HCC: A Hierarchical Chart Captioning Model for Enhanced Accessibility of Chart Data for Visually Impaired Users

Yoojeong Song[1], Kanghyeon Seo[2], Svetlana Kim[3], Joo Hyun Park[4]
Department of Computer Software Engineering, Soonchunhyang University, Asan, Republic of Korea[1]
Department of Artificial Intelligence, Republic of Korea Naval Academy, Chanwon-si, Republic of Korea[2]
Department of AI Convergence and Engineering, Open Cyber University of Korea, Seoul, Republic of Korea[3]
ICT Convergence Research Institute, Sookmyung Women's University, Seoul, Republic of Korea[4]

***Abstract*—In educational settings, charts and graphs are commonly used to convey complex information in a simple and understandable manner. However, these visual representations often present accessibility challenges regarding Accessibility for Visually Impaired users, as they cannot be directly interpreted by screen readers without proper alternative text. This paper proposes a novel hierarchical captioning model (HCC: Hierarchical Chart Captioning) designed to facilitate effective Chart Interpretation. The model utilizes spatial token features to generate captions at multiple levels, each offering varying degrees of detail and abstraction, mimicking human cognitive processing. Three hierarchical levels are developed: *Level 1* offers basic and factual descriptions, *Level 2* presents more detailed information, and *Level 3* provides intuitive interpretations and inferences. By integrating a fine-tuned Transformer Models, this approach ensures efficient caption generation and supports user-selectable caption lengths. The model's effectiveness is evaluated through user surveys involving 20 instructors, confirming that Level 2 captions provide the most comprehensible descriptions. Experimental results demonstrate that the proposed method outperforms existing captioning approaches, improving both the efficiency and accessibility of educational materials for visually impaired students. These findings highlight the potential of hierarchical learning models to create more inclusive and accessible educational experiences.**

***Keywords*—Hierarchical captioning; accessibility for visually impaired; chart interpretation; transformer models**

I. Introduction

Various chart materials, including graphs, are used in educational curricula. These charts and graphs are presented in a visual format and are an effective means of conveying complex information in a simple and understandable way. However, materials presented in visual form cannot be directly conveyed to individuals with visual impairments, as they are transmitted in text form through services such as screen readers [1], [2]. For visually impaired learners, the inability to directly perceive these visual materials not only limits their comprehension during lessons but also creates a persistent accessibility gap in real-world educational settings, reducing their opportunities for equal participation and achievement.

First, screen readers deliver content composed of text and images to people with visual impairments in an audio form. For graphs and charts, typically provided as images, instructors or content creators must manually input alternative text. Without it, visually impaired users cannot access the image content [3], [4].

Second, in most classes, when graphs or charts appear in lecture materials, instructors often assume that students have already perceived the visual information from the chart and focus on explaining its features. As a result, visually impaired students, lacking basic information about the chart, must rely solely on the instructor's spoken explanation, which limits the completeness of information they receive.

Third, this situation creates additional challenges for instructors when preparing lecture materials, as they must manually provide explanations for every graph and chart. The laborious task of setting alternative text often leads to omissions, resulting in images with insufficient descriptions.

The visually impaired learners, thus, struggle to access information through visual materials, and lack equal educational opportunities compared to other learners. The need to interpret images such as charts used in educational materials is critical to improving the quality of education for people with disabilities. While charts and graphs help to deliver and understand information quickly, for individuals with disabilities, accessing data in this visual form may be restricted. Therefore, research that helps individuals with disabilities easily access and interpret chart and graph information carries significant social value and meaning. To achieve this, this paper proposes a new methodology for generating image descriptions for individuals with disabilities. The approach leverages the spatial features of tokens—small image patches extracted from the chart that preserve both their content and position within the overall layout. By analyzing these tokens not only for their visual appearance but also for their location and relative arrangement, our method captures the structural context of the chart, which is essential for generating accurate and context-aware descriptions across multiple layers. This approach leverages hierarchical structures divided into four layers based on principles of human cognitive processing—progressing from basic visual perception to higher-level abstraction. Each layer corresponds to a distinct stage in how humans interpret visual information: 1) recognizing simple visual elements, 2) identifying relationships between elements, 3) summarizing overall patterns or trends, and 4) inferring contextual or conceptual meaning. Captions at each level vary in length and detail to align with different user needs and preferences. This method is referred to as

HCC: Hierarchical **C**hart **C**aptioning, highlighting its focus on providing multi-level descriptions that enhance accessibility for visually impaired users.

For individuals with visual impairments who receive all information through auditory means, only intuitive and concise information will be provided to facilitate their understanding of sound-based data [22]. Existing research on caption generation for charts and graphs primarily focuses on the accuracy of the generated captions, typically evaluated by measuring the similarity between human-generated and machine-generated captions. For instance, models such as PReFIL [12] and MATCHA [18] emphasize improving caption accuracy ... without explicitly addressing varying user needs or accessibility for visually impaired users. **The main contributions of this work are as follows:** 1) This study proposes a hierarchical captioning framework that generates captions at multiple levels of detail, enabling user-selectable descriptions tailored for accessibility needs. 2) Principles of hierarchical human cognition are integrated into caption generation, producing captions that balance factual accuracy and interpretability. 3) A user-centered evaluation is conducted to determine the most effective caption level for visually impaired users. Unlike prior works that produce a single, static caption optimized for text-similarity metrics, the proposed method introduces a multi-level captioning framework designed specifically for accessibility. This approach enables users—especially those with visual impairments—to select captions that balance factual accuracy and interpretability according to their needs, an aspect largely overlooked in existing chart captioning research. The novelty lies in integrating hierarchical human cognition principles into caption generation, coupled with a user-centered evaluation to determine the most effective caption level. The primary goal of this study is to develop a hierarchical chart captioning model (HCC) that enhances accessibility for visually impaired users by providing multi-level, user-selectable captions that balance factual accuracy with interpretability. This approach allows for the exploration of the optimal level and length for caption generation for chart images.

