLM, ChartGemma (Masry et al., 2024), using five tag injection strategies. Lastly, Section 3.4 details the evaluation metrics used.

3.1 Tag schema

As outlined in Section 2.1, a complete alt text consists of information across three levels (L1-L3) (Lundgard and Satyanarayan, 2021; Tang and Boggust, 2023). L1-level alt text captures the fundamental visual elements of a chart: the type of chart, the title of the chart, labels of the horizontal and vertical axes, and the value ranges of both axes. L1-level alt text is included in the VisText dataset (Tang and Boggust, 2023). We also incorporated it into our own dataset (see Section 3.2.2).

As previously mentioned, L2L3-level alt text is more difficult to operationalize. As a result, no prior work has focused on building a tag schema that guides the structure of this part of an overall alt text, resulting in lower generation performance for L2L3-level information compared to L1-level information (see Section 1). To address this gap, we performed a clustering analysis on the widely used Statista dataset (Kanthara et al., 2022) (see Table 7 in Appendix B for examples). This dataset contains 34,811 real chart images and their corresponding captions extracted from the Statista website [2], covering four representations: bar chart, line chart, pie chart, and table. Our analysis focused on the three chart types (bar chart, line chart, and pie chart). The 500 most frequent words of these three chart types were extracted from the chart descriptions and subsequently clustered using the HDBSCAN algorithm (McInnes et al., 2017), a method that identifies clusters based on word embedding densities without requiring the definition of a number of clusters.

The process produced 56 clusters for bar charts, 54 clusters for line charts, and 46 clusters for pie charts. These clusters were further refined based on characteristics of the different chart types by Hardin et al. (2012): Bar charts are primarily used to *compare* different categories, revealing the high and low values in data, line charts are used to visualize *dynamic* changes and trends, and pie charts are often used to visualize the *proportions* of each segment. Based on these principles, less relevant clusters were removed, including clusters containing stop words, function words, and contextspecific words (e.g., country names, dates and years, health-related terms, industry-specific words, and geographical terms). As a result, eight clusters remained for bar charts, twelve clusters remained for line charts, and seven clusters remained for pie charts.

Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) was applied to visualize the results (see Figure 1 in Appendix C). More details on the clustering process are provided in Appendix C.

For L1-level tags, we adopted the same schema as used in the VisText dataset (Tang and Boggust, 2023) (see Table 1). For the L2 and L3 levels, we created the tags from the clustering results.

Across the three chart types of interest, we identified three common cluster groups: The first contained expressions of extreme values, including terms such as “highest”, “biggest”, “largest”, and “lowest”. The second group captured trend-related terms, such as “increase”, “up”, “decrease”, and “down”. The third group was associated with quantitative expressions such as “many”, “more”, “roughly”, “approximately”, and “estimated”.

In addition to these shared patterns, we identified several distinct clusters across the three chart types. For line charts, one group reflects prepo-

[2] https://www.statista.com

sitional and temporal expressions, including "in", "during", "within", "throughout", and "into", which are used to describe observations over entire or partial intervals. For pie charts, we observe a focus on proportional and fractional descriptions used to express part-whole relationships, such as "percentage", "percent", "half", and "quarter". The final tag schema is shown in Table 1.

3.2 Dataset creation and annotation

Two datasets were used for the experiments reported on in this paper: the existing VisText (Tang and Boggust, 2023) and our newly proposed Chart-Alt dataset. Data samples from our Chart-Alt dataset are shown in Table 9 in Appendix E. The dataset annotation process is described in the following subsections. Table 2 summarizes information about the two datasets.

3.2.1 VisText

The VisText dataset (Tang and Boggust, 2023) is the first large-scale dataset that incorporates multiple levels (L1–L3 according to the four-level semantic model by Lundgard and Satyanarayan (2021)) of alt text in chart accessibility research. It contains more than 8k synthetic chart images (in contrast to our proposed Chart-Alt dataset, which consists of real images), and covers three chart types: simple bar chart, simple line chart and simple area chart. Since simple area chart is not covered in the Statista dataset (Kanthara et al., 2022) (see Section 3.1), no corresponding tag schema was developed for this chart type; therefore, area charts were excluded from our experiments.

To minimize errors and bias introduced by a single model, two large language models (LLMs) were employed to automatically insert tags following the scheme introduced in Section 3.1 into the VisText dataset: Qwen2.5-3B-Instruct[3] and Gemma-1.1-2b-it[4]. The prompts used for tag insertion are shown in Table 12 and Table 13 in Appendix G. Only samples with mutually consistent tagging outputs from both models were retained. As a result, 1,870 alt text were included in the final dataset, which were subsequently split into train (80%), test (10%), and validation (10%) sets.

3.2.2 Chart-Alt

Since the VisText dataset (Tang and Boggust, 2023) only includes synthetic chart images, we manually constructed our own dataset, Chart-Alt. This dataset includes 600 real chart images extracted from three digital libraries: arXiv[5], Springer[6], and ACM[7]. All chart images were extracted from publications between January 1, 2025 and May 1, 2025. The extracted images were manually checked to verify that there were no duplicates, that they were self-contained and interpretable without additional context, and that they belonged to one of three targeted chart types: simple bar, simple line, or pie chart. The resulting dataset consists of 200 images per chart type, i.e., 600 chart images in total.

The corresponding alt text for each image follows the same structure as the VisText dataset (Tang and Boggust, 2023) in that it is composed of L1-, L2-, and L3-level content. The texts were manually authored by students according to guidelines (see Table 8 in Appendix D) that follow the theoretical model of Lundgard and Satyanarayan (2021) (see Section 2.1). Contrary to the procedure for the VisText dataset, the tags for the alt text in this dataset were then *manually* assigned by a digital accessibility expert in accordance with the tag schema described in Table 1. Table 9 in Appendix E shows three final examples of alt text with tags for each chart type.

3.3 Fine-tuning ChartGemma

Multiple chart-specific VLMs have been developed in recent years, such as MatCha (Masry et al., 2023), ChartLlama (Han et al., 2023), ChartAssistant (Meng et al., 2024), TinyChart (Zhang et al., 2024), Evochart (Huang et al., 2025), and ChartGemma (Masry et al., 2024). Among these models, ChartGemma (Masry et al., 2024) is an open-source model with a small number of parameters (2B) and has demonstrated strong performance (Huang et al., 2025). It employs the SigLIP vision encoder (Zhai et al., 2023) together with Gemma-2B (Chen et al., 2023) and is pretrained through instruction tuning on data directly generated from chart images without the involvement of underlying data tables.

For the experiments underlying the paper at hand, we fine-tuned ChartGemma, following the original training setup. Code was adopted from GitHub[8].

[3] https://huggingface.co/Qwen/Qwen2.5-3B-Instruct
[4] https://huggingface.co/google/gemma-1.1-2b-it
[5] https://arxiv.org
[6] https://link.springer.com
[7] https://www.acm.org
[8] https://github.com/vis-nlp/ChartGemma

Tag	Level	Applicable Charts	Definition
<CHART_TYPE>	L1	Bar, Line, Pie	Type of the visualization
<CHART_TITLE>	L1	Bar, Line, Pie	Title of the visualization
<X_AXIS_LABEL>	L1	Bar, Line	Label on the horizontal axis
<X_AXIS_SCALE>	L1	Bar, Line	Value ranges on the horizontal axis
<Y_AXIS_LABEL>	L1	Bar, Line	Label on the vertical axis
<Y_AXIS_SCALE>	L1	Bar, Line	Value ranges on the vertical axis
<SEGMENT_LABEL>	L1	Pie	Label for each segment
<MAX_EXTREME>	L2L3	Bar, Line, Pie	Highest value in the visualization
<MIN_EXTREME>	L2L3	Bar, Line, Pie	Lowest value in the visualization
<STATISTIC>	L2L3	Bar	Prominent data value or feature in the visualization
<TREND>	L2L3	Bar	Notable patterns change over a sequence
<OVERALL_TREND>	L2L3	Line	Value change patterns over the entire data range
<LOCAL_TREND>	L2L3	Line	Value changes within specific intervals
<PROPORTION>	L2L3	Pie	Prominent percentage of segments

Table 1: Overview of the existing (L1) and proposed (L2L3) tag schema across the three targeted chart types.

Dataset	Chart Types	No. of Alt Texts	L1 Tagging Method	L2L3 Tagging Method
VisText (Tang and Boggust, 2023)	Bar, Line	1,870	Preexisting	LLMs
Chart-Alt (Ours)	Bar, Line, Pie	600	Manual	Manual

Table 2: Dataset overview.

