Bachelor Thesis

Caption Quality Comparison with the DaTikZ Dataset

Jonas Gnauck
(matriculation number 1835105)

June 2, 2025

Submitted to
Data and Web Science Group
Prof. Dr. Steffen Eger
University of Mannheim

Abstract

Scientific figures can convey complex information in research effectively. Many captions in large-scale scientific datasets are vague, incomplete or dependent on contextual knowledge. This thesis addresses this challenge by introducing a structured evaluation framework to assess caption quality and proposing a visually grounded, context-independent approach to caption re-writing. Using a 200-sample subset from the DaTikZ-v3 dataset, original captions were evaluated along multiple dimension, including functional role, structural elements, completeness and usefulness. All captions were manually rewritten using only the content visible within the figures. Semantic similarity analysis confirmed strong alignment among annotators and figure generation experiments using GPT4o-mini showed that revised captions produced significantly more accurate and detailed figures. To assess the scalability of this approach, several multimodal models were evaluated for their ability to generate high-quality captions across the same dataset.

This work highlights that visually grounded captions can substantially improve dataset quality and figure generation outcomes.

Contents

1. Introduction

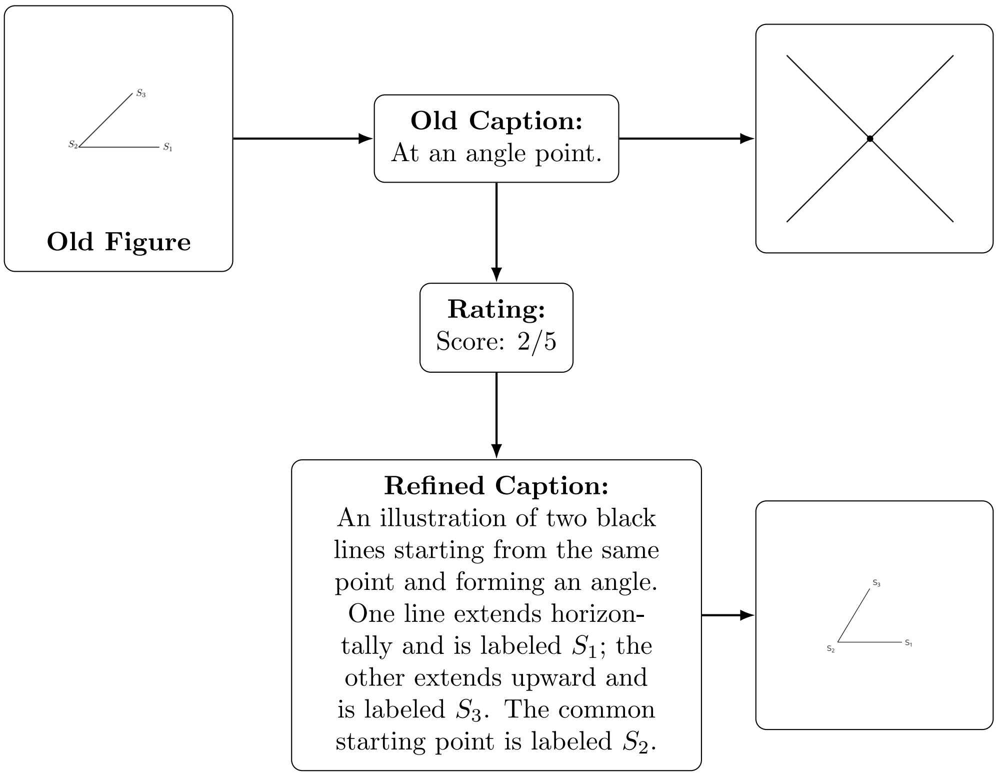


Figure 1.1.: The original figure with it's caption. Our evaluation and our revised description with the resulting generated figure.

To explain complex concepts, experimental setups and results, scientific figures serve as a powerful tool. They can turn large amount of technical detail into visualizations that improve the understanding. Their effectiveness depends not only on the quality of the figure, but also on the quality of their accompanying captions. A well-written caption should help the reader understand the figure more quickly and identify important elements.

Despite their important role, in large-scale scientific datasets, figures are often presented without their surrounding context, leaving only the caption. These captions are often incomplete, vague and rely heavily on contextual knowledge (Yang et al. 2023). This issue is especially problematic in datasets that are used to train and evaluate vision-language models, where insufficient captions can hinder model performances.

This thesis addresses this challenge by proposing and alternative captioning approach, with creating standalone, visually grounded and context-independent descriptions that rely solely on the information provided within the figure itself (Figure 1.1). This removes the dependence on surrounding text or interpretive knowledge and instead focused on detailed, objective description of visible components. This not only improves caption clarity but also enhances the utility of scientific datasets for downstream tasks such as figure generation from text, as demonstrated in this thesis.

To evaluate and improve caption quality, this thesis uses a representative 200 image-caption pair sample from the DaTikZ-v3 dataset (Belouadi et al. 2025). Each caption is assessed by a structured human evaluation framework introduced in this thesis. This framework is designed to capture multiple dimensions of caption quality, including functional role, presence of structural elements, completeness and usefulness. The results of this evaluation reveal lack of detail and clarity in the captions. Multiple annotators were involved in the evaluation process to validate the robustness and calculate Inter-annotator agreement using Cohen's Kappa (Cohen 1960) and Fleiss' Kappa (Fleiss 1971).

To improve the quality of captions, the thesis proceeds to manually rewrite all 200 captions in the sample, by only using the visual content of each figure with no external references, interpretations or speculations. The resulting description are self-contained, grammatically correct and visually complete. This was also validated by the semantic similarity between the new descriptions and those produced by other human annotators.

Both the original and revised captions were used to generate TikZ figures via GPT4o-mini. The generated figures were then compared to the original DaTikZ-v3 ground-truth figures using both human judgment and automated metrics. Results show that figures generated from the revised caption align more closely in structure and content with the originals.

While manual rewriting provides high-quality results, for large-scale dataset this is not scalable. To explore alternatives, this thesis evaluates the performance of multimodal models in generating figure captions. Results suggest that automated caption refinement using large language models is a scalable approach to improve dataset quality.

This work demonstrates that scientific mutimodal dataset, like DaTikZ, lack caption quality and presents a systematic, visually grounded approach to evaluate and enhance figure descriptions. Through human evaluation, manual caption refinement and model-based scalability, this work provides a validated methodology for improving the clarity, usefulness and utility of scientific figure datasets, ultimately enabling better performance in vision-language tasks and more robust figure generation.

2. Related Work

Our work relates to several key areas: evaluation metrics, vision language alignment for caption refinement and scientific captioning. Prior work in each domain is listed below.

2.1. Conventional Evaluation Metrics:

Evaluating the quality of image captions poses a challenge in vision-language research. Early work has been evaluated using metrics that rely heavily on automatic text similarity metrics adapted from machine translation. These metrics measure **n-gram overlaps** between a candidate caption and one or more reference captions. An example is **BLEU** (Papineni et al. 2002), which computes the precision of n-grams in the candidate that appear in the reference. Such overlap-based metrics (Lavie and Agarwal 2007; Vedantam et al. 2015; Anderson et al. 2016) are inexpensive and language independent, but they emphasize exact word match. This means they often fail to capture semantic accuracy or the descriptive richness of a caption when phrasing differs from the references. Existing reference-based measures tend to fall short of providing a complete assessment of caption quality that aligns with human judgment (Chan et al. 2023). For instance, a caption that is factually accurate but uses synonyms or paraphrasing may receive a low BLEU score despite conveying the same meaning as the reference. These limitations underscore the need for more semantically oriented evaluation approaches.

2.2. Vision-Language Alignment Metrics

To address these limitations of overlap metrics, researchers have proposed vision-language alignment metrics that rely on pretrained multimodal models. **CLIPScore** (Hessel et al. 2022) is a reference-free metric which uses a neural network called **CLIP** (Radford et al. 2021) that was trained on image-text pairs to directly measure the semantic similarity between an image and its caption. This rewards caption that are more semantically aligned with the image, if the wording isn't similar to the references. Studies showed that CLIPScore achieves a substantially higher correlation with human judgments of caption quality than reference-based metrics. However, CLIPScore focuses mainly on image-text compatibility and cannot ensure that the caption is contextual or insightful. But the introduction of CLIPScore shifted the focus from textual overlap to visual-semantic alignment when evaluating captions.