Unlike prior works that produce a single, static caption optimized for text-similarity metrics, our method introduces a multi-level captioning framework designed specifically for accessibility. This framework enables users—especially those with visual impairments—to select captions that balance factual accuracy and interpretability according to their needs, an aspect largely overlooked in existing chart captioning research. The novelty lies in integrating hierarchical human cognition principles into caption generation, coupled with a user-centered evaluation to determine the most effective caption level. The primary goal of this study is to develop a hierarchical chart captioning model (HCC) that enhances accessibility for visually impaired users by providing multi-level, user-selectable captions that balance factual accuracy with interpretability.

II. Related Research

A. Image Captioning Datasets

Datasets related to chart image interpretation and captioning have expanded from Question-Answering datasets [5], [6] to Image-caption pair datasets [7], [8]. Compared to large-scale natural language processing datasets [9], [10], the number of image captioning-related datasets is relatively small, limiting the use of large datasets like Transformer models [11]. PReFIL [12] has been proposed as a solution for Question-Answering data and has become a mainstream model for chart explanations. However, issues such as the complexity of Question-Answering data remain unresolved. Most natural language processing models primarily utilize the powerful capabilities of Transformers. These efforts have led to the successful integration with Large Language Models (LLMs) [13], [14], but practical applications of chart explanation models still focus mainly on their relation to labels. To build an effective caption generation model for charts, an approach based on human cognition is needed to facilitate smooth information delivery. Additionally, while image-based data should be factually accurate, overly narrow information delivery should be avoided.

B. Attention Mechanism for Natural Language Process

The attention mechanism primarily works through three components: K (key), V (value), and Q (query). Its main goal is to recognize information related to the input data and understand dependencies within sequences. This attention mechanism is used in various artificial intelligence (AI) models, primarily in encoder-decoder architectures. For example, GPT (Generative Pre-trained Transformer) [15], which is based on a decoder structure, utilizes attention mechanisms to perform unidirectional learning. This model calculates attention for input sequences and generates subsequent words. In contrast, BERT (Bidirectional Encoder Representations from Transformers) [16], based on an encoder structure, performs bidirectional learning by applying attention to both directions of the input sequence to understand context and learn effective representations. These models use the attention mechanism to dynamically detect and process relevant information in its input data, demonstrating excellent performance in natural language processing and other AI tasks.

C. Deep Learning-based Image Captioning Models

Image captioning has drawn significant attention with the advancement of deep learning techniques. Models that generate captions for images combine visual feature extraction and natural language processing. The most commonly used deep learning-based image captioning models are those that combine Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) architectures. CNN models such as VGG [26] and ResNet [27] extract visual features from images, which are then used as input for natural language generation models. After obtaining the visual features of the image, an RNN-based model generates sentences, and model performance is evaluated using automatic metrics (such as BLEU [24], METEOR [25]) and human evaluation. Deep learning-based image captioning models learn meaningful relationships between images and text and are utilized to generate descriptions for images. However, most existing caption generation models focus on creating captions at a single level by considering the overall context of the image. However, people's perception and information processing involve multiple stages of abstraction, gradually moving from detailed information to higher-level concepts [19].

Recent large-scale vision-language models, such as BLIP [29] and DePlot [30], have achieved state-of-the-art performance in general image captioning and chart-to-table translation. These models primarily focus on maximizing data extraction accuracy and reasoning capabilities, often treating chart captioning as a "one-shot" generation task optimized for text-similarity metrics (e.g., BLEU, CIDEr). However, from an accessibility perspective, these approaches exhibit a critical limitation: they produce a single, static output that fails to account for the cognitive constraints of visually impaired users receiving information via audio.

In an auditory environment, presenting a lengthy, unstructured description generated by these models can induce cognitive overload, making it difficult for users to grasp the overall context before diving into details. Unlike these general-purpose models, the HCC model is fundamentally distinct in its architectural assumption and task formulation. Chart captioning redefined not as a static translation task but as a user-centric, hierarchical scaffolding process. By explicitly modeling human cognitive stages—from fragmented perception to concrete interpretation—HCC offers variable-level captions that allow users to control the granularity of information. This "accessibility-first" design ensures that the generated captions are not just factually accurate, but also cognitively digestible for screen reader users, addressing a dimension of usability that standard vision-language models overlook.

III. Hierarchical Chart Caption Generation Model

A. Model Design for Hierarchical Chart Extraction

The goal of this section is to present the design and methodology of our proposed Hierarchical Chart Captioning (HCC) model, which generates multi-level captions for chart images to enhance accessibility for visually impaired users.In this work, the term hierarchical captions refers to a structured set of captions describing the same chart image at multiple levels of detail and abstraction, where lower levels focus on basic, factual elements and higher levels provide progressively richer, more interpretive information. The dataset and model architecture used for hierarchical caption extraction are first outlined, followed by the strategy for generating captions at varying levels of detail. This study is guided by the following research question: *"Can a transformer-based captioning model, when fine-tuned to generate descriptions at multiple hierarchical levels, improve the comprehensibility and accessibility of chart data for visually impaired users compared to single-level captioning approaches?"* The working hypothesis is that generating captions across multiple levels—aligned with human cognitive processing—will provide flexibility for different user needs while preserving factual accuracy and enhancing interpretability.

This study develops a model that generates hierarchical captions for chart images based on the LineCAP dataset [28]. The dataset is a comprehensive collection of annotated line chart images designed to enhance automatic chart analysis and processing in the fields of computer vision and machine learning. It includes various types of line charts with annotations for chart type, labels, axes, and data points, which can be used to extract hierarchical captions. For training the proposed model, pre-existing multi-level captions in the dataset were not relied upon; instead, three separate training subsets—one for each hierarchical level—were constructed by programmatically transforming the original LineCAP annotations. Specifically, Level 1 captions were generated by extracting only the most basic factual elements from the annotations (e.g., axis labels, line identifiers), Level 2 captions incorporated more detailed relationships and trends between elements, and Level 3 captions added shallow inferential statements derived from the annotated data. These level-specific subsets were then used to fine-tune the model in a multi-task learning setup, where each input chart image was paired with three corresponding captions, one per level. Fig. 1 shows the structure of the hierarchical caption generation model based on the Transformer architecture.