Fine-tuning consisted of injecting tags as part of the alt text generation task. We evaluated five different tag injection strategies for each level (L1, L2L3, and concatenated L1L2L3) as summarized in Table 3. The strategies were defined by varying four factors:

- **Embedded tag (E):** Tags were embedded inline within the ground truth alt text for each image.
- **List tag (L):** Tags were provided separately from the alt text as an ordered list for each image.
- **Tag schema (S):** The tag schema definition was explicitly included in the prompt.
- **Two-stage (2stage):** An additional stage for classifying sentences with tags was introduced before generating the alt text.

Additionally, a **baseline (B)** configuration was added for each level, consisting of applying pretrained ChartGemma directly. More specific training prompts, training labels, and inference examples are provided in Table 10 in Appendix F.

3.4 Evaluation metrics

Going beyond previous research (see Section 2.3), three categories of evaluation metrics were used to assess the generated alt text. The first category consisted of reference-based metrics commonly used in the chart summarization task and previously applied in the alt text generation task: BLEU-4 (Papineni et al., 2002), ROUGE_L (Lin, 2004), Translation Edit Rate (TER) (Snover et al., 2006) and SBERT (Reimers and Gurevych, 2019). These metrics measure textual overlap and semantic similarity between alt text candidates and references.

The second, newly added, category comprised reference-free metrics that evaluate the alignment between generated alt text and chart image directly, without relying on reference text, and provide a notion of factual consistency. For the experiments at hand, we used ChartVE (Huang et al., 2024) and CLIPScore (Hessel et al., 2021).

The third, novel, category employed four LLMs/VLMs as judges. Two evaluation criteria were considered:

- **Structural clarity (SC)**: This criterion measures information organization, including the *logical ordering* of information. Two LLMs were applied: Gemma-2-2B-IT [9] and Llama-3.2-3B-Instruct [10]. A five-point evaluation scale was used. The prompt is shown in Table 14 in Appendix H.
- **Information concentration (IC)**: This crite-

[9] https://huggingface.co/google/gemma-2-2b-it
[10] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

Level	Exp ID	Tag Injection Strategy			
		Embedded Tag	List Tag	Tag Schema	Two Stage
L1	L1_B	✗	✗	✗	✗
	L1_E	✓	✗	✗	✗
	L1_E_S	✓	✓	✗	✗
	L1_L_S	✗	✓	✓	✗
	L1_L_S_2stage	✓	✓	✓	✓
L2L3	L2L3_B	✗	✗	✗	✗
	L2L3_E	✓	✗	✗	✗
	L2L3_E_S	✓	✓	✗	✗
	L2L3_L_S	✗	✓	✓	✗
	L2L3_L_S_2stage	✓	✓	✓	✓
L1L2L3	L1L2L3_B	✗	✗	✗	✗
	L1L2L3_E	✓	✗	✗	✗
	L1L2L3_E_S	✓	✓	✗	✗
	L1L2L3_L_S	✗	✓	✓	✗
	L1L2L3_L_S_2stage	✓	✓	✓	✓

Table 3: Experimental configurations for different tag injection strategies.

rion measures how well an alt text focuses on the *key visual elements* and *avoids unnecessary or low-importance details*. Two VLMs were employed: Qwen2-VL-2B-Instruct [11] and Phi-3.5-Vision-Instruct [12]. A five-point evaluation scale was used, as shown in Table 15 in Appendix H.

Paired bootstrap resampling (Koehn, 2004) was applied for statistical significance testing between the baseline ($Level_B$) and the best-performing strategy at each alt text level; this is a widely used method that provides reliable results for small sample sizes. We adopted existing code[13] for the implementation.

4 Results

In this section we present the results of fine-tuning ChartGemma with and without tags on two datasets. Table 4 and Table 5 show the performance on the VisText and Chart-Alt datasets, respectively. The complete significance test results for both datasets are provided in Table 16 and Table 17 in Appendix I. Note that for VisText, comparison with previous approaches is not possible due to differences in sample sizes (We used 1,870 of the over 8,000 chart image/alt text pairs of the VisText dataset for which two LLMs agreed on the structural tags assigned) and evaluation metrics.

[11] https://huggingface.co/Qwen/Qwen2-VL-2B-Instruct

[12] https://huggingface.co/microsoft/Phi-3.5-vision-instruct

[13] https://github.com/neubig/util-scripts/blob/master/paired-bootstrap.py

4.1 Alt text generation on VisText

The first two parts of the table show performance gaps when generating alt text separately by levels. More specifically, L2L3-level alt text yielded lower generation performance than L1-level alt text.

The four reference-based metrics showed relatively consistent results across all three input levels, suggesting that the tag injection strategy that provided tags as a separate list together with a schema in the prompt ($Level_L_S$) led to the best performance improvements over the baseline, with the exception of the TER score at the L2L3-level. Both ChartVE and CLIPScore scores indicated that incorporating tags improved the alignment between the chart image and the generated alt text only at L2L3-level.

For evaluating the structural clarity of the output text, both LLM judges agreed that improvements were observed when using either L1 or L2L3 inputs independently, but their judgments diverged on the concatenation of all levels. Regarding information concentration, both VLM judges agreed that incorporating tags in the generation of alt text resulted in higher scores, although the degree of improvement varied across different tag injection strategies.

Significance testing revealed that the tag injection strategy that provided tags as a separate list together with a schema in the prompt ($L1_L_S$) improved the generation results significantly at L1 level based on the four reference-based metrics and Qwen-as-a-judge on information concentration. However, based on ChartVE and CLIPScore, the baseline ($L1_B$) performed significantly better than other strategies. For L2L3-level alt text, the strategy ($L2L3_L_S$) that provided tags as a

separate list together with a schema in the prompt showed significant improvements based on BLEU-4, CLIPScore, and Qwen-as-a-judge on information concentration. For concatenated L1L2L3-level alt text, similar results were observed as for L1-level alt text: Four reference-based metrics showed significant improvements for the strategy ($L1L2L3_L_S$) that provided tags as a separate list together with a schema in the prompt. The baseline strategy still performed significantly better in other metrics, except Gemma-as-a-judge on structural clarity.

4.2 Alt Text generation on Chart-Alt

The evaluation results on the Chart-Alt dataset behaved similarly as on the VisText dataset in that the results also showed a performance gap between L1-level and L2L3-level alt text. The four reference-based metrics showed improvement across all text levels, the best scores were observed according to the $L1_L_S$ and the $L2L3_L_S$ strategies (the strategies that provided tags as a separate list together with a schema in the prompt). Improvements on ChartVE and CLIPScore were observed in strategies when tags were injected, while the opposite was the case for the $L1L2L3_B$ (baseline) strategy.

For the structural clarity evaluation, Gemma and Llama judges produced consistent results: For L1- and L2L3-level alt text, the best scores were observed with embedded tags ($Level_E$ and $Level_E_S$), whereas the separating tag strategies ($L1L2L3_E_S$) showed the best performance for L1L2L3-level alt text. Interestingly, the two VLM judges exhibited substantial discrepancies in the information concentration evaluation: Qwen appeared less sensitive, producing repeated or identical scores, while Phi showed opposite evaluation results compared to Qwen for several strategies, including $L1_E_S$, $L2L3_L_S$, $L1L2L3_E$, and $L1L2L3_E_S$.

Statistically significant improvements were observed at the L2L3 and L1L2L3 levels under the $Level_L_S$ strategy, with no statistically significant evidence showing on the L1 level.

5 Discussion

As shown in both tables 4 and 5, there is a large gap between L1-level and L2L3-level alt text generation performance. These findings are consistent with previous experiments on the VisText dataset (Tang and Boggust, 2023; Moured et al., 2024b), providing further evidence that generating L2L3-level alt text is more challenging than generating L1-level alt text.

The evaluation results on both datasets indicate that injecting tags in the generation process improves performance. More specifically, we found that injecting tags as a separate list and including the schema in the prompt ($Level_L_S$) achieves the best performance compared to other injection strategies. This is likely due to the fact that the tag schema explicitly incorporated into the prompt offers the strongest guidance for the model. In contrast, embedded tags require the model to infer structure from lexical output, which leads to limited improvement.

Another interesting finding is that incorporating tags has limited impact on image-text alignment, as evaluated by the two reference-free metrics. One possible reason for this is that the tags were mainly applied at the text level; therefore, these reference-free metrics may not fully capture the improvements.

In terms of evaluation results from four LLM/VLM judges, more consistent results were observed on the VisText dataset, which contains a larger number of samples, while less consistent results were found on the Chart-Alt dataset likely due to its smaller sample size. We also observed divergent results for the same tag injection strategies across datasets. This suggests that the VLM/LLM-as-a-judge paradigm requires further investigation for the task observed, that of alt text generation.