2.3. Scientific Captioning

General image captioning methods and metrics may not directly apply to scientific figures, which present challenges. In research paper, figures often contain complex diagrams or graphs. The captions reference to details that are not obvious directly from the figure, such as experimental context or abbreviations defined in the text. Therefore, many scientific figures have captions which are either too generic or need prior knowledge.

SciCap Dataset: In this research SciCap (Hsu et al. 2021) was introduced, which is a large-scale figure-caption dataset. It contains over two million figure-caption pairs extracted from over 290,000 papers and is based on the arXiv dataset. Preprocessing was required to create this dataset: figure-type classification was used to distinguish between figures like graphs, diagrams and tables. Scientific figures containing subfigures were removed from the dataset. Text was converted to lowercase and figure numbers, such as "Figure 1:" or "Fig. 1", which are very common in scientific papers, were removed to keep only the main text. The most common figure type in the dataset are graph plots, which make up around 19% of figures. Captioning models were trained on this specific subset, but only achieved low BLEU scores, indicating that an image-to-caption approach left much room for improvement. This happens because captions often need to convey scientific insights about a figure, which may require understanding trends, reading textual labels, or integrating information from the surrounding text. SciCap found that models trained on general image captions have difficulty with the technical vocabulary and implicit knowledge necessary for scientific images (Hsu et al. 2021).

SciCap+: To tackle the above mentioned challenges, SciCap+ (Yang et al. 2023) was introduced, which is a knowledge-augmented extension of SciCap. It provides additional context for each figure by including the paragraph from the paper that discusses or reference the figure and as well as OCR-extracted text from the figure itself. With these informations, SciCap+ can leverage external knowledge. The results show that providing the context paragraph yields better captions compared to models that only see the figure. This means that having access to the paper context improved the models output, as reflected in higher BLEU, METEOR and other scores (Yang et al. 2023). This supports that many scientific figures are clarified only with the surrounding text. However, even with these improvements human evaluators still identified gaps and errors in the generated caption. This shows how difficult it is to produce highly informative captions for complex scientific figures.

DaTikZ Dataset: The DaTikZ (Belouadi et al. 2024) dataset is a large-scale TikZ dataset of approximately 120k paired TikZ drawings and captions. These are extracted from websites, online repositories, TeX Stack Exchange, arXiv papers and also artificial examples. One approach for improving caption quality was data augmentation. Captions containing fewer than 30 tokens were labeled as poor quality and therefore leveraged into short descriptions by LLaVAR (Zhang et al. 2024). This increased the diversity in the dataset while retaining the original captions as well. It also increased the

CLIPScore notably for these captions from 24.76 to 29.12, even exceeding the score of 27.06 for captions over the 30 token threshold. **DaTikz-v3** (Belouadi et al. 2025) even further expands this approach. It systematically extracts captions alongside its TikZ graphics whenever possible , which results in over 450k instances where approximately 170k include captions.

Unlike SciCap and SciCap+, which primarily focus on the dataset scale and contextual enrichment, this thesis takes a fundamentally different approach to improve caption quality. Rather than relying on indirect signals like caption length or surrounding content, we design and apply a structured human evaluation framework that explicitly measure the captions clarity and structure. SciCap+ improves figure caption by providing context, but does not rewrite the caption itself. Our method includes direct caption rewriting as a core part of the process. These rewritten captions are not only grammatically improved, but explicitly written in a way with no reliance on outside text or assumed knowledge. We focus only on the content visible within the figure. Our approach provides a practical and scalable way to improve the descriptive quality of scientific figures. It complements existing efforts like SciCap, but shifts the focus from context to descriptiveness.

3. Chapter Guidelines

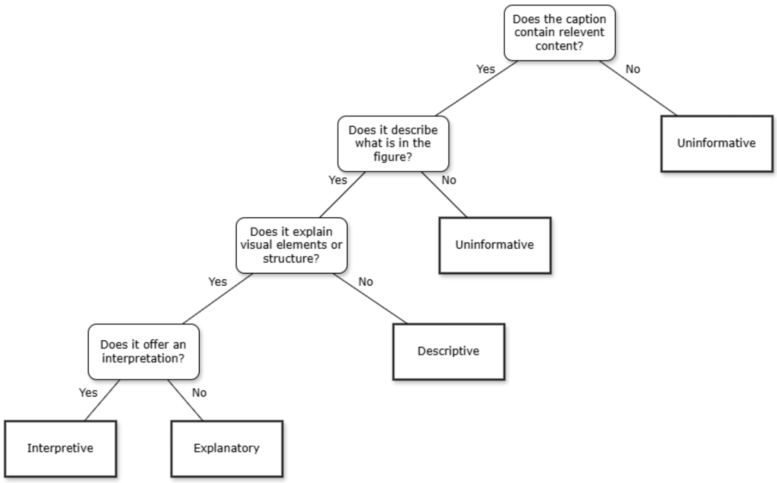


Figure 3.1.: Decision tree for functional role

3.1. Manual Classification Framework

To evaluate each caption with the same guidelines, we developed a manual annotation framework. This framework captures multiple aspects. We examined each caption along four dimension:

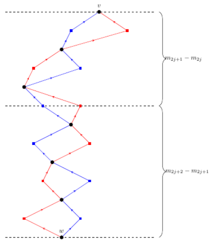
1. its *functional role* in relation to the figure
2. the presence of *structural elements*
3. the caption's *length and completeness* and
4. an overall *quality score.*

Each caption was independent evaluated by multiple annotators. To ensure consistency, annotators were provided with detailed guidelines, that are described below.

Functional Role

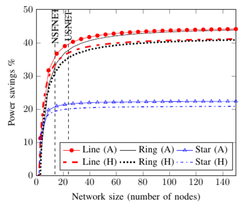
The functional role of a caption refers to its primary purpose. We defined four distinct functional roles for scientific figure captions, adapted from best practices in academic writing and prior research taxonomies (Hsu et al. 2024; Tang et al. 2023). Each caption was assigned one primary role from the following:

Uninformative



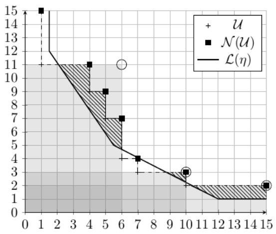
Caption:
"Schematics"

Descriptive



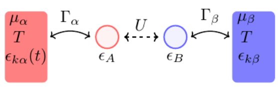
Caption:
"Workflow to obtain a Feasible Planning Region"

Explanatory



Caption:
"The lower bound set L(n), depicted by the solid line, is partially dominated by the upper bound set U, represented by the crosses. In this situation, it is possible to split the problem into three disjoint sub-problems in the objective space. Each sub-problem is highlighted by the hatched areas, and its corresponding super local upper bound is depicted by its closest large circle."

Interpretive



Caption:
"Schematic representation of the model investigated. The two quantum dots are coupled to noninteracting electron reservoirs and coupled to each other by Coulomb interaction. Within the investigation, energies of reservoir A are driven harmonically, resulting in a nonzero current between the dots and reservoirs and energy transfer between the reservoirs."

Figure 3.2.: Examples of the four functional roles assigned to scientific figure captions.

- **Uninformative:** The caption provides little useful information about the figure. Often these captions are too vague, overly brief or entirely off-topic, failing to describe what is shown.

- **Descriptive:** The caption only describes what is shown in the figure, without a deeper explanation or interpretation.

- **Explanatory:** The caption clarifies key visual aspects of the figure and may provide contextual details, but doesn't draw conclusions. It points out important features, trends or relationships shown in the figure, but focuses only on the content itself.

- **Interpretive** The caption provides a interpretation or insight about the figure. It explains the meaning or importance. These captions often convey the figures key message explaining its meaning or importance in context.

Figure 3.1 shows a tree for the functional role decision. Every caption receives a role label best deciding its main function.

Structural Elements

To analyze even more than just the general purpose of the figure, we looked at specific structural elements in each caption. Each caption was broken down into five common structural elements. These are simple yes/no checks, but together they give a detailed breakdown of the captions structure. (Hsu et al. 2024):

- **Figure type:** assesses whether the caption specifies the type of figure presented. This could be something as direct as saying "graph," "tree," "workflow," or any other high-level category.