This model uses the attention mechanism to focus on the features of the chart image and generate captions focused on three different types, allowing for hierarchical caption generation. The hierarchical generation of captions in this study was achieved by fine-tuning the T5 model [17], a large-scale, multi-purpose pre-trained text-to-text Transformer, with sequence lengths limited to 16, 32, and 64 tokens. These lengths were chosen to balance accessibility for visually impaired users—avoiding overly long captions that hinder auditory processing and overly short captions that omit key details—with empirical stability observed during preliminary experiments. The T5-base configuration was utilized, which contains 12 layers in the encoder and 12 layers in the decoder, each with a hidden size of 768 and 12 attention heads. The base architecture was preserved, but the input embedding layer was modified to accept spatial token features from the chart image encoder instead of standard text embeddings, and the output projection layer was adjusted to generate captions at three predefined sequence lengths corresponding to the hierarchical levels. No other structural changes were made to the Transformer blocks.

Other lengths such as 24 and 48 tokens were also tested, but they provided no significant improvement in comprehension or training stability, leading to the adoption of the 16–32–64 configuration. Multiple experiments were conducted to explore the optimal setup, starting with an initial learning rate of 3e-5, weight decay of 0.01, batch size of 32, and gradient clipping at a maximum norm of 1.0. Hyperparameters were tuned through an iterative manual adjustment process guided by validation loss trends. Common defaults from prior Transformer-based captioning studies were initially employed, after which each parameter—learning rate, weight decay, and batch size—was adjusted individually while monitoring validation loss and caption quality on a held-out set. Configurations yielding the lowest validation loss without overfitting signs were selected as the final values. When validation loss plateaued for three consecutive epochs or increased by more than 10.

Despite the powerful tokenizer and architecture of the original model, instability during training led to adjustments to the learning rate and repeated implementations. If divergence was observed, training parameters were adjusted to prevent information loss due to sequence length limitations. T5, a text-to-text model, uses both the original image caption data and the input sequences limited in length during training. As a

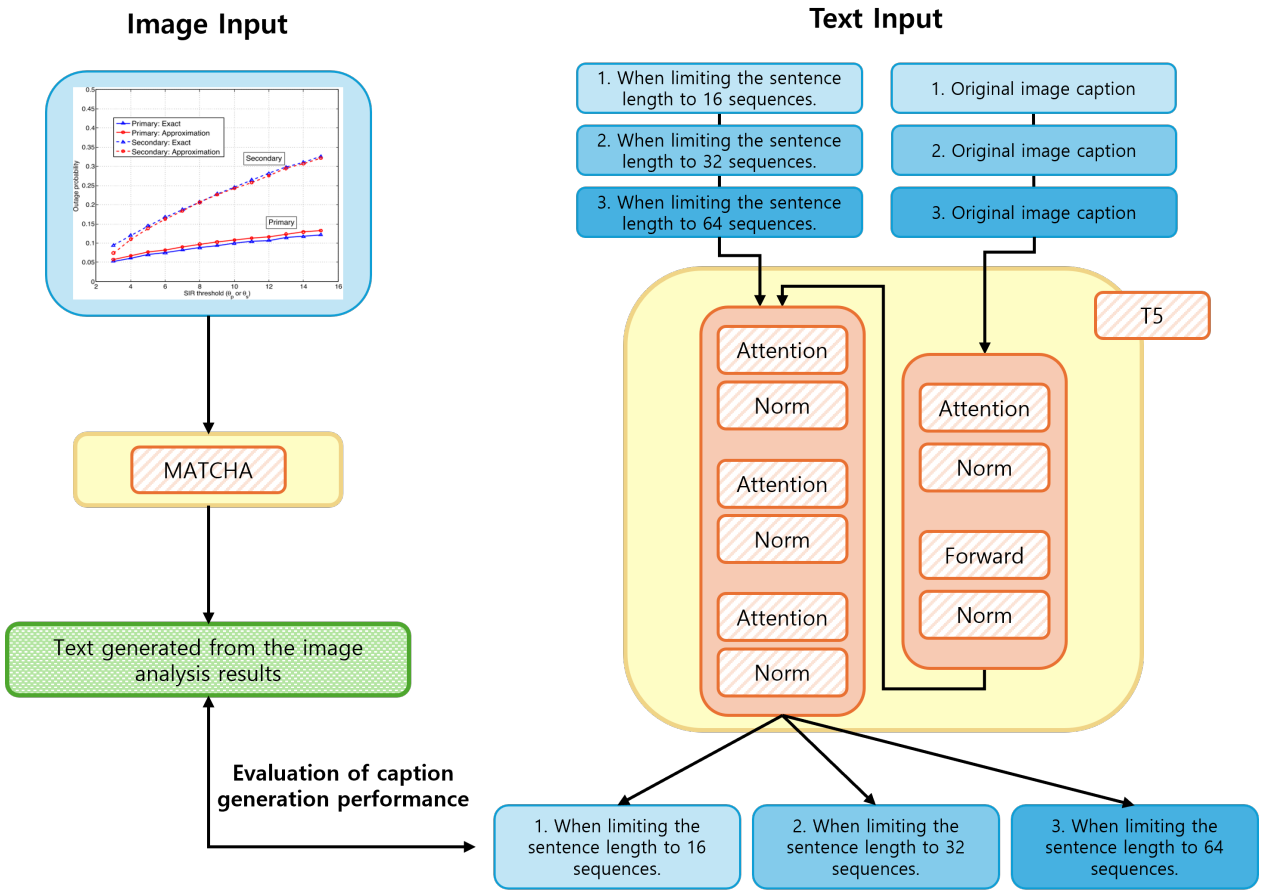


Fig. 1. Overall architecture of the proposed Hierarchical Chart Captioning (HCC) model. The visual encoder (left) processes spatial token features extracted from the input chart image, capturing both visual content and positional structure. These features are fed into the hierarchical caption generator (center, based on the T5 Transformer), where the encoder stack refines the representations and the decoder stack produces captions at three target sequence lengths (16, 32, 64 tokens) corresponding to different abstraction levels (Level 1–3). The encoder and decoder interact via cross-attention, allowing the decoder to selectively attend to relevant encoded features while generating the caption. MATCHA is used as a baseline comparison model, and the generated captions are evaluated for accessibility and comprehension.

result, captions of lengths 16, 32, and 64 are generated as outputs. These outputs are compared with MATCHA's image feature description text for performance evaluation. In this study, MATCHA is used solely as a benchmarking baseline to evaluate the quality of generated captions. It is not involved in the training process of our model. Specifically, after generating hierarchical captions with our fine-tuned T5-based model, The descriptive quality of the generated captions is compared against captions produced by MATCHA, following the same chart image inputs, to provide a consistent reference for performance assessment. MATCHA adopts the same structure as Pix2Struct [18], where the image encoder processes the input image through a series of Transformer layers to extract features, while the text decoder generates token sequences representing the structural layout of the image.