Although we cannot directly compare our results with previous experiments using the VisText dataset due to differences in sample sizes (we used only the chart image/alt text pairs for which two LLMs agreed in their structural tags assigned for our experiments, which amounted to 1,870 of the over 8,000 pairs of the VisText dataset) and evaluation metrics, our results indicate that providing a tag schema can guide VLMs to improve alt text generation, which bridges the gap between raw image input and generated text.

We note that these results are based on a single chart-specific VLM (ChartGemma); it therefore remains unclear how they generalize on other VLMs. In addition, only three chart types were included in the present experiments; results may differ and may require an even more fine-grained tag schema for more complex chart types, such as scatter plots, box plots, or heatmaps.

Exp ID	Tag	Reference-based Metrics				Reference-free Metrics		Structural Clarity		Information Concentration	
		BLEU-4 ↑	ROUGE-L ↑	TER ↓	SBERT ↑	ChartVE ↑	CLIPScore ↑	Gemma ↑	Llama ↑	Qwen ↑	Phi ↑
L1_B	✗	0.4310	0.6212	50.1627	0.9037	**0.5191***	**0.3956***	3.0403	4.2581	3.5376	3.2558
L1_E	✓	0.4429	0.6263	49.9316	0.9082	0.3596	0.3918	3.0064	4.0728	3.5215	2.5000
L1_E_S	✓	0.4320	0.6213	49.8363	0.9159	0.5186	0.3952	3.0258	4.2468	3.4247	2.9322
L1_L_S	✓	**0.6102***	**0.7824***	**30.6904***	**0.9513***	0.2500	0.3864	**3.3797**	**4.2981**	**3.7204***	**3.3896**
L1_L_S_2stage	✓	0.5850	0.7660	31.4412	0.9459	0.2775	0.3858	3.3155	4.2190	3.6505	2.8970
L2L3_B	✗	0.0326	0.2717	**92.2388**	0.6044	0.6239*	0.3544	1.6824	3.1260	3.1398	1.2188
L2L3_E	✓	0.0400	0.2701	97.7233	0.5976	0.4239	**0.3614**	1.8144	3.1923	**3.2366**	1.6944
L2L3_E_S	✓	0.0309	0.2657	92.8798	0.5918	**0.6494**	0.3499	1.6593	**3.2806**	3.1613	1.1935
L2L3_L_S	✓	**0.0442***	**0.2765**	96.8748	**0.6078**	0.4664	0.3601*	**2.0101**	3.0614	3.2258*	**2.0233***
L2L3_L_S_2stage	✓	0.0264	0.2659	92.4764	0.5936	0.6339	0.3476	1.9510	3.0593	3.1344	1.7941
L1L2L3_B	✗	0.3233	0.5247	60.3368	0.8948	**0.3084***	**0.3976***	3.1220	**4.1231**	3.9677*	3.2340
L1L2L3_E	✓	0.3317	0.5084	61.8944	0.8808	0.2156	0.3862	3.1321	3.9542	3.9570	**3.5500**
L1L2L3_E_S	✓	0.3310	0.5097	61.8029	0.8906	0.2319	0.3972	3.1667	4.0190	3.9409	3.4447
L1L2L3_L_S	✓	**0.4476***	**0.6374***	**48.0984***	**0.9284***	0.1496	0.3856	**3.392***	4.0606	3.8925	3.2000
L1L2L3_L_S_2stage	✓	0.4207	0.6252	50.4640	0.9226	0.1777	0.3895	3.3191	4.1069	**3.9731**	3.2182

Table 4: Performance comparison across different tag injection strategies on the VisText dataset. For the experiment configurations, see Table 3. Significance tests were run between two strategies, highlighted in gray. * marks significant values.

Exp ID	Tag	Reference-based Metrics				Reference-free Metrics		Structural Clarity		Information Concentration	
		BLEU-4 ↑	ROUGE-L ↑	TER ↓	SBERT ↑	ChartVE ↑	CLIPScore ↑	Gemma ↑	Llama ↑	Qwen ↑	Phi ↑
L1_B	✗	0.2931	0.6199	58.8215	0.8798	0.1615	0.3610	3.3333	4.0870	3.8525	2.2500
L1_E	✓	0.3103	0.6108	60.1346	0.8886	0.1634	**0.3651**	**3.4082**	3.3448	**3.9016**	2.6667
L1_E_S	✓	0.2587	0.5728	64.0200	0.8720	0.1756	0.3612	3.3860	**4.2903**	**3.9016**	2.0000
L1_L_S	✓	**0.3129**	**0.6487**	**56.4374**	**0.8889**	0.1840	0.3622	3.2787	4.1034	3.8689	2.400
L1_L_S_2stage	✓	0.2706	0.6038	62.2837	0.8303	**0.2161**	0.3526	3.2333	3.9286	3.7705	**2.8182**
L2L3_B	✗	0.1491	0.4147	84.3105	0.7740	0.2568	0.3365	2.8596	3.8333	3.7869	2.4286
L2L3_E	✓	0.1262	0.4228	81.0638	0.7676	0.3403	0.3336	**2.9286**	**4.0571**	**3.9016**	**3.000**
L2L3_E_S	✓	0.1215	0.4218	80.5558	0.7489	**0.2686**	**0.3381**	2.8491	3.9737	3.8033	2.4445
L2L3_L_S	✓	**0.1834***	**0.4747***	**75.0562***	**0.7831**	0.2507	0.3334	2.9074	4.0286	3.8852*	1.500
L2L3_L_S_2stage	✓	0.1220	0.4117	78.0494	0.7518	0.2785	0.3260	2.8400	4.0556	3.6886	2.400
L1L2L3_B	✗	0.2010	0.4767	74.1343	0.8659	**0.1402**	**0.3732**	3.7917	4.1250	**4.0000**	2.0500
L1L2L3_E	✓	0.2389	0.5049	**66.3942**	**0.8822**	0.1004	0.3694	3.4667	4.200	3.9016	2.5000
L1L2L3_E_S	✓	0.2195	0.4884	68.4889	0.8661	0.0960	0.3624	3.4412	4.2075	3.9672	**3.2500**
L1L2L3_L_S	✓	**0.2515***	**0.5308***	67.4959*	0.8752	0.1250	0.3696	**3.8444**	**4.4815***	**4.0000**	1.6667
L1L2L3_L_S_2stage	✓	0.2344	0.5175	66.8508	0.8500	0.1263	0.3673	3.6852	4.1957	**4.0000**	1.9286

Table 5: Performance comparison across different tag injection strategies on the Chart-Alt dataset. Experiment settings are explained in Table. 3. Significance tests were run between two strategies, highlighted in gray. * marks significant values.

6 Conclusion

In this paper, we introduced tags as guidance in the alt text generation task for scientific charts, an underexplored task that is of high relevance for blind persons and persons with visual impairments. We developed a tag schema across three commonly used chart types by means of a clustering analysis. We compared fine-tuning results with and without tags on two datasets and evaluated them using three categories of metrics, thereby going beyond prior evaluations of this task.

The evaluation results demonstrated that incorporating tags can improve alt text generation performance, particularly in terms of lexical overlap, semantic similarity, structural clarity, and information concentration. Among the different tag injection strategies explored, providing tags as a separate list together with schema information in the prompt achieved the strongest overall performance. Our work not only shows that incorporating intermediate representations such as tag schema can improve alt text generation performance, but it also suggests that such tag schema can offer guidance for further chart accessibility research and human alt text generation practices.

6.1 Limitations

Our work has three main limitations: (1) Three chart types were included, leaving other chart types (such as histograms, scatter plots, or box plots) as well as more complex charts that contain multiple

data series (such as grouped bar charts, multiple-line charts) unexplored. (2) Two datasets were used for fine-tuning and evaluation in this work; other datasets could be considered, such as ChartSumm (Rahman et al., 2023) or Pew (Kanthara et al., 2022). (3) Evaluation relied on automatic evaluation metrics and LLMs/VLMs as judges, while we acknowledge that human evaluation still remains crucial.

Ethical considerations

We acknowledge that the models used in the experiments underlying this paper, both the VLM that was fine-tuned as well as the LLM and VLM that acted as judge, may exhibit bias (Pu et al., 2025; Szymanski et al., 2025; Li et al., 2025; Laskar et al., 2025).

The Chart-Alt dataset was authored by student annotators. All students signed consent forms before they started their task, and they were compensated at a standard hourly rate. This task did not expose any harmful content to participants.