- **Visual details:** capture whether the caption refers to specific visual elements of the figure. This include references to colors ("red line"), shapes ("triangular region"), axes or how different parts are arranged. This element is important because it shows whether the caption describes the figure visually.

- **OCR:** checks if the caption mentions any text that's visible in the figure itself. That could be axis labels, figure labels or even small annotations of mathematical expressions. Including this kind of content in a caption helps models label the figure correctly.

- **Contextual reference:** looks at whether the caption points to information outside the figure, like "as seen in Section 3" or "using the method described above." While this can be useful in papers, it reduces the stand-alone usefulness of the caption.

Structural Elements Highlighted by Category
• **[Contextual], [Figure Type]**: The *T.3.1 graph* • **[Figure Type], [OCR]**: *Workflow* to obtain a *Feasible Planning Region* • **[Figure Type], [Visuals], [Quantitative]**: A *tree T* on *5 vertices* • **[Visuals], [Quantitative]**: The contracting process. At the end, the intersection number between two *red lines* is $\frac{n-1}{2}$. The intersection number of *red and black lines* is just one. • **[OCR],[Quantitative],** : General system pipeline. W_x is the input to the system and O_x is the output before the pre-training task layer (MLM / NSP). • **[Quantitative], [Visuals]**: Pendulum constrained to $\mathcal{S}^1$ manifold with the tangent space $T_{x_0}M$ *in blue and* T_xM *in red.* • **[Figure Type], [Visuals], [OCR], [Quantitative]**: *Schematic reperesentation* of the two solitary waves R_1 and R_2, of the functions W_1 and W_2, and of the function P with a *plateau between the solitary waves*

Figure 3.3.: Examples of captions containing structural elements, labeled by category: figure-type, Visuals, OCR, contextual and quantitative references.

- **Quantitative content:** checks if the caption includes numbers, formulas, code or anything else that adds mathematical or technical substance to the caption. This often is in combination with OCR.

Figure 3.3 shows examples for each element visible marked with colors. These five structural markers help us examine the kind of information the caption wants to convey and whether it's detailed enough. They're not meant to judge quality by themselves, but they're useful for analyzing what a caption includes and what it might be missing. For each caption, the annotators indicated whether each of the above structural elements was present or absent. In addition to these five attributes, we evaluated how useful each caption was to the annotator.

Usefulness was rated on a three-point scale: *Not Helpful*, *Partially Helpful*, *Helpful*. This scale rates the annotator's subjective judgment of how well the caption clarifies the figure. A "helpful" caption clarifies the figure or adds valuable information. A "not helpful" caption may be confusing or provide no additional insight. The usefulness assessment complements the structural element checks by providing an overall sense of the caption's practical utility in interpreting the figure.

Caption Helpfulness Examples

Label:	Phrase:
"Graph X"	"The domain Ω with boundary conditions."
"Schematics"	"Translation performance on different instruction scales."
"The graph G_k"	

Sentence:
"Analytic structure of the full non-perturbative amplitudes $\Phi_i(s, t, u)$ in the complex plane s at a fixed value of $t < 0$. There are two branch cuts along the real axis, namely, for $s > 0$ and for $u = -s - t > 0$. [..]",

"An example of a subgraph G_i. The black vertices are included in the set S constructed in the proof of Lemma (ref). The cycle with the solid edges is a Hamiltonian cycle of G_i', and the dashed lines within the cycle represent edges of G_i' not in the Hamiltonian cycle. The dotted lines depict a possible partition of the vertices considered in steps 1-3 to obtain an upper bound on $|S|$."

Figure 3.4.: Examples illustrating the helpfulness of figure captions. The top box contains short labels and phrases that offer limited context, while the lower box shows full sentences providing specific, descriptive information about the figure content.

Length & Completeness

The length and completeness of each caption are assessed because these are basic characteristics that help to judge how detailed a caption is. We categorized **caption length** as short or long based on a threshold of 20 words.

$$\begin{array}{ll} \textit{Short} & \leq 20 \text{ words} \\ \textit{Long} & > 20 \text{ words} \end{array}$$

Captions containing 20 words or fewer were labeled short and those exceeding 20 words were labeled long. This threshold was chosen in order to identify short captions and this process can also be automated. Caption length can serve as an indicator of quality, it does not necessarily ensure it. Therefore, completeness was evaluated as a separate dimension.

Caption completeness refers to the grammatical and structural form of the caption. We classified captions into three levels: *Label*, *Phrase*, *Sentence*. A label is just a title or classification that often lacks a verb. That can be a noun or short noun phrase, such as "Experimental Setup". A phrase is a caption that forms a short phrase that lacks a complete sentence structure. It may contain descriptive words but isn't grammatically a complete independent sentence. A caption marked as a complete sentence is grammatically correct with a subject and verb in it. This categorization to label, phrase and sentence helps to identify how detailed each caption is. In scientific papers the captions are often written as phrases or titles instead of full sentences.

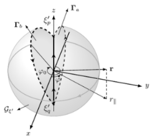
Scoring Method

Each caption received an overall quality score on a five point scale (Narins et al. 2023; Hsu et al. 2024; Kasai et al. 2022). The scoring criteria aligned with the previously mentioned aspects (functional role, structural elements, completeness) to provide a quantitative measure of the caption quality. The scale is defined as follows:

- **1 - Very poor:** The caption does not meaningfully describe the content of the figure. It may contain only a figure label, an unrelated code snippet or other irrelevant or misleading text. In practice, captions in this category are often labeled "Uninformative" as their functional role.
- **2 - Poor:** Although the caption is somewhat relevant to the figure, it is still vague, unclear and incomplete. While it may mention the general topic or name an element of the figure, it does not adequately explain what is shown. For instance, a caption that provides only a partial description without clarifying the context, would fall into this category. These captions provide some information, but not enough for a clear understanding.
- **3 - Average:** The caption offers a description of the figure. It usually describes the main content of the figure, but lacks depth or specificity. It states what the figure is, but doesn't focus on important details.
- **4 - Good:** The caption is well-written and relevant. It is nearly complete in its description. It captures important details of the figure's visual and quantitative content and is clear and specific. A caption of this quality explains the figure's structure and highlight important details.
- **5 - Very Good:** The caption is precise, informative and insightful. It describes the figure in detail and can function without the figure itself. The caption also effectively explains important key elements.

Figure 3.5 shows two examples of captions rated as "very poor" and "poor". Figure 3.6 shows two examples of "average" and "good". No caption in the evaluated sample was rated as "very good", therefore no example for this category is shown. Annotators used these definitions as guidelines when assigning scores to ensure consistent and objective scoring. Each caption in the sample was rated independently by multiple annotators. This provided a measure of reliability.

1 - Very Poor

	**Caption:** "cityscape at dusk"
	**Caption:** "Path on integral ???and ???."

2 - Poor

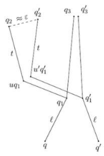
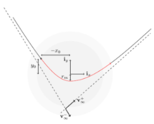
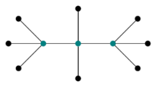
	**Caption:** "Illustration of Eskin–Mirzakhani scheme."
	**Caption:** "The Stark fly-by geometry."

Figure 3.5.: Two representative examples each from the 'Very Poor' and 'Poor' caption score.

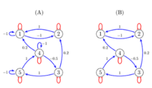
3 - Average

Caption: "The original graph Γ (left); the "doubled graph" Γ_2 (centre); an Eulerian cycle C (right), which forms a closed cycle in Γ_2, traversing every edge exactly once."

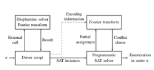


Caption: "A graph $G = (V, E)$ with 10 edges having $\mathrm{VC}(G) = 3$. The unique vertex cover of size 3 is highlighted in teal."

4 - Good



Caption: "Mixed graph (A) representing a model with $d = 5$ variables and with a diagonal C-matrix. The edge labels of the directed (blue) edges are the values of the non-zero entries in the M-matrix, while all the blunt (red) edges correspond to the diagonal entries of C all being 1. The mixed graph (B) is the same as (A) but with all directed self-loops removed. We call (B) the base graph of (A)."



Caption: "Outline of our algorithm for enumerating Williamson sequences of order n. The boxes on the left correspond to the preprocessing which encodes and decomposes the original problem into SAT instances. The boxes on the right correspond to an SMT-like setup where the system that computes the discrete Fourier transform takes on the role of the theory solver."