B. Hierarchical Caption Extraction Method

The hierarchical structure of the caption generation model proposed in this study consists of three levels: *Level 1, Level 2,* and *Level 3*, as shown in Fig. 2. This figure illustrates the link between these caption levels and information granularity. Specifically, the graph on the left (PDR% vs. percentage of source nodes) demonstrates how selecting different caption levels allows for balancing interpretability, detail, and performance under varying data conditions. The reason for dividing the caption into three levels is that human information recognition is performed in four stages, and Layers 1–4 correspond to fragmented information, refined information, specific information, and abstract information [19]. Therefore, captions were initially derived in four types to align with these cognitive stages. However, Level 4 captions were excluded from the final analysis because, despite being intended to represent the "abstract" stage of cognition, they exhibited substantial textual redundancy with the human-written reference captions. This meant that Level 4 did not introduce a genuinely new cognitive layer but instead reproduced the reference captions with minimal lexical changes, thereby failing to capture an additional abstraction process beyond Level 3. Quantitative evidence supports this decision: the average BLEU-4 score between Level 4 captions and the references was 0.92, and over 85% of sentences matched almost verbatim. For instance, in many samples, Level 4 captions duplicated entire reference sentences except for minor punctuation differences. Given this lack of conceptual distinction, retaining Level 4 would confound the analysis of hierarchical abstraction. Consequently, the focus

remained exclusively on Levels 1–3, which demonstrated clearer, progressive gradations in detail and interpretability. The captions of Level 1 explain the basic and factual elements of the chart in simple and intuitive language, focusing only on the basic expression of the data. The Level 2 captions provide more detailed expressions than the Level 1, offering information that can be directly extracted from the chart. Lastly, the captions of the Level 3 add interpretations to the chart by providing shallow inferences based on intuitive facts. This hierarchical expression structure expands the model's

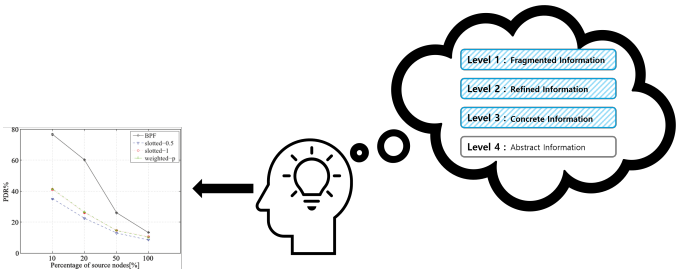


Fig. 2. Hierarchical caption levels in the proposed HCC model and their link to information granularity.

practical application in chart explanation, reflecting an attempt to reflect subjective human interpretation. This is a strategic approach to appropriately interpret the complexity of data visualization and offers deeper insights. In this study, the sequence length was limited to 16, 32, and 64 during training to maintain model learning efficiency while considering the necessary information to generate appropriate captions related to the input image.

In the context of the attention mechanism in neural networks, computational complexity is closely related to the number of tokens (n). In our model, controlling n is directly tied to the sequence length settings (16, 32, 64 tokens) used for different hierarchical caption levels. This allows us to balance descriptive detail with computational feasibility for accessibility-focused caption generation. Despite not performing computations for all token pairs, the computational burden increases proportionally to the square of n, as described in Eq. (1).

$$T(n) = O(n^2 \cdot d) \tag{1}$$

Moreover, the model's processing capacity in the input space increases linearly with n, as outlined in Eq. (2), where c represents a model-dependent constant. In practical terms, this reflects how increasing caption length expands the model's ability to capture detailed chart features, but at the cost of higher processing demands.

$$T(n) = c \cdot n \tag{2}$$

Considering inductive bias within the framework of Ridge Regression [23], the cost function (J) can be formulated by setting the attention mechanism as the hypothesis function.Here, inductive bias is interpreted as a form of regularization that constrains the model's attention weights when generating captions of varying lengths. When computing dependencies between text sequences x and y, the parameters (θ) and the regularization parameter (λ) can be derived from the number of training samples (m) and the number of features (n). By controlling n through sequence length limits, consequently, the inductive bias is implicitly regulated, which helps maintain stable learning across hierarchical levels without overfitting to any single caption length. This suggests that by controlling the number of tokens (n) within a constrained space of m, x, and y, the inductive bias can be effectively regulated [Eq. (3)].

$$J(\theta) = \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2 + \lambda \sum_{j=1}^{n} \theta_j^2 \tag{3}$$

IV. Experiment and Evaluation

A. Experiment and Results

The primary objective of our experiments is to comprehensively evaluate the proposed HCC model in terms of four key aspects: 1) its ability to control the input space while preserving the intrinsic properties of the attention mechanism, 2) its usability for visually impaired users through clarity-focused caption generation, and 3) the validity and effectiveness of the newly introduced performance metric M.

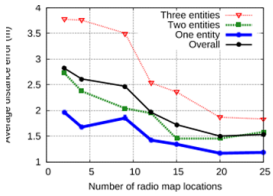
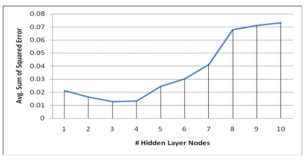
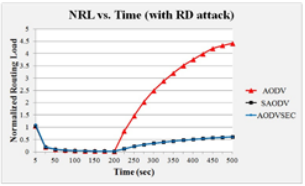
This paper proposes a model that can effectively control the input space while preserving the intrinsic properties of the attention mechanism. To develop this model, a pre-trained T5-base Transformer model [17] is utilized, originally trained on the Colossal Clean Crawled Corpus (C4) dataset [20] containing 60,000 tokens.

All encoder and decoder layers were fine-tuned, while only the input embedding layer was modified to accept spatial token features from the chart image encoder and the output projection layer was adjusted to support multiple sequence length targets (16, 32, 64 tokens) corresponding to the proposed hierarchical captioning levels. This fine-tuning was necessary to adapt the model from its original text-to-text objective to the chart captioning task, ensuring it could effectively map spatial visual features to multi-level textual descriptions.