The Chart-Alt dataset is an extension of the VisText dataset from Tang and Boggust (2023), and the annotations added to the raw VisText dataset are consistent with its intended use case. The dataset is distributed under the GNU General Public License v3.0, which was also used for the original VisText dataset. Our code and data will be made available with the camera-ready version.

In this work, ChatGPT was employed for devicing the prompts for the LLM/VLM-as-a-judge paradigm as well as for writing assistance, table formatting, and code cleanup. More specifically, the prompts used were “rephrase this [sentence]", “does this [sentence] make sense", “format this [table] by adding a midline between [line] and [line]", and “clean this [code] and remove redundancy parts". All generated content was validated by the first author.

References

Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, and 1 others. 2023. Pali-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*. Doi: 10.48550/arXiv.2310.09199.

Sanjana Shivani Chintalapati, Jonathan Bragg, and Lucy Lu Wang. 2022. A dataset of alt texts from hci publications: Analyses and uses towards producing more descriptive alt texts of data visualizations in scientific papers. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–12. Doi: 10.1145/3517428.3544796.

Jiwon Choi and Jaemin Jo. 2022. Intentable: A mixed-initiative system for intent-based chart captioning. In *2022 IEEE Visualization and Visual Analytics (VIS)*, pages 40–44. IEEE. Doi: 10.1109/VIS54862.2022.00017.

Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. 2023. Chartllama: A multimodal llm for chart understanding and generation. *arXiv preprint arXiv:2311.16483*. Doi: 10.48550/arXiv.2311.16483.

Maila Hardin, Daniel Hom, Ross Perez, and Lori Williams. 2012. Which chart or graph is right for you. *Tell Impactful Stories with Data. Tableau Software*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528. Doi: 10.18653/v1/2021.emnlp-main.595.

Ting-Yao Hsu, C Lee Giles, and Ting-Hao'Kenneth' Huang. 2021. Scicap: Generating captions for scientific figures. *arXiv preprint arXiv:2110.11624*. Doi: 10.18653/v1/2021.findings-emnlp.277.

Chan-Wei Hu, Reuben Luera, Franck Dernoncourt, Huanjiang Liu, Seon Gyeom Kim, Ashish Singh, Puneet Mathur, Bo Ni, Samyadeep Basu, Hongjie Chen, and 1 others. 2026. Figure captioning with large generative models: A survey.

Kung-Hsiang Huang, Mingyang Zhou, Hou Pong Chan, Yi Fung, Zhenhailong Wang, Lingyu Zhang, Shih-Fu Chang, and Heng Ji. 2024. Do lvlms understand charts? analyzing and correcting factual errors in chart captioning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 730–749. Doi: 10.18653/v1/2025.emnlp-main.138.

Muye Huang, Han Lai, Xinyu Zhang, Wenjun Wu, Jie Ma, Lingling Zhang, and Jun Liu. 2025. Evochart: A benchmark and a self-training approach towards real-world chart understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3680–3688.

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*, 28(1):1095–1105. Doi: 10.1109/TVCG.2021.3114846.

Shankar Kanthara, Rixie Tiffany Ko Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq R. Joty. 2022. Chart-to-text: A large-scale benchmark for chart summarization. In *Annual Meeting of the Association for Computational Linguistics*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966. PMLR.

Md Tahmid Rahman Laskar, Mohammed Saidul Islam, Ridwan Mahbub, Ahmed Masry, Mizanur Rahman, Amran Bhuiyan, Mir Tafseer Nayeem, Shafiq Joty, Enamul Hoque, and Jimmy Huang. 2025. Judging the judges: Can large vision-language models fairly evaluate chart comprehension and reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 1203–1216. Doi: 10.18653/v1/2025.acl-industry.83.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791. Doi: 10.18653/v1/2025.emnlp-main.138.

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.

Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Eisenschlos. 2023. Matcha: Enhancing visual language pretraining with math reasoning and chart derendering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12756–12770. Doi: 10.18653/v1/2023.acl-long.714.

Mengsha Liu, Daoyuan Chen, Yaliang Li, Guian Fang, and Ying Shen. 2024. Chartthinker: A contextual chain-of-thought approach to optimized chart summarization. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3057–3074.

Alan Lundgard and Arvind Satyanarayan. 2021. Accessible visualization via natural language descriptions: A four-level model of semantic content. *IEEE transactions on visualization and computer graphics*, 28(1):1073–1083. Doi: 10.1109/TVCG.2021.3114770.

Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. 2023. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14662–14684. Doi: 10.18653/v1/2023.emnlp-main.906.

Ahmed Masry, Megh Thakkar, Aayush Bajaj, Aaryaman Kartha, Enamul Hoque, and Shafiq Joty. 2024. Chartgemma: Visual instruction-tuning for chart reasoning in the wild. *arXiv preprint arXiv:2407.04172*. Doi: 10.48550/arXiv.2407.04172.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205. Doi: 10.21105/joss.00205.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*. Doi: 10.48550/arXiv.1802.03426.

Fanqing Meng, Wenqi Shao, Quanfeng Lu, Peng Gao, Kaipeng Zhang, Yu Qiao, and Ping Luo. 2024. ChartAssistant: A universal chart multimodal language model via chart-to-table pre-training and multitask instruction tuning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7775–7803. Association for Computational Linguistics. Doi: 10.18653/v1/2024.findings-acl.463.

Omar Moured, Shahid Ali Farooqui, Karin Müller, Sharifeh Fadaeijouybari, Thorsten Schwarz, Mohammed Javed, and Rainer Stiefelhagen. 2024a. Alt4blind: A user interface to simplify charts alt-text creation. In *International Conference on Computers Helping People with Special Needs*, pages 291–298. Springer. Doi: 10.1007/978-3-031-62846-7_35.

Omar Moured, Jiaming Zhang, M Saquib Sarfraz, and Rainer Stiefelhagen. 2024b. Altchart: Enhancing vlm-based chart summarization through multi-pretext tasks. In *International Conference on Document Analysis and Recognition*, pages 349–366. Springer. Doi: 10.1007/978-3-031-70533-5_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 13th International Conference on Natural Language Generation*, pages 138–147. Doi: 10.18653/v1/2020.inlg-1.20.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318. Doi: 10.3115/1073083.1073135.

Shu Pu, Yaochen Wang, Dongping Chen, Yuhang Chen, Guohao Wang, Qi Qin, Zhongyi Zhang, Zhiyuan Zhang, Zetong Zhou, Shuang Gong, and 1 others. 2025. Judge anything: Mllm as a judge across any modality. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5742–5753. Doi: 10.1145/3711896.3737409.

Raian Rahman, Rizvi Hasan, Abdullah Al Farhad, Md Tahmid Rahman Laskar, Md Hamjajul Ashmafee, and Abu Raihan Mostofa Kamal. 2023. Chartsumm: A comprehensive benchmark for automatic chart summarization of long and short summaries. *arXiv preprint arXiv:2304.13620*. Doi: 10.21428/594757db.0b1f96f6.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. Doi: 10.48550/arXiv.1908.10084.

Ashish Singh, Prateek Agarwal, Zixuan Huang, Arpita Singh, Tong Yu, Sungchul Kim, Victor Bursztyn, Nikos Vlassis, and Ryan A Rossi. 2023. Figcaps-hf: A figure-to-caption generative framework and benchmark with human feedback. *arXiv preprint arXiv:2307.10867*. Doi: 10.48550/arXiv.2307.10867.

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 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Annalisa Szymanski, Noah Ziems, Heather A Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A Metoyer. 2025. Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks. In *Proceedings of the 30th international conference on intelligent user interfaces*, pages 952–966. Doi: 10.1145/3708359.3712091.

Benny J. Tang and Angie Boggust. 2023. Vistext: A benchmark for semantically rich chart captioning. In *Annual Meeting of the Association for Computational Linguistics*. Doi: 10.18653/v1/2023.acl-long.401.

Peixin Xu, Yujuan Ding, and Wenqi Fan. 2024. Chartadapter: Large vision-language model for chart summarization. *arXiv preprint arXiv:2412.20715*. Doi: 10.48550/arXiv.2412.20715.

Chuqiao Yan, Hans-Peter Hutter, Felix M Schmitt-Koopmann, and Alireza Darvishy. 2025. Chart accessibility: A review of current alt text generation. *IEEE Access*. Doi: 10.1109/ACCESS.2025.3571626.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11941–11952.