Figure 3.6.: Two representative examples each from the 'Average' and 'Good' caption score.

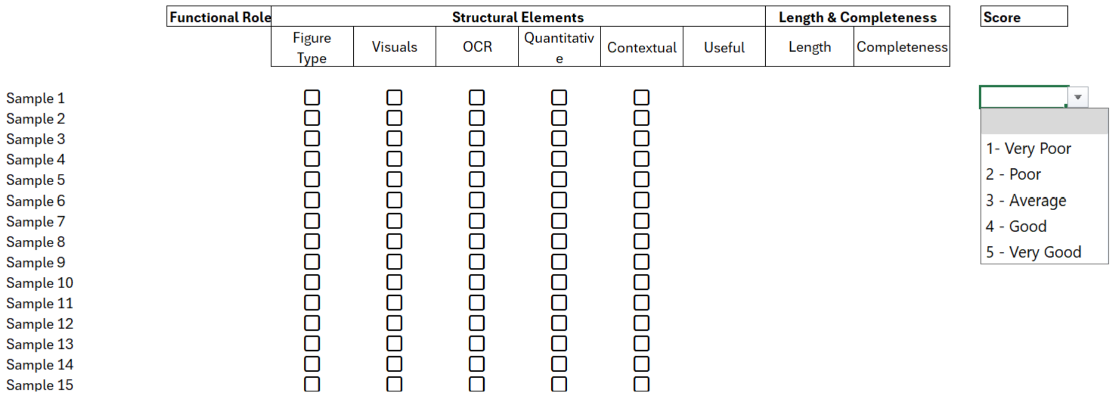


Figure 3.7.: Evaluation Sheet for annotations

3.2. Evaluation Sheet

For a a structured and efficient annotation process, we used a evaluation sheet to record all the above mentioned dimensions. The evaluation sheet was designed with pre-defined fields and options to help enforce consistency across annotators and samples. Figure 3.7 shows the evaluation sheet and provided following features for each caption entry:

- **Functional Role:** A drop down menu with the four categories listed: Uninformative, Descriptive, Explanatory and Interpretive. Annotators selected the option that best fit the caption.
- **Structural Elements:** A set of checkboxes for the five structural elements: figure type, visuals, OCR, contextual and quantitative content. A separate drop-down field was included for the three-level usefulness rating (not helpful, partially helpful, helpful) to capture the helpfulness judgment.
- **Length & Completeness:** Two drop-down fields, one for length and one for completeness. Annotators simply selected the appropriate category.
- **Score:** A drop down field for the scores 1 to 5..

The evaluation sheet presented results more structured and reduced ambiguity and variation in how the evalaution was recorded.

3.3. Caption Refinement Process

To improve the quality of captions, we implemented a human-guided caption refinement workflow. Annotators re-wrote each caption by hand, following best practices for scientific figure descriptions (Tufte 2001). The goal was to clarify and improve the captions while preserving factual accuracy and staying strictly within the information visible in the figure. In other words, no "hallucinated" content was added. Each caption was

Caption Refinement Example 1		
Original: *"(Left) The polytope P_A for Example **??**, (Right) The completable region and curves defined by $f(u,v) = 0$ and $g(u,v) = 0$."*	$\longrightarrow$	**Refined:** *"The figure shows a geometric figure of a triangle with labeled vertices. The triangle is formed by connecting three black dots labeled 1, 2 and 3 with undirected black lines. Inside the triangle there are two additional black dots labeled 4 and 5 positioned along the vertical axis."*

Original: *"The function $D^{\lambda}_{\phi}(\cdot, \frac{1}{2})$"*	$\longrightarrow$	**Refined:** *"A plot of a blue curve over $x \in [-2, 2]$ (horizontal axis) and $y \in [0, 2.25]$ (vertical axis). The curve begins at [0,0.25] dips down to [1/2,0] and then goes up to [2,2.25]. An identical curve is mirrored on the negative side of the x-axis, creating a symmetry about the y-axis. "*

Original: *"A generic diagram of a deep SSM with L layers. "*	$\longrightarrow$	**Refined:** *"A grid like composed of connected nodes. At the top row the nodes $x_{1,t-1}$, $x_{1,t}$ and $x_{1,t+1}$ are connected by horizontal arrows. From each of these top nodes a vertical arrow descends, indicated with dots for continuation, to the corresponding nodes $x_{L,t-1}$, $x_{L,t}$ and $x_{L,t+1}$. Each of these then connects downward to y_{t-1}, y_t and y_{t+1}, which are colored gray."*

Figure 3.8.: Side-by-side comparisons of original and refined figure captions.

edited into a self-contained, descriptive paragraph with correct grammar and a neutral, academic tone. The editors followed structured guidelines to ensure consistency and quality in the refinements. These guidelines are outline below.

- **Describing Visual Components:** All important visual elements of the figure (object, shapes, arrows, axes, etc.) are explicitly described in the caption. The purpose of the caption is to provide a detailed description of the figure's visual structure and layout.
- **Including Text and Data:** Any text within the figure, such as labels, titles or mathematical expression, as well as key quantitative information, is described in the caption.
- **Completeness Without Hallucination:** Only information provided within the figure should be described. No external information or content may be introduced. This intends to prevent misleading or speculative information (Cheng et al. 2025; Huang et al. 2025).
- **Grammar and Tone:** The captions were rewritten as complete and grammatically correct sentences written in a formal, scientific tone. Well-structured fluent captions improve readability and ensure that information is conveyed clearly.

4. Methodology

4.1. Inter-Annotator Agreement

Automatic metrics such as BLEU (Papineni et al. 2002), METEOR (Lavie and Agarwal 2007) and image similarity scores offer quick and consistent evaluations, they often lack the depth necessary to fully assess caption quality, particularly in scientific figures. This is why we conducted human evaluation in this work. However human evaluation can differ from evaluator to evaluator. It is important to determine inter-annotator agreement to calculate the consistency of these human ratings. In our work we use Cohen's Kappa (two raters) (Cohen 1960) and Fleiss' Kappa (more than two raters) (Fleiss 1971) to calculate this agreement. These metrics measure how much observed agreement exceeds what would be expected by chance.

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{4.1}$$

Here, p_0 represents the observed agreement and p_e represents the expected agreement by chance. These agreement values are typically interpreted using the Landis and Koch scale (Landis and Koch 1977), as shown in Table 4.1. Values above 0.80 indicate near-perfect agreement, while values below 0.40 suggest only fair or slight agreement.

4.2. Semantic Similarity Measure

After rewriting the captions, we measured how much their semantic content changed compared to the original. To do this, we used a semantic textual similarity analysis based on Sentence-BERT (sBERT) embeddings (Reimers and Gurevych 2019). sBERT is a neural network model that encodes sentences into high-dimensional vectors, placing semantically similar sentences closer together in the embedding space. For our analysis, we

Kappa	Interpretation
< 0	Poor agreement
0.00 - 0.20	Slight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 1.00	Almost perfect agreement

Table 4.1.: Landis and Koch Scale for interpreting Kappa scores (Landis and Koch 1977).

used the pre-trained *all-MiniLM-L6-v2* model, which produces 384-dimensional sentence embeddings. This model is a compact transformer variant known for its strong performance on semantic similarity tasks while remaining computationally efficient (Reimers and Gurevych 2019). We then compared the original and revised caption by calculating the cosine similarity between their vectors. In our case, scores usually ranged between 0 and 1. A score closer to 1 suggests that our revised description didn't significantly change the meaning. This approach is limited to only semantic overlap. For example, a refined description could drastically improve in terms of grammar and fluency, yet the cosine similarity score can still be near 1.0. (Sarto et al. 2025).

4.3. Evaluation of Generated Figures

To visually confirm the improvements of our revised description, we conducted an analysis on the generated figures via. TikZ-Code. GPT4o-mini was used to generate TikZ-code from captions, which then got evaluated with automated metrics and human judgment.