For fine-tuning, the learning rate was set to 3e-5, the batch size to 32, and the model was trained for 20 epochs using the AdamW optimizer with a weight decay of 0.01. These parameters were selected based on preliminary experiments to ensure stable convergence and reproducibility. Cross-entropy loss was utilized as the objective function. Prior to training, all chart images were resized to 512×512 pixels, normalized to the range [0, 1], and converted into spatial tokens via patch embedding. The LineCAP dataset was split into 80% training, 10% validation, and 10% testing sets, ensuring that charts from the same source did not appear in multiple splits to prevent data leakage. All experiments were implemented in Python 3.10 using PyTorch 2.1.0 and HuggingFace Transformers 4.36.2, with supporting libraries including NumPy 1.26.4, OpenCV 4.8.1, and scikit-learn 1.3.2. Training was performed on Ubuntu 22.04 with an NVIDIA RTX 3090 GPU (24 GB VRAM) and CUDA 12.1.

This extensive dataset ensures that the model can recognize and represent complex patterns and relationships in the data. As shown in Table I, hleach hierarchical level offers

TABLE I. COMPARISON OF HIERARCHICAL CAPTIONS GENERATED BY HCC MODEL AND ORIGINAL CAPTIONS

Chart ID	Chart Image	Captions
1	 Original Caption: Lines 2 and 4 intersect a total of three times. They first intersect at xlabel 12, then at xlabel 20, and lastly at xlabel 25.	**Layer 1**: Lines 2 and 4 intersect at xlabel 12, then at xlabel 20, and lastly at x. **Layer 2**: Lines 2 and 4 intersect at xlabel 12, then at xlabel 20, and lastly at x. **Layer 3**: Lines 2 and 4 intersect three times at xlabel 12, then at xlabel 20, and lastly at x.
2	 Original Caption: ylabel at first decreases at a decreasing rate, then increases roughly at an increasing rate, and then increases at a decreasing rate.	**Layer 1**: The line at first decreases, then increases. **Layer 2**: The line at first decreases at a decreasing rate, then increases at an increasing rate. **Layer 3**: The line at first decreases at a decreasing rate, then increases roughly at an increasing rate.
3	 Original Caption: All lines share the same trend from xlabel 5 to 200. Lines 2 and 3 even share the same trend for all values of xlabel.	**Layer 1**: All lines have the same trend from xlabel 5 to 200. **Layer 2**: All lines have the same trend from xlabel 5 to 200. **Layer 3**: All lines have the same trend from xlabel 5 to 200. Lines 2 and 3 share the same trend for all xlabel values.

a distinct balance between brevity and detail. Specifically, Layer 1 captions provide concise factual descriptions, Layer 2 adds moderate detail for improved clarity, and Layer 3 incorporates the most comprehensive explanation, sometimes adding interpretive insights. Layer 2 captions, for example, provide sufficient context without excessive complexity, which aligns with our usability study finding that Layer 2 is the most comprehensible for visually impaired users. This supports our claim that an intermediate level of abstraction maximizes accessibility. It is noteworthy that the captions at each layer convey the same information from diverse perspectives, and all of the explanations highlight the tendency of the third and fourth lines to be similar. Overall, in Layer 1, it is challenging to discern the extent of data representation in the chart due to the uniformity of the lines in their description of the trend. Layer 2 offers slightly more descriptive capabilities than Layer 1. Layer 3 augments the information in Layer 2, incorporating an explanation of the data represented by the chart.

To apply the generated captions as descriptions for chart materials provided to individuals with visual impairments, usability evaluation is crucial. To evaluate the proposed hierarchical chart caption generation model, a within-subjects survey was conducted with 20 actual instructors. The participants rated three caption layers for five samples, including those shown in Table I, using a 5-point Likert scale. The remaining samples were randomly selected from the LINECAP dataset. Each respondent was tasked with selecting the most comprehensible caption from the provided options for each layer, with a 5-point Likert scale utilized to assess the perceived clarity of each caption. The 5-point Likert scale was composed of 1 = Very Dissatisfied and 5 = Very Satisfied. The results of this survey indicated that the captions generated by Layer 2 seem to be the most comprehensible, as shown in Fig. 3. Fig. 3 presents a graph showing the average understanding score for each caption layer, evaluated on a 5-point Likert scale.

According to Fig. 3, the captions generated by the Layer 3 were also rated as highly relevant, while Layer 1 had the lowest score (Layer 2=3.35, Layer 3 = 2.88, Layer 1 =2.75). In addition, an analysis of variance (ANOVA) was performed to determine if the mean comprehension scores between each layer were statistically significant. The P value of 0.019 confirmed that the mean comprehension scores between the groups were different and statistically significant. In fact, the Diagram Center Image Description [21], which is used by the U.S. Department of Education as a guideline for writing descriptions to make visuals accessible and usable by students with disabilities, suggests that when writing descriptions for diagrams, the captions should be minimal and informative such as the captions in the layer 2. The hierarchical captions generated by this model not only automatically generates