Liang Zhang, Anwen Hu, Haiyang Xu, Mingshi Yan, Yichen Xu, Qin Jin, Ji Zhang, and Fei Huang. 2024. Tinychart: Efficient chart understanding with program-of-thoughts learning and visual token merging. In *Conference on Empirical Methods in Natural Language Processing*. Doi: 10.18653/v1/2024.emnlp-main.112.

Jonathan Zong, Crystal Lee, Alan Lundgard, JiWoong Jang, Daniel Hajas, and Arvind Satyanarayan. 2022. Rich screen reader experiences for accessible data visualization. In *Computer Graphics Forum*, volume 41, pages 15–27. Wiley Online Library. Doi: 10.1111/cgf.14519.

Appendix

A Four-level semantic model

This example (see Table 6) explains the Four-Level semantic model developed by Lundgard and Satyanarayan (2021). The image is from VisText dataset (Tang and Boggust, 2023). Since its original alt text only contains three levels, we re-authored the alt text to better demonstrate the semantic model.

B Statista dataset example

Table 7 shows two examples from Statista dataset (Kanthara et al., 2022). We only used its bar chart and line chart descriptions to develop the tag schema for the experiments reported in this paper.

C Clustering analysis

We applied unsupervised clustering for each chart type using the top 500 frequent words from Statista chart descriptions (Kanthara et al., 2022). We first embedded these words using Sentence-BERT all-MiniLM-L6-v2 [14] and then clustered using HDBSCAN algorithm (McInnes et al., 2017). The clustering parameters used were min_cluster_size=3, min_samples=1, with Euclidean distance as similarity metric. The UMAP visualization is shown in Figure 1.

D Chart-Alt dataset authoring

Nine students with at least a college degree were recruited to author the alt text, and they were compensated at a rate of 25 CHF per hour. Before starting, each participant received a handout describing the annotation tasks, procedures, intended dataset usage, and annotation guidelines. The students were informed that the dataset would be used to fine-tune a chart-specific Vision Language Model for chart accessibility research. The annotation guidelines were developed based on survey findings from Yan et al. (2025) and are summarized in Table 8. The annotation task involved no known risks, and participants were informed that they could withdraw at any time if they felt uncomfortable.

E Examples from Chart-Alt dataset

Table 9 shows three data samples from our Chart-Alt dataset, one example per chart type.

[14] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

F Tag injection strategy

We evaluated five tag injection strategies across each alt text level. Table 10 summarizes the training prompts, input labels, and inference examples, which demonstrate how tags are injected under different configurations, The complete prompts are listed in Table 11.

G Prompt for tagging VisText

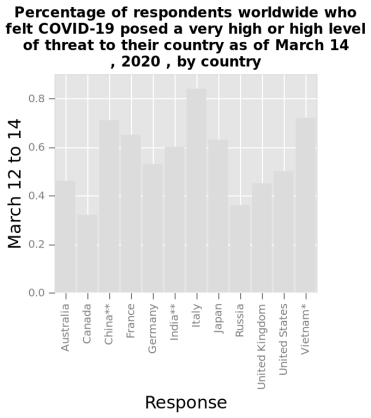
Two LLMs were employed to tag the VisText dataset (Tang and Boggust, 2023). Table 12 shows the prompt for tagging line charts and one example before tagging and after tagging. Table 13 shows the prompt and an example for tagging bar charts.

H Evaluation prompt

Two LLMs and two VLMs were employed to evaluate structural clarity (see Table 14) and information concentration (see Table 15) of the output text. The same prompts were used for each evaluation criterion.

I Significance test

Paired-Bootstrap resampling (Koehn, 2004) was used to conduct significance tests between two strategies ($Level_B$ and $Level_L_S$) for each alt text level. The results are presented in Table 16 and Table 17 for the two datasets.



[L1] This is a bar chart titled "Percentage of respondents worldwide who felt COVID-19 posed a very high or high level of threat to their country as of March 14, 2020, by country." The y-axis is labeled "March 12 to 14" on a linear scale ranging from 0.0 to 0.8, while the x-axis represents country responses using a categorical scale ranging from Australia to Vietnam.

[L2] Italy has the highest percentage of respondents who perceived COVID-19 as a very high-level threat, with a value of 0.8. Canada reported the lowest percentage, with a value of 0.3.

[L3] The chart reveals that perceptions of the COVID-19 threat varied across countries.

[L4] This variation may be related to differences in the timing of the first confirmed COVID-19 cases in each country, although the WHO declared the outbreak a public health emergency on January 30, 2020.

Table 6: An example of alt text following the four-level semantic model. The image is from the VisText dataset (Tang and Boggust, 2023). The alt text was composed by the first author of the present paper.

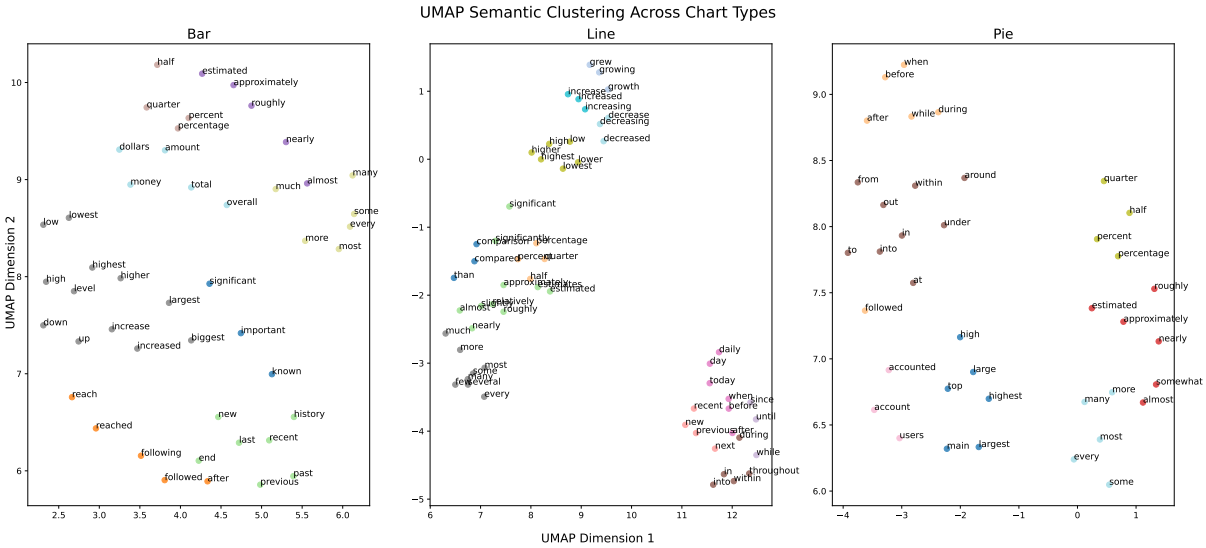


Figure 1: UMAP cluster visualization across three chart types, showing the distribution of keyword clusters after removing less relevant clusters such as stop words, function words, and domain/context-specific terms.