Quantitative Evaluation: We used three commonly adopted metrics to compare the generated figures to the original (ground-truth) DaTikZ-v3 figures: LPIPS (Learned Perceptual Image Patch Similarity), FSIM (Feature Similarity Index), and HOG (Histogram of Oriented Gradients). Each metric captures different aspects of visual similarity:

1. **LPIPS**(Zhang et al. 2018) quantifies perceptual similarity by comparing the activations of deep neural network features. Lower values indicate greater similarity, and a score of 0 denotes identical images.
2. **FSIM**(Zhang et al. 2011) focuses on low-level image features, such as edges and phase consistency. Values range from 0 to 1, with those closer to 1 indicating higher similarity.
3. **HOG similarity**(Dalal and Triggs 2005) captures the distribution of edge directions, providing useful structural alignment comparisons for geometric figures.

A metric win was achieved, if our figure outperformed the figure generated by the original caption in at least two of the three metrics. This majority-vote approach ensured robustness across complementary measures. We deliberately avoided metrics such as SSIM (Structural Similarity Index) and PSNR (Peak Signal-to-Noise Ratio) because they are sensitive to minor geometric changes. As with figures like those in DaTikZ-v3, small shifts in text labels or line positions can have a disproportionate effect on these metrics despite having minimal semantic impact (Wang et al. 2004)(Gonzalez and Woods 2002)(Fardo et al. 2016).

Human Evaluation: To complement the automated evaluation, we conducted a blind study with three human annotators. Each annotator was shown the original figure alongside two generated versions, one based on the original caption and one based on the revised caption, in a random order and without labels indicating which figure was generated with what caption. The annotators were asked to select the version that best matched the original figure. We used a simple majority vote to determine the winner. If at least two out of three annotators selected the figure based on the revised caption, that figure was considered a winner.

5. Experimental Setup

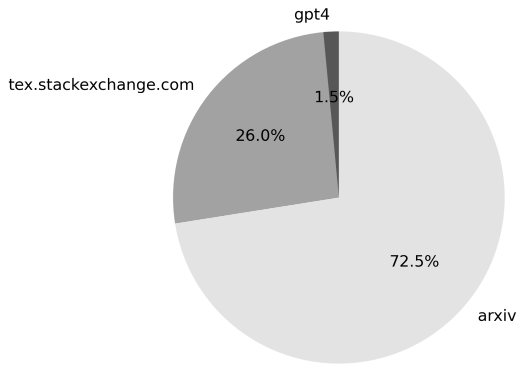


Figure 5.1.: Origin Distribution of the Sample Dataset

5.1. Tools and Environment

All experiments were conducted in a controlled computing environment. Specifically, we ran our benchmarks on the Data and Web Science Group's student GPU server at the University of Mannheim. The server is equipped with six NVIDEA RTX A6000 GPUs, each with 48GB of VRAM.

For human annotations, we employed a cloud-based spreadsheet (Google Sheets) as the evaluation platform. The annotation sheet had predefined fields to minimize ambiguity. As noted in Section 4.2, annotators selected from drop-down menus for each quality dimension, as well as an overall quality score.

5.2. Dataset for Experiments

To manually evaluate the quality of captions, we selected a representative sample of 200 image-caption pairs from the DaTikZ-v3 dataset (Belouadi et al. 2025). This number was chosen to be large enough to capture a wide range of caption styles and quality, while still being within the scope of this thesis to analyze. Since not all figures in DaTikZ-v3 come with captions, we curated the sample ourselves to reflect the original distribution of captioned figures across different sources. Specifically, 72.5% of the samples were taken from arXiv papers, 26% from TeX Stack Exchange posts, and 1.5% from GPT. This way, the sample stays true to the distribution of DaTikZ-v3, rather than over-representing the subset with existing captions. These 200 figure-caption pairs formed the basis for all further manual analysis.

5.3. Model Selection

A variety of language and vision-language models were employed to generate improved captions and TikZ-code. For caption generation, we selected four state-of-the-art multimodal models that balanced performance with computational feasibility. Due to resource constraints, models with parameter counts exceeding 8 billion were excluded. Our focus was on models within the 6–7 billion parameter range.

1. **Qwen2.5-VL-7B-Instruct (Bai et al. 2025):** is a vision-language variant of the Qwen model family that has been fine-tuned for instruction following. This 7 billion parameter model can process images and text, integrating visual understanding into language generation. Given its tailoring for following detailed instructions, we anticipated that Qwen2.5-VL would produce thorough captions when prompted appropriately.

2. **LLaVA-1.5-7B-hf (Liu et al. 2023):** is an open-source, mutimodal model based on the LLaMA architecture. LLaVA (Large Language and Vision Assistant) combines a vision encoder and a 7B LLaMA language model. A projection layer feeds visual features into the language model. This enables the model to reason about images in a conversational or instruction-following manner.

3. **BLIP-2-opt-6.7B (Li et al. 2023):** combines a pretrained vision encoder with a 6.7 billion parameter OPT language model. It uses the vision encoder to extract image features, which are then passed through a lightweight mapping network to the language model that generates the caption. We used the BLIP-2 variant based on the OPT-6.7B model. While BLIP-2 has shown strong performance on general image captioning benchmarks, it tends to produce relatively short and concise descriptions, which aligns with its design and training on datasets with brief captions (Li et al. 2023). Given this, we expected its outputs to be less detailed than those from models explicitly tuned for richer or domain-specific captions, making it a useful baseline for comparison in the context of scientific diagrams.

4. **GPT4o-mini (Radford et al. 2018, 2019)** is a compact version of OpenAI's GPT-4 model and can handle visual and text inputs in a unified transformer architecture. Although it is not as large as the GPT-4 model, GPT4o-mini aims to approximate its capabilities on a smaller scale. We included this model to evaluate the performance of a GPT4o-style approach in captioning, given that GPT-4o is renowned for its high-quality text generation.

Throughout our experiments, we used abbreviated names, Qwen, LLaVA, BLIP-2 and GPT4o-mini, to refer to these models. All models were accessed via the HuggingFace

Transformers library using their official checkpoints. The exception is GPT-4o-mini, which was accessed through the OpenAI API. Each model was run on an RTX A6000 GPU, which met the memory and computing requirements of all the models.

Caption Generation Procedure: We tasked each model with generating improved captions for all 200 figures in our evaluation sample. To ensure comparability, we standardized the input prompt and output configuration across models. The prompt used was:

> *"Describe this image precisely so someone could redraw it."*

Each model received the figure and prompt as input and could produce up to 256 output tokens. This token cap ensured descriptive completeness while controlling inference cost, which is particularly relevant for API-based models, such as GPT4o-mini.

Figure Generation: For the figure generation experiments in Section 6.3, we used GPT4o-mini to generate TikZ code. The token limit for each output was set to 1000 and we then compiled the resulting code into 512×512 images. To enable side-by-side comparison of each figure, we resized the original DaTikZ-v3 figures to 512×512. This enabled direct comparison for the metric-based evaluation.

6. Experiments

6.1. Caption Quality Evaluation Results

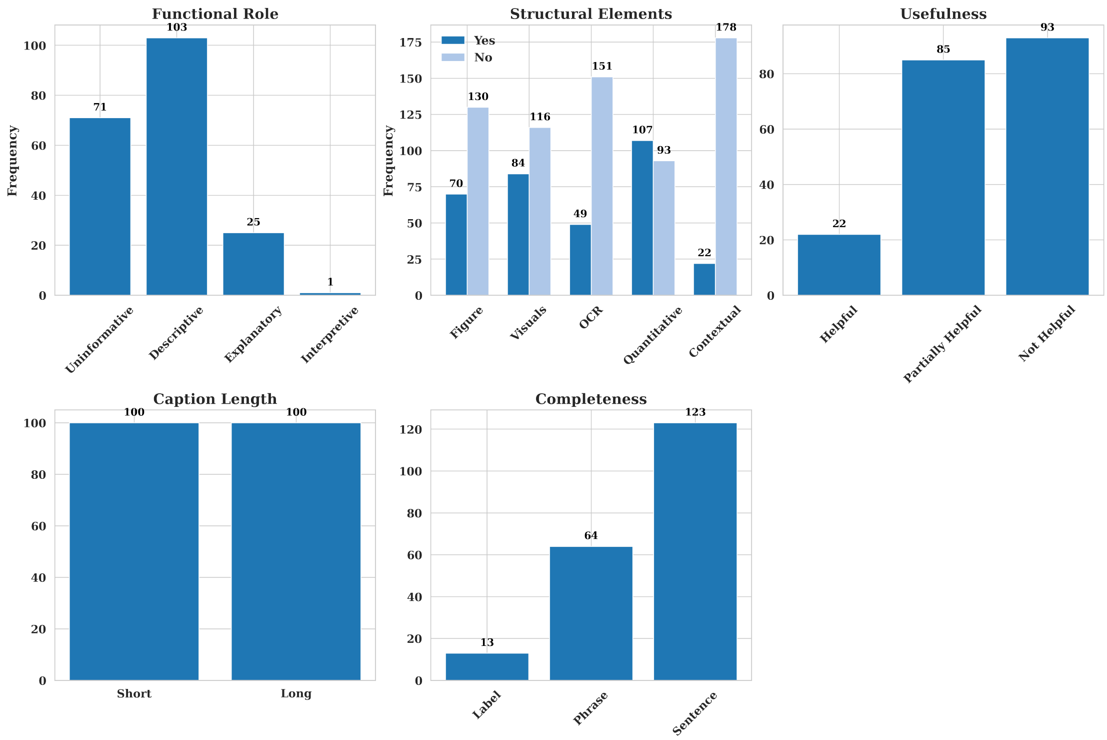


Figure 6.1.: Evaluation Results across 5 criteria: functional role, structural elements, usefulness, caption Length, completeness and overall score.