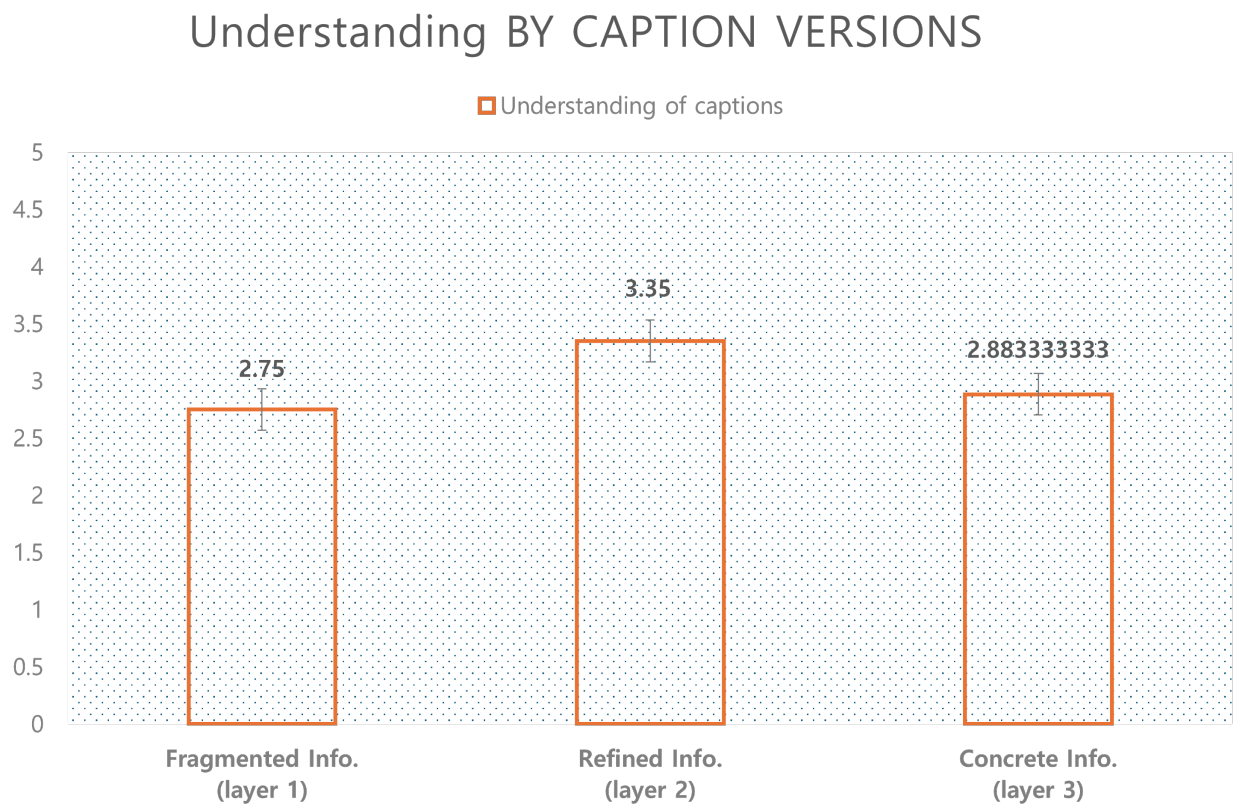


Fig. 3. Graph showing the average understanding score for each caption layer, evaluated on a 5-point Likert scale.

captions when an image is entered, but also automatically assigns the layer 2 captions selected by the usability evaluation so that it can be entered.

B. Caption Generation Performance Evaluation

Before introducing the new metric M, the data preparation process used for this evaluation is first described. The datasets LineCAP, DVQA, and FigureQA were selected because they provide diverse chart types and high-quality annotations necessary for training a robust chart captioning model, covering both low-level data extraction and high-level reasoning tasks. For each of these datasets, the same preprocessing pipeline was applied as in model training: all chart images were resized to 512×512 pixels, normalized to the range [0,1], and converted into spatial tokens via patch embedding. For each test set, hierarchical captions (Levels 1–3) were generated using the fine-tuned HCC model. The corresponding attention masks from the decoder cross-attention layers were stored to enable the computation of SV_{am} in the metric formula. No training or validation samples were included, ensuring that all metric computations were performed on unseen data to maintain evaluation integrity.

Traditional captioning metrics such as BLEU or CIDEr are designed to reward lexical overlap between a generated caption and a single reference caption. In our hierarchical setting, captions at different levels (e.g., Level 1 vs. Level 3) are intentionally diverse in length, detail, and abstraction, meaning that a low lexical match does not imply poor quality. For example, a concise Level 1 caption can be perfectly accurate yet achieve a low BLEU score simply due to its brevity. Similarly, CIDEr's term-frequency weighting favors n-gram overlap, which biases against higher-level captions that incorporate interpretive language absent from the reference. These metrics are therefore discarded because they conflate intended stylistic variation with error, and metric M is adopted instead to directly assess feature preservation and attention distribution properties that remain meaningful across levels

Therefore, the objective of this research is not to ensure high performance on the correct answer labels generated by the model, but rather to prioritize stable learning through caption generation to enhance the practical utility of the model. To this end, the following metric was established, given in Eq. (4).

$$M = \cos(f_\theta(x), y) \cdot k + \sigma(SV_{am}) \tag{4}$$

This metric comprises two terms. The first utilizes cosine similarity to ensure that the model's output preserves the features of the input tokens without compromising their characteristics. Here, "input" refers to the vectorized embedding representation of the spatial tokens extracted from the chart image encoder, and "output" refers to the embedding representation of the generated caption tokens produced by the decoder. Both are mapped into the same latent space using the final projection layer before similarity computation. This ensures that cosine similarity measures alignment between semantic feature vectors rather than raw token IDs or unprocessed

TABLE II. EVALUATION OF CAPTION GENERATION PERFORMANCE ACROSS THREE DATASETS (LINECAP, DVQA, FIGUREQA) USING THE PROPOSED METRIC M. HIGHER SCORES INDICATE BETTER SEMANTIC FEATURE PRESERVATION AND BALANCED ATTENTION FOCUS. 'M(OURS)' REPRESENTS THE PROPOSED HCC MODEL, WHILE 'M(NEW BASELINE)' DENOTES A STANDARD FINE-TUNED T5 MODEL

M/dataset	**LineCAP**	**DVQA**	**FigureQA**
Existing performance evaluation metrics	**BLEU-4/CIDEr**	Structure/Data/Reasoning	Accuracy
Existing evaluation metrics results	0.418/1.096	94.47/65.40/44.03	72.54

TABLE III. COMPARE THE NEW "M (BASELINE)" RESULTS WITH OUR "M (OURS)" RESULTS

M/dataset	**LineCAP**	**DVQA**	**FigureQA**
M(New baseline)	15.98	13.24	16.98
M(Ours)	10.92	9.84	7.34

sequences, avoiding ambiguity between "input tokens" and "output tokens."

The second term examines the standard deviation of the singular values (SV_{am}) of the attention mask. SV_{am} denotes the set of singular values obtained by applying Singular Value Decomposition (SVD) to the attention mask matrices extracted from the decoder's cross-attention layers. For each generated caption, the attention mask $A \in \mathbb{R}^{n \times n}$ is collected, where n is the sequence length, and SVD is performed: $A = U\Sigma V^\top$. The diagonal entries of Σ represent the singular values, which capture the distribution of energy across attention components. Singular values are aggregated across all heads and layers for each caption, and the standard deviation is computed to quantify the variability in attention focus. This process is repeated for all samples in the test set to produce the final SV_{am} value used in the metric M. Rather than implying that lower variance always equates to better expressiveness, a moderate reduction is interpreted as desirable because it suggests a balanced allocation of attention that preserves relevant structural information while avoiding noise-induced dispersion.