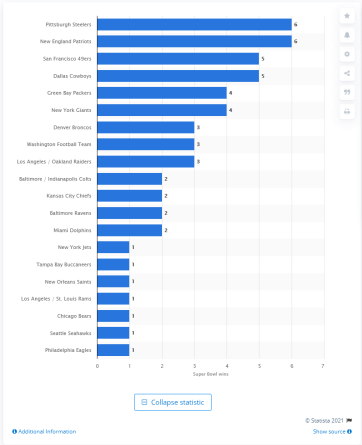
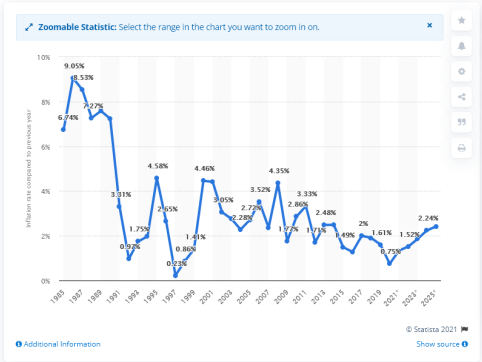
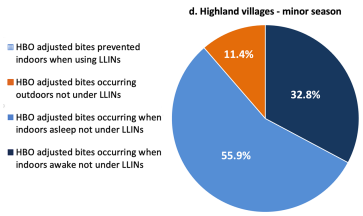
Image	Description
	The New England Patriots have won the Super Bowl a record six times, most recently Super Bowl LIII in February 2019. Joining them at the top of this list is the Pittsburgh Steelers, whose most recent of their six titles came in 2009. The franchise's most successful period came between 1975 and 1980, during which they claimed four titles. What is the Super Bowl? The Super Bowl is the annual championship game of the National Football League (NFL), pitting the champions of the NFL two conferences (AFC and NFC) against each other. The 2020 Super Bowl saw the Kansas City Chiefs emerge victorious against the San Francisco 49ers in a thrilling final quarter comeback, a spectacle that was watched by 99.9 million viewers in the United States. The most watched Super Bowl in history is Super Bowl XLIX in 2015, when a record 114.4 million viewers saw the Patriots beat the Seattle Seahawks. Legendary quarterbacks The New England Patriots defeated the Los Angeles Rams in Super Bowl LIII in 2019 with a score of 13-3, which made quarterback Tom Brady the most decorated player of all time with six Super Bowl victories. While Tom Brady is certainly a legend of the game, he does not top the list of quarterbacks with the most passing yards in NFL history. This honor belongs to fellow quarterback legend, Drew Brees, who has thrown a total of over 74 thousand yards in his career. However, Brady is the undoubted master of the playoffs he has won a total of 30 playoff games during his career, almost double that of his closest rival on the list, Joe Montana, making him by far the quarterback with the most playoff wins. Brady is also peerless when it comes to total passing yards for the Patriots franchise his over-74,000 yards are more than double those of Drew Bledsoe, who occupies second place on the all-time list.
	The statistic shows the inflation rate in Australia from 1985 to 2019, with projections up until 2025. The inflation rate is calculated using the price increase of a defined product basket. This product basket contains products and services, on which the average consumer spends money throughout the year. They include expenses for groceries, clothes, rent, power, telecommunications, recreational activities and raw materials (e.g. gas, oil), as well as federal fees and taxes. In 2019, the average inflation rate in Australia was at about 1.61 percent compared to the previous year. Australia's economy Australia has one of the world's largest economies and is a significant global importer and exporter. It is also labeled as one of the G20 countries, also known as the Group of Twenty, which consists of 20 major economies around the globe. The Australian economy is highly dependent on its mining sector as well as its agricultural sector in order to grow, and it exports the majority of these goods to eastern Asian countries, most prominently China. Large quantities of exports have helped Australia maintain a stable economy and furthered economic expansion, despite being affected by several economic obstacles. Australia's GDP has seen a significant increase over the past decade, more than doubling its value, and experienced a rather quick recovery from the 2008 financial crisis, which indicates that the country experienced economic growth as well as higher productivity. One of the primary reasons is the further development of the nation'™s mining industry coupled with the expansion and success of many Australian mining companies.

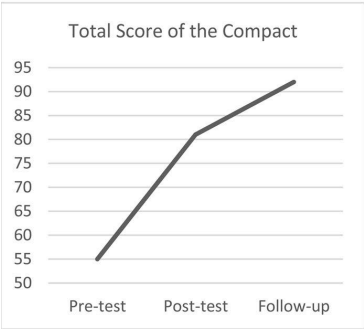
Table 7: Two examples from the Statista Dataset (Kanthara et al., 2022). We include only bar and line charts, which were used for clustering.

Category	Explanation
1. Structure and Elements	
General to Specific	Describe features from general to specific details.
Two-Part Description	Include a two-part description for each image: a high-level overview (L1-level) and detailed description (L2L3-level).
High Level Overview (L1)	Mention chart type, title/content, axis labels/range, and legend.
Detailed Description (L2L3)	Provide key takeaways, statistical data, comparisons, or trends.
Omission of Visual Attributes	Omit visual attributes such as colors unless it is critical for data understanding.
2. Content and Language	
Avoid Redundancy	Focus on information that is not covered in the main text; reference main text if needed.
Avoid Repetition	Do not repeat the caption in the alt text.
Abbreviations	Write out abbreviations and symbols in full.
Accuracy	Provide accurate information.
Consistency	Maintain consistent terminology and style with the main text.
Conciseness	Keep it around 61-82 words, and less than 8 sentences. Complex charts can be longer.
Grammar	Ensure correct grammar and spelling throughout.

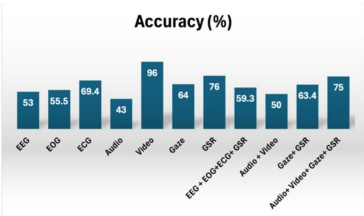
Table 8: Guidelines for authoring Chart-Alt ground truth.



Image_id: 1.png
Chart_type_1: Pie,
Chart_type_2: Pie,
L1_description: This pie chart is titled “Highland villages - minor season” and shows the distribution of four different conditions: ’HBO adjusted bites prevented when using LLINs’, ’HBO adjusted bites occurring outdoors not under LLINs’, ’HBO adjusted bites occurring when indoors asleep not under LLINs’, and ’HBO adjusted bites occurring when indoors awake not under LLINs’.
L2L3_description: The largest segment is ’HBO adjusted bites occurring when indoors asleep not under LLINs’, with 55.9%. The smallest segment is ’HBO adjusted bites occurring outdoors not under LLINs’, with 11.4%.
L1_description_tag: This <CHART_TYPE> pie chart is <CHART_TITLE> titledĤighland villages - minor seasonänd shows the distribution of <SEGMENT_LABEL> four different conditions: ’HBO adjusted bites prevented when using LLINs’, ’HBO adjusted bites occurring outdoors not under LLINs’, ’HBO adjusted bites occuring when indoors asleep not under LLINs’ and ’HBO adjusted bites occuring when indoors awake not under LLINs’. Each condition is represented by one slice of the pie."
L2L3_description_tag: <MAX_EXTREME> The largest segment is ’HBO adjusted bites occurring when indoors asleep not under LLINs’, with 55.9%. <MIN_EXTREME> The ’HBO adjusted bites occurring outdoors not under LLINs’ has the lowest percentage, with 11.4%. <PROPORTION> The ’HBO adjusted bites occurring when indoors awake not under LLINs’ has 32.8%."
L1_tags: <CHART_TYPE>, <CHART_TITLE>, <SEGMENT_LABEL>,
L2L3_tags: <MAX_EXTREME>, <MIN_EXTREME>, <PROPORTION>



Image_id: 2.png,
Chart_type_1: Line
Chart_type_2: Simple Line
L1_description: This line chart is titled ’Total Score of the Compact’. The y-axis shows the value of the total score from 50 to 95, and the x-axis lists three categories: Pre-test, Post-test, and Follow-up.
L2L3_description: "The chart shows the total score of the compact increased from the ’Pre-test’ condition with a value of 55 to the ’Follow-up’ condition with a value of 93. The total score of the compact shows a sharp increase from the ’Pre-test’ condition to the ’Post-test’ condition. The total score of the compact shows a less rapid increase from the ’Post-test’ condition to the ’Follow-up’ condition.
L1_description_tag: This <CHART_TYPE> line chart is <CHART_TITLE> titled ’Total Score of the Compact’. The y-axis shows the <Y_AXIS_SCALE> value of the total score <Y_AXIS_SCALE> from 50 to 95, and the x-axis lists <X_AXIS_SCALE> three categories: Pre-test, Post-test, and Follow-up.
L2L3_description_tag: <OVERALL_TREND> The chart shows the total score of the compact increased from the ’Pre-test’ condition with a value of 55 to the ’Follow-up’ condition with a value of 93. <LOCAL_TREND> The total score of the compact shows a sharp increase from the ’Pre-test’ condition to the ’Post-test’ condition. <LOCAL_TREND> The total score of the compact shows a less rapid increase from the ’Post-test’ condition to the ’Follow-up’ condition.
L1_tags: <CHART_TYPE>, <CHART_TITLE>, <Y_AXIS_SCALE>, <Y_AXIS_SCALE>, <X_AXIS_SCALE>
L2L3_tags: <OVERALL_TREND>, <LOCAL_TREND>, <LOCAL_TREND>



Image_id: 3.png,
Chart_type_1: Bar
Chart_type_2: Simple Bar
L1_description: This bar chart shows the accuracy of various data sources. The x-axis lists eleven different data sources, and the y-axis represents the accuracy percentage, ranging from 0% to 100%.
L2L3_description: The chart shows that "Video" category has the highest accuracy at 96%. The "Audio" category has the lowest accuracy at 43% Other categories have accuracy between 50% and 76%.
L1_description_tag: This <CHART_TYPE> bar chart shows the accuracy of various data sources. The x-axis lists <X_AXIS_LABEL> eleven different data sources, and the y-axis represents the <Y_AXIS_LABEL> accuracy percentage, <Y_AXIS_SCALE> ranging from 0% to 100%."
L2L3_description_tag: <MAX_EXTREME> The chart shows that "Video" category has the highest accuracy at 96%. <MIN_EXTREME> The "Audio" category has the lowest accuracy at 43% <STATISTIC> Other categories have accuracy between 50% and 76%.
L1_tags: <CHART_TYPE>, <X_AXIS_LABEL>, <Y_AXIS_LABEL>, <Y_AXIS_SCALE>
L2L3_tags: <MAX_EXTREME>, <MIN_EXTREME>, <STATISTIC>

Table 9: Data samples per chart type from our Chart-Alt dataset.