The first experiment was to evaluate the quality of the original captions in our sample dataset using the guidelines outlined in Section 3. The goal was to assess how well each caption described its corresponding figure.

Our evaluation revealed that many of the captions were missing important structural elements. In a lot of cases, captions didn't mention the type of figure, left out mathematical labels, or ignored any text shown in the image. Figure 6.1 gives an overview of the results: the majority of captions were rated as either "Uninformative" or just broadly "Descriptive," without offering much useful content. 93 captions were marked as not helpful at all, 85 were seen as only partially helpful, 130 didn't mention what kind of figure was shown, 116 failed to reference visual elements, and 151 ignored any

Annotator Pair	Functional	Figure Type	Visuals	OCR	Quantitative	Contextual	Useful	Completeness
Annotator 1 + Ourselves	0,405	0,718	0,67	0,556	0,732	1	0,545	1
Annotator 2 + Ourselves	0,479	0,737	0,727	0,867	0,444	0,634	NaN	NaN
Annotator 1 + Annotator 2	0,49	0,737	0,5	0,609	0,333	0,634	NaN	NaN
∅	0,458	0,731	0,632	0,594	0,503	0,756	0,545	NaN

Table 6.1.: Cohen's Kappa scores across annotator pairs. Only overlapping caption subsets (n = 15) were evaluated. NaN = not evaluated by annotator.

embedded text or labels. On top of that, more than half of the captions were shorter than 20 words, which made it hard for them to describe much about the image. This points to a common issue, most of the original captions simply aren't detailed enough to stand on their own.

To validate the consistency of our evaluation approach, we had two additional annotators independently assess overlapping subsets of the caption sample. We provided them with the same guidelines and then computed inter-annotator agreement using Cohen's kappa (Cohen 1960), following the interpretation scale proposed by **Landis and Koch** (Landis and Koch 1977). Since not all annotators rated the full set, agreement was calculated only on 15 overlapping samples.

As shown in Table 6.1, the agreement was stronger for more objective categories. Figure Type, Visuals, and OCR showed substantial agreement ($\kappa > 0.7$), indicating that annotators could reliably identify the presence or absence of these elements. This high consistency relates to the fact that these categories are less dependent on interpretations. More subjective dimensions, such as Functional Role and Usefulness, resulted in only moderate agreement. These categories require interpretive judgment and are more likely to vary among annotators. Nevertheless, the agreement scores remained within an acceptable range, which shows the robustness of our evaluation guidelines.

Score: In addition to binary and categorical judgments, each caption was rated on a scale of 1 to 5, with 1 being "Very Poor" and 5 being "Very Good." As shown in Figure 6.2, the distributions of these scores reveal a clear trend toward low ratings. The majority of ratings fell into the Poor or Very Poor categories. It is also notable that there were differences in how individual annotators applied the rubric. Annotator 1 demonstrated a stricter pattern, achieving an average score of 1.53. In contrast, Annotator 2 showed more variability and leniency, achieving an average rating score of 2.20. These patterns underscore the inherent subjectivity of qualitative scoring. Although all annotators recognized the general lack of quality, they had different thresholds for what constituted an acceptable caption. This variability is common in human annotation tasks and further supports the use of multiple annotators and agreement metrics for reliable evaluations.

These consistently low scores also highlight a broader issue: the original captions in the DaTikZ dataset frequently failed to provide thorough, standalone descriptions. One of the reasons for this is that many captions were derived from academic publications (Belouadi et al. 2025), which often assume the reader has access to the surrounding text

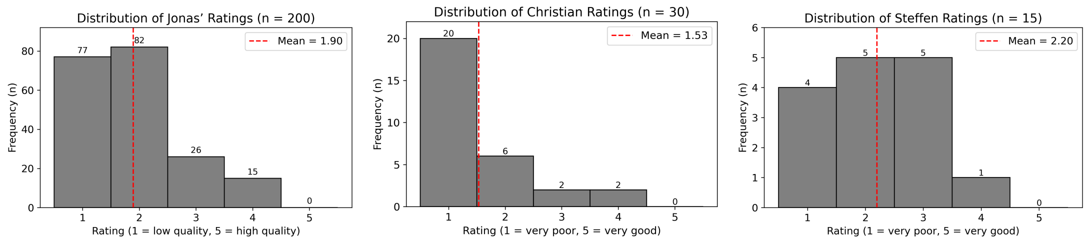


Figure 6.2.: Score distributions from ourselves (n = 200), Annotator 1 (n = 30) and Annotator 2 (n = 15). Red lines indicate mean score.

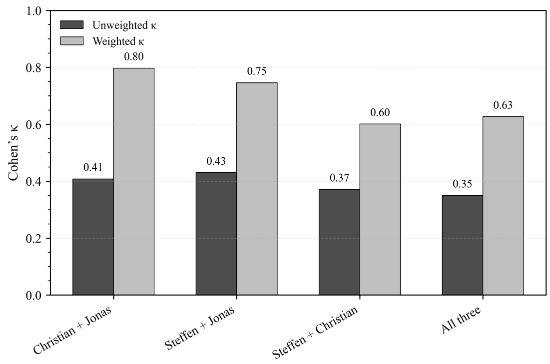


Figure 6.3.: Comparison between unweighted Kappa and Quadratic Weighted Kappa (QWK) across annotators pairs.

and prior domain knowledge. Consequently, these captions tend to focus on contextual interpretation rather than providing detailed visual descriptions.

To evaluate the reliability of our scoring system among annotators, we compared Cohen's unweighted Kappa and Quadratic Weighted Kappa (QWK) across all annotator pairs. Unweighted kappa measures simple agreement by treating all disagreements equally and QWK accounts for the severity of disagreement by assigning lower penalties to minor score differences. This distinction is particularly relevant in ordinal rating tasks like ours, where a one-point difference (e.g., "Poor" vs. "Average") is less meaningful than a three-point difference (e.g., "Very Poor" vs. "Good"). Figure 6.3 illustrates this comparison across the three annotator pairs and the combined set of all three annotators. In all cases, the QWK values are substantially higher than the unweighted values, showing that most disagreements occurred between adjacent rating categories.

Specifically, we observed the highest level of agreement between Annotator 1 and ourselves, with a weighted Kappa (QWK) of 0.80, compared to a unweighted Kappa of 0.41. This indicates a high degree of alignment in judgment, with only slight variations in score assignments. Similarly, Annotator 2 and ourselves achieved a QWK of 0.75, despite an unweighted Kappa of just 0.43. This indicates that, while the exact ratings did not match, the differences were small and consistent. The lowest level of agreement was observed between Annotators 1 and 2, who had a QWK of 0.60 and a Kappa of 0.37.

This reflects a reasonable level of consistency, but also suggests a greater variability in the scoring behavior of these two evaluators. The average agreement across all three annotators was 0.63 for QWK and 0.35 for unweighted kappa. These values reflect a moderate level of agreement overall, which is improved by the QWK's into a substantial agreement.