The earlier discussion of attention complexity in Eq. (1) and (2) directly informs the interpretation of metric M. Controlling the sequence length n in the hierarchical captions not only constrains the $O(n^2 \cdot d)$ computational cost but also influences the structure of the attention masks from which SV_{am} is computed. Shorter sequences reduce potential attention dispersion, leading to a more concentrated singular value spectrum, whereas longer sequences increase complexity and variability in attention focus. By incorporating SV_{am} into M, this trade-off between descriptive richness and computational efficiency is captured.

This value $k = 10$ was selected through preliminary sensitivity analysis, where k was varied across 1, 5, 10, 20 and the resulting metric distributions were inspected. It was observed that $k = 10$ provided a balanced scaling between the cosine similarity term and the singular value variability term.

Table II and III demonstrate that our model consistently achieves lower M scores across all datasets compared to the baseline, with reductions of up to 56.8% (FigureQA). Since a lower M indicates better semantic alignment and more balanced attention distribution, these results provide quantitative evidence that the proposed HCC model more effectively preserves key features of the input space while controlling complexity.

Focusing on the results of the newly proposed 'M(baseline)' and our 'M(Ours)', the following observations can be made. In the LineCAP dataset, the baseline achieved 15.98, while our model achieved 10.92. Similarly, for DVQA, the baseline scored 13.24, and our model scored 9.84, while in FigureQA, the baseline result was 16.98, and our model scored 7.34. In all datasets, our model outperformed the baseline. Notably, in FigureQA, the M value decreased by almost half, indicating an improvement in performance.

V. DISCUSSION : ENHANCING ACCESSIBILITY THROUGH USER-COGNIZANT HIERARCHICAL CAPTIONING

The Hierarchical Chart Captioning (HCC) model is fundamentally distinguished from prior works, which primarily focused on maximizing lexical similarity metrics, by aiming to resolve the practical accessibility gap through a user-centered approach. This section discusses the behavioral and cognitive implications of our hierarchical framework.

A. The Mechanism of Hierarchical Captioning in Mitigating Cognitive Load

HCC's effectiveness is rooted in its ability to translate the complex stages of human visual information processing—which progresses from visual perception to higher-level abstraction—into a manageable auditory information delivery structure. This mechanism directly addresses the significant cognitive load associated with processing lengthy, undifferentiated auditory descriptions in non-visual environments.

The multi-level structure segments chart information into ascending cognitive levels: Level 1 (basic facts), Level 2 (relationships and trends), and Level 3 (shallow inferences). This managed segmentation allows users to assimilate the chart data incrementally, supporting a fundamental principle of Universal Access by providing flexible information delivery based on individual processing needs. Furthermore, the approach of offering user-selectable sequence lengths (16, 32, 64 tokens) introduces a critical HCI benefit, empowering the visually impaired user to tailor the information depth to their immediate needs, thereby overcoming the inefficiency of auditory searching through overly long, static narratives.

B. Level 2 as the Optimal Threshold for Information Comprehension

The user evaluation, which involved a survey with 20 instructors, empirically confirmed that the Layer 2 caption

achieved the highest comprehensibility score (3.35), positioning it as the optimal level for driving positive information comprehension behavior. This finding is central to our work, highlighting a critical trade-off between descriptive detail and cognitive feasibility. The inadequacy of extremes was clearly demonstrated, as Level 1 captions, due to their limited and fragmented information, scored the lowest (2.75) because they failed to establish the necessary overall context or data relationships. Conversely, Level 3 captions, while relevant due to the inclusion of inferences, imposed a higher cognitive burden due to their increased length and complexity, resulting in a comprehension score (2.88) lower than Level 2. Level 2 is thus identified as the optimal point of cognitive equilibrium. By detailing the relationships and trends directly extractable from the chart, Level 2 provides sufficient context while judiciously avoiding the excessive complexity and subjective interpretation of Level 3. This balanced delivery strategy aligns precisely with accessibility best practices, such as those recommended by the U.S. Department of Education's Diagram Center, which advocates for minimal yet informative descriptions. The superior performance of Level 2 captions provides strong empirical evidence that effective accessibility solutions must be driven not merely by technological performance metrics, but by an HCI design balance that explicitly accounts for user-specific cognitive capacities to yield successful behavioral outcomes.

Consequently, this finding challenges the prevailing assumption in captioning research that higher text similarity scores (e.g., BLEU) automatically equate to better utility. By confirming that instructors—the domain experts in education—prioritize the 'refined' information of Level 2 over the more exhaustive Level 3, this study provides a concrete HCI design guideline: accessibility tools should prioritize cognitive load management over mere data exhaustion. This shift from technical optimization to human-centered validation is the core contribution of our work, demonstrating that the 'best' model is not the one with the most detailed output, but the one that aligns with the pedagogical needs of the users.

VI. Conclusion

Visually impaired individuals, particularly those who are completely blind, face significant challenges in accessing and understanding descriptive information about graph or chart images. If an image lacks recorded descriptions, screen readers are unable to convey its content; therefore, when a graph or chart image is provided, an accompanying explanation is essential to enhance accessibility for visually impaired users. However, conventional graph captions are often overly detailed, deviating from their original intent and hindering practical usability and scalability. This highlights a critical need to re-examine the purpose of such descriptions by refining the levels of human cognitive processing and applying this to chart explanation methodologies.

This study proposed the HCC (Hierarchical Chart Captioning) model to address this accessibility gap by generating multi-level, user-selectable captions. Unlike traditional captioning methods that provide information at a single layer, the HCC model utilizes spatial token features to generate three hierarchical levels of captions that mimic human cognitive processing. This design ensures that the model provides varying degrees of detail and abstraction, allowing users to select the description length that best suits their needs.The model's effectiveness was rigorously evaluated through user surveys involving 20 instructors, which revealed a critical insight: Layer 2 captions were the easiest to understand, scoring 3.353. This result confirms that the most effective solution lies in achieving an optimal balance between factual detail and cognitive load, aligning with the "minimal yet informative" approach advocated by guidelines like the Diagram Center Image Description4444. Consequently, this finding underscores that technical complexity does not guarantee practical utility. Future accessibility metrics must therefore evolve beyond statistical correlation to prioritize human-centric clarity. By identifying Level 2 as the optimal granularity, this study establishes a practical standard that bridges AI capabilities with real-world educational needs.