Exp ID	Tag	Training Prompt	Training Label	Inference Text
L1L2L3_B	no	Prompt 1	L1_description + L2L3_description	This bar chart . . .
L1L2L3_E	yes	Prompt 1	L1_description_tag + L2L3_description	This <CHART_TYPE> . . .
L1L2L3_E_S	yes	Prompt 2	L1_description_tag + L2L3_description	This <CHART_TYPE> . . .
L1L2L3_L_S	yes	Prompt 3	L1_description + L2L3_description	This bar chart . . .
L1L2L3_L_S_2	yes	Prompt 4 + Prompt 3	L1_description + L2L3_description	This bar chart . . .

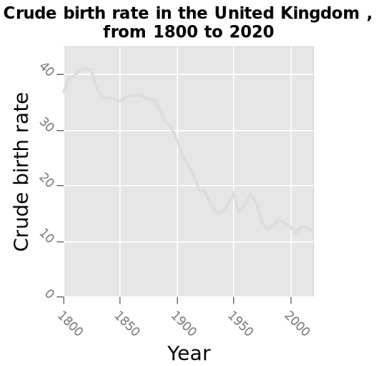
Table 10: Table of training prompts, input label and inference text examples for each experiment setting. The complete prompts are provided in Table 11. Training labels can be derived from data examples in Table 9. We only present examples for the L1L2L3-level alt text setting; the same procedures are applied to the L1 and L2L3 levels.

Prompt 1	<image> Generate a L1L2L3 description of the chart.
Prompt 2	<image> Determine the chart type. The chart type must be one of: line chart, bar chart, or pie chart. Then write the full L1L2L3 description using the appropriate tag schema below for that chart type. Schema for Bar chart: <CHART_TYPE>, <CHART_TITLE>, <X_AXIS_LABEL>, <X_AXIS_SCALE>, <Y_AXIS_LABEL>, <Y_AXIS_SCALE>, <MAX_EXTREME>, <MIN_EXTREME>, <STATISTIC>, <TREND> Schema for Line chart: <CHART_TYPE>, <CHART_TITLE>, <X_AXIS_LABEL>, <X_AXIS_SCALE>, <Y_AXIS_LABEL>, <Y_AXIS_SCALE>, <OVERALL_TREND>, <LOCAL_TREND>, <MAX_EXTREME>, <MIN_EXTREME> Schema for Pie chart: <CHART_TYPE>, <CHART_TITLE>, <SEGMENT_LABEL>, <MAX_EXTREME>, <MIN_EXTREME>, <PROPORTION> <PROPORTION> Rules: -Use only tags from the correct chart type. -Do not use tags from other chart types. Output
Prompt 3	<image> You are given a chart, tags schema for each chart type, and a list of guidance tags. Determine the chart type. The chart type must be one of: line chart, bar chart, or pie chart. Then write the full L1L2L3 description using the appropriate tag schema below for that chart type. Schema for Bar chart: <CHART_TYPE>, <CHART_TITLE>, <X_AXIS_LABEL>, <X_AXIS_SCALE>, <Y_AXIS_LABEL>, <Y_AXIS_SCALE>, <MAX_EXTREME>, <MIN_EXTREME>, <STATISTIC>, <TREND> Schema for Line chart: <CHART_TYPE>, <CHART_TITLE>, <X_AXIS_LABEL>, <X_AXIS_SCALE>, <Y_AXIS_LABEL>, <Y_AXIS_SCALE>, <OVERALL_TREND>, <LOCAL_TREND>, <MAX_EXTREME>, <MIN_EXTREME> Schema for Pie chart: <CHART_TYPE>, <CHART_TITLE>, <SEGMENT_LABEL>, <MAX_EXTREME>, <MIN_EXTREME>, <PROPORTION> Rules: - Write exactly one sentence for each tag. - Follow the same order as the provided tags. - Each sentence must correspond to one tag. - Use only tags from the correct chart type. - Do not use tags from other chart types. - Do NOT output the tags themselves. - Do NOT add extra information. Now write the L1L2L3 description for this chart. Tags: f"tags Output
Prompt 4	<image> Given this chart and sentence: {sentence} Predict the tag for the sentence.

Table 11: Table of the four prompts used across different tag injection strategies. We present only L1L2L3-level examples in this table. Similar prompts are applied for L1- and L2L3-level alt text by adapting the corresponding schema.

Prompt for Tagging Line Charts

```
You are a text-tagging assistant for simple (single-series) bar charts.

You must assign ONLY ONE tag to EACH sentence.

Available tags:
{tags_list}

Definitions:
<OVERALL_TREND> – sentence describes the global trend over the entire sequence.

<LOCAL_TREND> – sentence describes a local or partial trend over a specific sequence.

<MAX_EXTREME> – sentence describes the maximum value or peak of the series.

<MIN_EXTREME> – sentence describes the minimum value or trough of the series.

Sentences (numbered):
numbered_sentences

Output format: For EACH sentence i, output ONE line with:
i<TAB><TAG1>[,<TAG2>,<TAG3>...]

Where each TAG is one of:
tags_inline

Important:
- Do NOT rewrite or repeat the sentences.
- Do NOT output any words from the sentences.
- Do NOT output anything except these numbered tag lines.
```



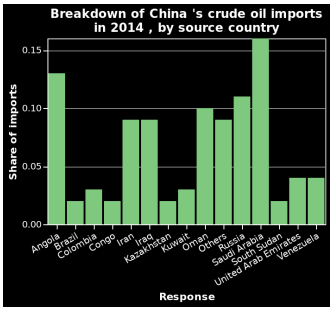
Input: The chart shows that for the most of time between 1800 and 2020 the rate of crude births in the UK was gradually decreasing. The chart shows that there were slight increases in the initial years of 1800s and around 1950s.

Output: <OVERALL_TREND> The chart shows that for the most of time between 1800 and 2020 the rate of crude births in the UK was gradually decreasing. <LOCAL_TREND> The chart shows that there were slight increases in the initial years of 1800s and around 1950s.

Table 12: The top row shows the prompt used for tagging line charts in the VisText dataset (Tang and Boggust, 2023). The bottom row shows an input alt text and output alt text with tags for a corresponding chart image.

Prompt for Tagging Bar Charts

```
You are a text-tagging assistant for simple (single-series) bar charts.

You must assign ONLY ONE tag to EACH sentence.

Available tags:
{tags_list}

Definitions:
<MAX_EXTREME> - sentence describes the highest value.

<MIN_EXTREME> - sentence describes the lowest value.

<STATISTIC> - sentence reports or summarizes the prominent numeric or statistical information.

<TREND> - sentence describes a notable changing pattern across categories.

Sentences (numbered):
{numbered_sentences}

Output format:
For EACH sentence i, output ONE line with:
i<TAB><TAG1>[,<TAG2>,<TAG3>...]

Where each TAG is one of:
{tags_inline}

Important:
- Do NOT rewrite or repeat the sentences.
- Do NOT output any words from the sentences.
- Do NOT output anything except these numbered tag lines.
```



Input: Most of China's crude oil imports go to South Arabia. Least of imports go to Brazil, Kongo, Kazakshtan and South Sudan. The top 3 countries for crude oil import from China are South Arabia, Angola and Russia.

Output: <MAX_EXTREME> Most of China's crude oil imports go to South Arabia. <MIN_EXTREME> Least of imports go to Brazil, Kongo, Kazakshtan and South Sudan. <STATISTIC> The top 3 countries for crude oil import from China are South Arabia, Angola and Russia.

Table 13: The top row shows the prompt used for tagging bar charts in the VisText dataset (Tang and Boggust, 2023). The bottom row shows an input alt text and output alt text with tags for a corresponding chart image.