6.2. New Descriptions

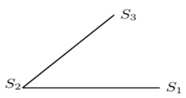
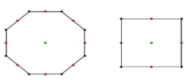
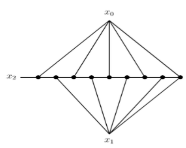
Figure	Original	Revised Description
	At an angle point.	An illustration of two black lines starting from the same point and forming an angle. One line extends horizontally to the right and is labeled S_1 at its end. The other extends upwards at an angle from the same point and is labeled S_3 at the end. The common starting point of both lines is labeled S_2.
	Draw octagon with color node at middle edges I would like to draw three pictures in TikZ but I don't know how to do it.	Two shapes adjacent to each other (left octagon and right square). Both shapes have a small green dot in the middle. Furthermore, the vertices of each shapes have small black dots. The midpoint of the lines connecting the vertices of each shape, is marked wit red dots. All dots have the same size.
	Example of a zigzag integral, with integration vertices marked with dots.	A geometric structure resembling a diamond. Along a horizontal black line labeled x_2, nine black dots are placed. Above the center of this line is a point labeled x_0 and below the line a point labeled x_1. Each dot on the line is connected by a straight line either to x_0 or x_1, in alternating order. The leftmost dot connects to x_0, the next to x_1 and so on, with the rightmost dot connecting to x_1.

Table 6.2.: Comparison between the original captions and revised descriptions.

Sample	Ours + Ann. 1	Ours + Ann. 2	Ann. 1 + Ann. 2
1	0.85	0.79	0.69
2	0.62	0.66	0.81
3	0.88	0.79	0.68
4	0.63	0.60	0.86
5	0.67	0.68	0.72
6	0.93	0.82	0.81
Mean	0.76	0.72	0.76

Table 6.3.: Pairwise semantic similarity (cosine) between our revised captions and two independent human annotators, computed with Sentence-BERT (higher is better).

With the poor quality of captions identified in our sample, we created new descriptions that improved the descriptiveness. We used the Guidelines provided in Section **??** and described exactly what is depicted in the figure, without adding unnecessary detail or contextual knowledge. We therefore also didn't interpret the figure and only described it. We followed this approach for all 200 captions in the sample dataset. On average, the original captions were approximately 61 words long, while our revised descriptions averaged around 95 words and had a consistent structure. Table 6.2 shows examples for our re-writes. In row 1 the original caption (from arxiv) provides only a minimal and vague description, omitting key visual elements. Our revised caption explicitly includes all of these elements. It also explicitly mentions the labels (S_1, S_2, S_3). In row two the original caption is a question. The reason for that is that this caption was sourced from TeX. We revised this into a description, by naming again key elements like "octagon" "square" and "green dot". The goal was to improve the information a LLM can get from the caption about visual details of the figure. To validate that our description are precise, we additionally compared a smaller subset (n = 6) to other human-written descriptions. Table 6.3 shows the semantic similarity between each annotator pair. The similarity scores range between 0.6 and 0.93 with an average score around 0.75. These values indicate a consistent high similarity among annotators. The strongest alignments were observed in comparisons between our description and Annotator 1's description (e.g. 0.93 and 0.88), which suggest a strong semantic overlap of the descriptions. Even the lower scoring pairs (e.g 0.6) reflect a meaningful degree of alignment. This suggests that the annotators consistently provided similar descriptions of the figures. It also suggests that our descriptions align with those of other human annotators, which gives our revised description more credibility. It is important to note that this semantic similarity metric measures only the textual similarity rather than correctness. High similarity indicates agreement in descriptive content among annotators but does not necessarily mean that the description is completely accurate in relation to the figure.

The process of re-writing 200 captions manually is a time-consuming task that is not scalable large scale. To address this, we investigated whether large language models can

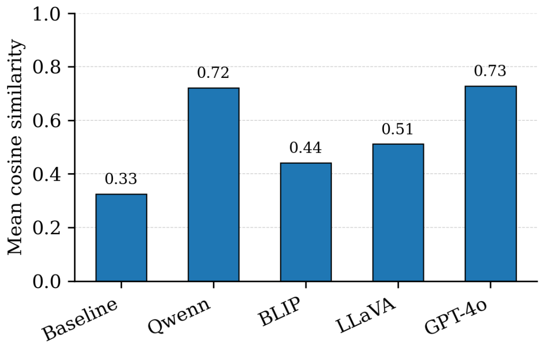


Figure 6.4.: Semantic similarity scores between the model-generated captions and the revised human-written descriptions. The baseline represents the average semantic similarity between the original DaTikZ-v3 captions and our revised versions. Higher scores indicate greater alignment with the revised descriptions.

come close to human re-written descriptions. We used Qwen, GPT-4o-mini, LLaVA and BLIP and tasked them to generate descriptions for all 200 figures in our sample dataset.

Figure 6.4 presents the semantic similarity between the model generated descriptions and our revised descriptions. The baseline compares our descriptions against the original captions in DaTikZ-v3. This low baseline score of 0.33 indicates that the original captions are less semantically similar to our revised captions. Among the tested model, Qwen and GPT4o-mini produced the highest score. This score of over 0.7 is even comparable to the semantic similarity between human annotators. In terms of descriptive content, these models come closely to our revised description. In contrast, LLaVA and BLIP score notably lower (0.44 and 0.51), but still higher than the original captions. This demonstrate that LLMs can generate description that can come close semantically to human written ones. But if the content of these descriptions is correct, still needs to be checked separately. This shows us that using model generated captions can help in describing figures large scale, though human oversight is still necessary to ensure accuracy.

6.3. Figure Generation Results

To demonstrate the impact of our revised captions, we used GPT-4o-mini to generate TikZ code. For comparison, we generated TikZ code for the original DaTikZ-v3 captions as well. We then compared each resulting figure to the original reference figure to evaluate how closely they aligned visually.

We conducted a side-by-side evaluation of a subset of 50 figures. For each sample, we used GPT4o-mini to generate TikZ code based on the original caption and based on our revised description. The resulting figures were then presented in a randomized

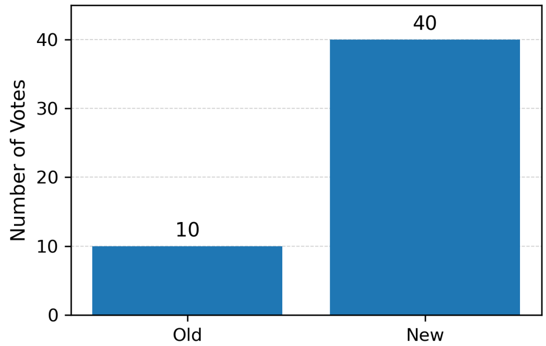


Figure 6.5.: Number of sample where the figure generated from the new or original caption was judged to more closely resemble the reference figure (n = 50).

layout. Only the original figure was labeled and the two generated versions were shown in a randomized order. This setup minimized potential bias. The goal was to determine which figure more closely resembles the original. Due to the evaluated and rated original captions, we expected that our revised versions perform better. Results in Figure 6.5 validates that. Out of the 50 judged captions, our revised version was selected 40 times and the original caption only 10 times. This clearly shows the impact of our refinements.

To further validate that, we focused on a subset of 10 samples. We applied on this subset two types of evaluation:

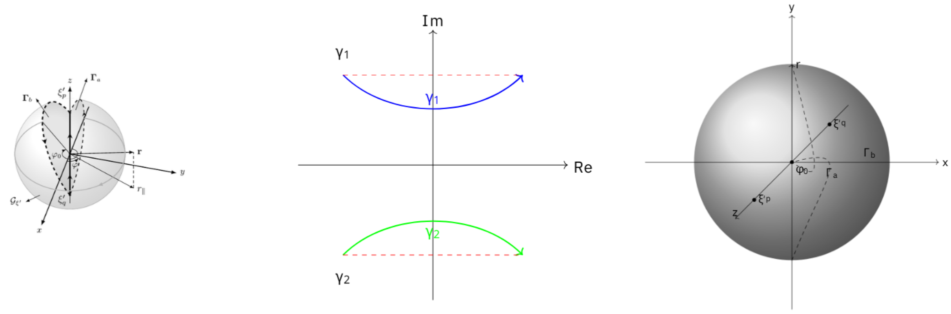
1. **Quantitative Evaluation:** We computed three commonly used image similarity metrics (LPIPS, SSIM, and HOG) to measure the visual similarity between the generated and original figures for 10 selected samples. We considered a sample a "win" for the revised caption if the generated figure outperformed the original caption figure on at least two of the three metrics. Detailed metric values are reported in Appendix 1.