Furthermore, the model's performance was validated using a newly introduced evaluation metric, M, which confirmed that the proposed methodology achieved performance levels comparable to existing baseline architectures5. This demonstrates that our approach ensures both reliable performance and effective learning, highlighting its strong potential applicability to hierarchical learning frameworks.

While this study demonstrates the efficacy of the HCC model, several limitations remain. Currently, our validation is primarily focused on line charts; thus, further research is required to assess the model's adaptability to more complex or composite scientific visualizations. Future work will address this by expanding the dataset diversity and conducting in-depth end-user testing directly with visually impaired participants to gather qualitative feedback. Additionally, we plan to explore the real-time integration of our hierarchical captioning system with Text-to-Speech (TTS) interfaces to support live learning environments.

In conclusion, the HCC model offers a novel, user-centered approach to enhancing digital accessibility by translating complex visual data into cognitively manageable auditory descriptions. Ultimately, this work aims to advance Universal Design for Learning (UDL), ensuring that educational insights are equally accessible to every learner regardless of visual ability.

Acknowledgment

The authors would like to acknowledge the assistance of ChatGPT by OpenAI, which was used for English translation and refinement of several sections of this manuscript. This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2023-00211205)

References

[1] Murillo-Morales, T., Miesenberger, K., "Audial: A natural language interface to make statistical charts accessible to blind persons," in *Computers Helping People with Special Needs: 17th International Conference, ICCHP 2020*, Lecco, Italy, 2020, pp. 373–384.

[2] Mukherjee, A., Garain, U., Biswas, A., "Experimenting with automatic text-to-diagram conversion: A novel teaching aid for the blind people," *J. Educ. Technol. Soc.*, vol. 17, no. 3, pp. 40–53, 2014.

[3] Balik, S. P., others, "Including blind people in computing through access to graphs," in *Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility*, 2014, pp. 91–98.

[4] Torres, M., Barwaldt, R., "Approaches for diagrams accessibility for blind people: a systematic review," in *2019 IEEE Frontiers in Education Conference*, IEEE, 2019, pp. 1–7.

[5] Kushal, K., others, "DVQA: Understanding Data Visualizations Via Question Answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5648–5656.

[6] Kahou, S. E., others, "FigureQA: An Annotated Figure Dataset for Visual Reasoning," in *Sixth International Conference on Learning Representations*, 2017.

[7] Anita, M., Kostic, Z., Tanner, C., "Lincap: Line Charts for Data Visualization Captioning Models," in *2022 IEEE Visualization and Visual Analytics (VIS)*, 2022, pp. 35–39.

[8] Hsu, T., Giles, C. L., Huang, T., "SciCap: Generating Captions for Scientific Figures," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 3258–3264.

[9] Devlin, J., others, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint*, 2018.

[10] Yang, Z., others, "XLNet: Generalized autoregressive pretraining for language understanding," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[11] Vaswani, A., others, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[12] Kushal, K., others, "Answering Questions about Data Visualizations Using Efficient Bimodal Fusion," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 1498–1507.

[13] Liu, F., others, "MatCha: Enhancing Visual Language Pretraining with Math Reasoning and Chart Derendering," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023, pp. 12756–12770.

[14] Liu, F., others, "DePlot: One-shot Visual Language Reasoning by Plot-to-table Translation," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 10381–10399.

[15] Luciano, F., Chiriatti, M., "GPT-3: Its Nature, Scope, Limits, and Consequences," *Minds and Machines*, vol. 30, pp. 681–694, 2020.

[16] Devlin, J., others, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *The 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019, pp. 4171–4186.

[17] Colin, R., others, "Exploring the Limits of Transfer Learning with a Unified Text-to-text Transformer," *The Journal of Machine Learning Research*, vol. 21, no. 140, pp. 5485–5551, 2020.

[18] Lee, K., others, "Pix2struct: Screenshot parsing as pretraining for visual language understanding," in *International Conference on Machine Learning*, 2023, pp. 18893–18912.

[19] Elbarougy, R., Akagi, M., "Improving speech emotion dimensions estimation using a three-layer model of human perception," *Acoustical Science and Technology*, vol. 35, no. 2, pp. 86–98, 2014.

[20] Dodge, J., others, "Documenting large webtext corpora: A case study on the colossal clean crawled corpus," *arXiv preprint*, 2021.

[21] Diagram Center Image Description, "Diagram Center Image Description," 2023, Available: http://diagramcenter.org/making-images-accessible.html.

[22] Park, J. H., others, "Mobile Voice Web Browser for the Low Vision," *Journal of Korea Multimedia Society*, vol. 23, no. 11, pp. 1418–1427, 2020.

[23] Arashi, M., Roozbeh, M., Hamzah, N. A., Gasparini, M., "Ridge regression and its applications in genetic studies," *PLOS One*, vol. 16, no. 4, pp. e0245376, 2021.

[24] Papineni, Kishore, Roukos, Salim, Ward, Todd, Zhu, Wei-Jing, "BLEU," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[25] Banerjee, Satanjeev, Lavie, Alon, "METEOR: An automatic metric for MT," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005, pp. 65–72.

[26] Karen Simonyan, Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Deep Residual Learning for Image Recognition," *arXiv preprint arXiv:1512.03385*, vol. , no. , pp. , 2016.

[28] Anita Mahinpei, Zona Kostic, Chris Tanner, "LineCap: Line Charts for Data Visualization Captioning Models," in *2022 IEEE Visualization and Visual Analytics (VIS)*, 2022, pp. 35–39.

[29] Li, J., Li, D., Xiong, C., Hoi, S. C. H., "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pp. 12888–12900, 2022.

[30] Liu, F., et al., "DePlot: One-shot visual language reasoning by plot-to-table translation," *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 10381–10399, 2023.