Structural Clarity Evaluation Prompt	
L1 level	**L2L3 level**
You are an unbiased and fair evaluator. Your task is to evaluate **STRUCTURAL CLARITY**. Structure clarity refers to how well the description organizes information for easy understanding, including: - Logical ordering of information - Grouping of related information into coherent units or sentences - Smooth progression of ideas without abrupt jumps - Clear separation of distinct concepts or visual elements The description may contain different structures of information depending on the chart type. For **Bar Charts**, the structure may follow: - The chart type - The chart title, if present - The label and value range of the horizontal axis - The label and value range of the vertical axis For **Line Charts**, the structure may follow: - The chart type - The chart title, if present - The label and value range of the horizontal axis - The label and value range of the vertical axis For **Pie Charts**, the structure may follow: - The chart type - The chart title, if present - The category labels of the segments Rate the description on a scale from 1 to 5: 1 = very poor structure (confusing order, scattered ideas, no grouping) 2 = weak structure (some grouping but inconsistent or unclear flow) 3 = acceptable structure (generally understandable but imperfect organization) 4 = good structure (clear grouping and logical progression with minor issues) 5 = excellent structure (well-organized, highly logical, and easy to follow Chart type: {chart_type} Description: {generated_text} Output format: Score:	You are an unbiased and fair evaluator. Your task is to evaluate **STRUCTURAL CLARITY**. Structural clarity refers to how well the description organizes information for easy understanding, including: - Logical ordering of information - Grouping of related information into coherent units or sentences - Smooth progression of ideas without abrupt jumps - Clear separation of distinct concepts or visual elements The description may contain different structures of information depending on the chart type. For **Bar Charts**, the structure may follow: - The category with the highest value - The category with the lowest value - Important category values or comparisons between categories For **Line Charts**, the structure may follow: - The overall direction or trend of a line - Significant short-term trend of a line - Peak values - Lowest values For **Pie Charts**, the structure may follow: - The category with the largest percentage - The category with the smallest percentage - Major proportions or percentage distributions Rate the description on a scale from 1 to 5: 1 = very poor structure (confusing order, scattered ideas, no grouping) 2 = weak structure (some grouping but inconsistent or unclear flow) 3 = acceptable structure (generally understandable but imperfect organization) 4 = good structure (clear grouping and logical progression with minor issues) 5 = excellent structure (well-organized, highly logical, and easy to follow Chart type: {chart_type} Description: {generated_text} Output format: Score:

Table 14: Prompt for evaluating structural clarity of L1- and L2L3-level alt text. The left column shows the prompt for the L1 level, the right column shows the prompt for the L2L3 level. The primary difference lies in the information structure for each level.

Information Concentration Evaluation Prompt	
L1 level	**L2L3 level**
You are an unbiased and strict evaluator. Your task is to evaluate **INFORMATION CONCENTRATION**. Information concentration measures how well the description focuses on the key visual in the chart, and avoids unnecessary or low-importance details. Important information depends on chart type and must be inferred from the IMAGE. Key Information Definition: For **Bar Charts**, important information may include: - The chart type - The chart title, if present - The label and value range of the horizontal axis - The label and value range of the vertical axis For **Line Charts**, important information may include: - The chart type - The chart title, if present - The label and value range of the horizontal axis - The label and value range of the vertical axis For **Pie Charts**, important information may include: - The chart type - The chart title, if present - The category labels of the segments Rate the description on a scale form 1 to 5: 1 = focuses mostly on irrelevant or minor details 2 = weak focus, many low-importance details 3 = mixed focus, some key info but noticeable noise 4 = mostly focused on key information 5 = highly focused on key information only Chart Type: {chart_type} Generated Description: {generated_text} Output format: Score:	You are an unbiased and strict evaluator. Your task is to evaluate **INFORMATION CONCENTRATION**. Information concentration measures how well the description focuses on the key visual in the chart, and avoids unnecessary or low-importance details. Important information depends on chart type and must be inferred from the IMAGE. Key Information Definition: For **Bar Charts**, important information may include: - The category with the highest value - The category with the lowest value - Important category values or comparisons between categories For **Line Charts**, important information may include: - The overall direction or trend of a line - Significant short-term trend of a line - Peak values - Lowest values For **Pie Charts**, important information may include: - The category with the largest percentage - The category with the smallest percentage - Major proportions or percentage distributions Rate the description on a scale from 1 to 5: 1 = focuses mostly on irrelevant or minor details 2 = weak focus, many low-importance details 3 = mixed focus, some key info but noticeable noise 4 = mostly focused on key information 5 = highly focused on key information only Chart Type: {chart_type} Generated Description: {generated_text} Output format: Score:

Table 15: Prompts for evaluating information concentration of L1- and L2L3-level alt text. The left column is for the L1 level, and the right column is for the L2L3 level. The primary difference lies in the information focus for each level.

	Reference-based Metrics				Reference-free Metrics		SC		IC	
System	**BLEU-4** ↑	**ROUGE-L** ↑	**TER** ↓	**SBERT** ↑	**ChartVE** ↑	**CLIPScore** ↑	**Gemma** ↑	**Llama** ↑	**Qwen** ↑	**Phi** ↑
L1_B	0.43 ± .026	0.62 ± .019	50.16 ± 2.24	0.90 ± .011	**0.52*** ± .046	**0.40*** ± .007	3.04 ± .061	4.26 ± .100	3.54 ± .073	3.26 ± .596
L1_L_S	**0.61*** ± .021	**0.78*** ± .015	**30.69*** ± 2.29	**0.95*** ± .009	0.25 ± .036	0.39 ± .006	**3.38** ± .085	**4.30** ± .107	**3.72*** ± .065	**3.40** ± .635
L2L3_B	0.03 ± .013	0.27 ± .022	**92.28*** ± 5.86	0.60 ± .029	**0.60*** ± .051	0.35 ± .008	1.68 ± .152	**3.13** ± .180	3.14 ± .059	1.22 ± .000
L2L3_L_S	**0.04*** ± .015	**0.28** ± .022	96.88 ± 4.73	**0.61*** ± .030	0.47 ± .026	**0.36*** ± .008	**2.01** ± .242	3.06 ± .180	**3.23*** ± .067	**2.02*** ± .523
L1L2L3_B	0.32 ± .019	0.53 ± .015	60.34 ± 1.65	0.90 ± .012	**0.31*** ± .040	**0.40*** ± .007	3.12 ± .085	**4.12** ± .067	**3.97*** ± .024	**3.23** ± .690
L1L2L3_L_S	**0.45*** ± .015	**0.64*** ± .013	**48.10*** ± 1.73	**0.93*** ± .010	0.15 ± .026	0.39 ± .006	**3.39*** ± .132	4.07 ± .096	3.90 ± .046	3.20 ± .690

Table 16: Paired bootstrap resampling test results for the VisText dataset. Values are reported as mean ± 95% confidence intervals. Best-performing results are shown in bold, and statistically significant results are marked with *.

	Reference-based Metrics				Reference-free Metrics		SC		IC	
System	**BLEU-4** ↑	**ROUGE-L** ↑	**TER** ↓	**SBERT** ↑	**ChartVE** ↑	**CLIPScore** ↑	**Gemma** ↑	**Llama** ↑	**Qwen** ↑	**Phi** ↑
L1_B	0.29 ± .045	0.62 ± .041	58.82 ± 5.39	0.88 ± .023	0.16 ± .040	0.36 ± .011	**3.33** ± .125	4.08 ± .107	3.85 ± .090	2.25 ± 1.250
L1_L_S	**0.31** ± .048	**0.65** ± .040	**56.44** ± 5.08	**0.89** ± .021	**0.18** ± .049	0.36 ± .012	3.28 ± .125	**4.10** ± .107	**3.87** ± .090	**2.40** ± 1.000
L2L3_B	0.15 ± .041	0.41 ± .038	84.31 ± 7.77	0.77 ± .032	**0.26** ± .062	**0.34** ± .013	2.86 ± .096	3.82 ± .167	3.79 ± .123	**2.43** ± 2.000
L2L3_L_S	**0.18*** ± .042	**0.47*** ± .042	**75.06*** ± 6.78	**0.78** ± .035	0.25 ± .050	0.33 ± .012	**2.91** ± .067	**4.03** ± .222	**3.89*** ± .082	1.50 ± 2.000
L1L2L3_B	0.20 ± .031	0.48 ± .026	74.13 ± 4.66	0.87 ± .023	**0.14*** ± .062	0.37 ± .013	3.79 ± .115	4.13 ± .173	4.00 ± .000	**2.05** ± 1.000
L1L2L3_L_S	**0.25*** ± .030	**0.53*** ± .034	**67.50*** ± 4.82	**0.88** ± .023	0.13 ± .034	0.37 ± .012	**3.84** ± .103	**4.48*** ± .143	4.00 ± .000	1.67 ± 1.000

Table 17: Paired bootstrap resampling test results for the Chart-Alt dataset. Values are reported as mean ± 95% confidence intervals. Best-performing results are shown in bold, and statistically significant results are marked with *.