2. **Qualitative Human Evaluation:** To complement the quantitative results, we conducted a blind human evaluation of the same ten samples. Three annotators were shown the original figure, as well as the two generated variants (from the original and revised captions), which were presented in a random order. The annotators were asked to select the version that more closely resembled the original. If two or more annotators selected the same version, it was marked as the winner for that sample.

In Figure 6.6, we present two examples where our revised captions significantly improved the quality of the generated figures. The first example demonstrates an ideal outcome. While the original caption includes basic details, such as color and the number of edges, it remains vague and fails to describe the layout. Our revised version addresses this by providing a detailed explanation of both the layout and the visual elements. As a result, the generated figure closely resembles the original. Since the original

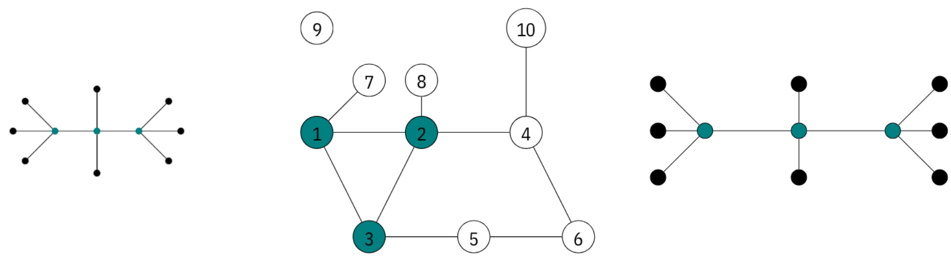
Caption Generation Comparison: Old vs. New

Original | **Old** | **New**



Old Caption: Path on integral 29 and 30.

New Description: The diagram is set in a 3D Cartesian coordinate system with axes labeled x, y, and z. There are two vectors r in the positive x-y-direction and $r_{||}$ in the negative x- and positive y-direction, which are connected through a dashed line. A large semi-transparent gray sphere denoted as $\mathcal{G}_{\xi'}$ is in the 3D coordinate system, with its midpoint being the origin of the coordinate system. Around the sphere at the middle, there is a gray line. The sphere is cut open from the top left and top right to its midpoint. The resulting surfaces are labeled as Γ_b (left) and Γ_a (right), with the two cutting lines depicted as dashed lines. The points where the cuts meet are labeled as $\xi_p^{'}$ (top) and $\xi_q^{'}$ (bottom).



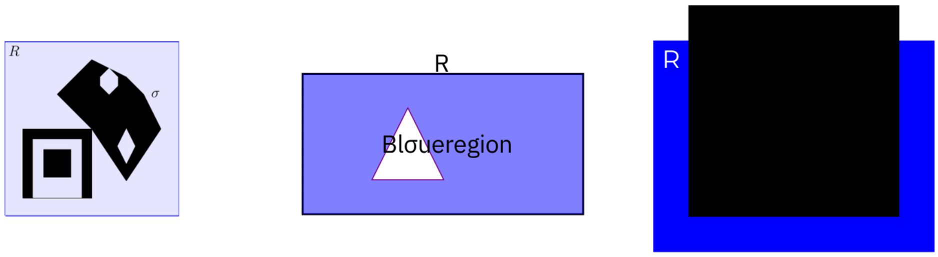
Old Caption: A graph $G = (V, E)$ with 10 edges having $\mathrm{VC}(G) = 3$. The unique vertex cover of size 3 is highlighted in teal.

New Description: A graph with three blue nodes arranged horizontally and connected by black lines. The leftmost blue node has three black lines extending to black nodes: one directly to the left, one diagonally upward to the left and one diagonally downward to the left. The center blue node connects vertically to two black nodes, one above and one below. The rightmost blue node has also three lines extending to black nodes: one directly to the right, one diagonally upward to the right and one diagonally downward to the right.

Figure 6.6.: Two examples illustrating that the figures generated using our revised captions (New) align more closely with the original ground truth figures than those generated using the original captions from DaTikZ-v3 (Old).

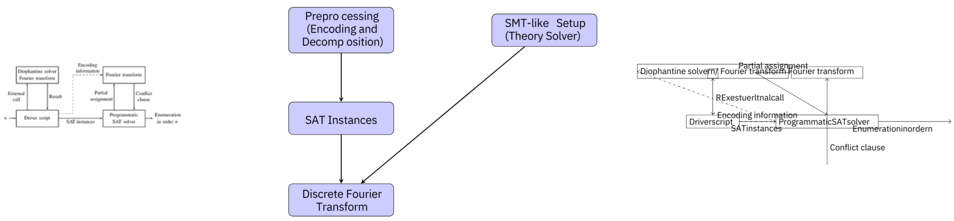
Caption Generation Comparison: Old vs. New

Original | **Old** | **New**



Old Caption: A set $\sigma \in PW^+$ inside a rectangle R. The blue region $R \setminus (\sigma \cup \partial R)$ can always be triangulated.

New Description: A blue rectangle labeled R in the top-left corner. Inside the rectangle, there are two black geometric figures. At the lower-left side, is a layered square pattern composed of three squares, a small black square at the center, surrounded by a blue square matching the background color of the rectangle, surrounded by a larger black square. Diagonally toward the upper-right is an irregular black polygon labeled σ. Inside the polygon two shapes have the black background color, one is hexagonal and the other is diamond shaped.



Old Caption: Outline of our algorithm for enumerating Williamson sequences of order n. The boxes on the left correspond to the preprocessing which encodes and decomposes the original problem into SAT instances. The boxes on the right correspond to an SMT-like setup where the system that computes the discrete Fourier transform takes on the role of the theory solver.

New Description: A block diagram with several components. There are four main labeled rectangular blocks connected by arrows indicating the direction. At the bottom left, there is an input labeled n entering a rectangular block titled "Driver script", which sends an arrow labeled "External call" upward to a block titled "Diophantine solver / Fourier transform". From this block another arrow labeled "Result" points downwards back to the "Driver script". From the "Driver script" a horizontal black arrow point to the right and is labeled "SAT instances" connected to a block titled "Programmatic SAT solver". It outputs a horizontal black arrow labeled "Enumeration in order n" pointing to the right out of the diagram. Above the "Programmatic SAT solver" is another block labeled "Fourier transform" and connected with an upward arrow labeled "Partial assignment" and a downward arrow labeled "Conflict clause". A dashed arrow labeled "Encoding information" points from the "Driver script" block back to the "Diophantine solver / Fourier transform".

Figure 6.7.: Two examples illustrating that the figures generated using the original captions from DaTikZ-v3 (Old) align more closely with the original ground truth figures than those generated using our revised captions (New).

figure is relatively simple, such high fidelity is not always guaranteed for more complex cases.

The second example involves a more complex figure. The original DaTikZ-v3 caption lacks specificity and fails to convey meaningful structural information. Consequently, GPT4o-mini generates a generic figure based on prior knowledge rather than from the caption itself, leading to a poor resemblance to the original. In contrast, our revised description includes detailed information about the coordinate system, vector directions, labels, the sphere, and color. Although replicating the original precisely remains challenging due to its complexity, the generated figure aligns much more closely with the original in structure and content.

However, not all revised descriptions led to improvements. Figure 6.7 illustrates two cases where the original captions resulted in better figures. In the first case, the figure consists of numerous geometric elements. Our revised caption described these in detail, but the model still struggled to translate this into accurate TikZ code. As a result, while the generated figure based on the original caption was simplistic and incorrect, it more closely resembled the structure of the original. A similar outcome occurred in the second case. Despite following a block diagram structure similar to the original, the complexity introduced by many labeled boxes and arrows led to overlapping content in the generated figure. The simpler version generated from the old caption is less detailed, but resembles the original more.

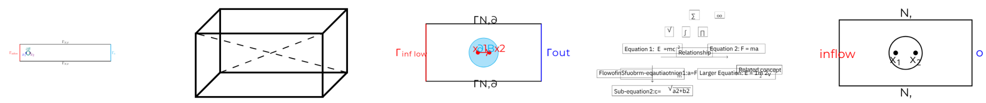
We find that in most cases, our revised descriptions enabled GPT4o-mini to produce better TikZ code, yielding figures more similar to the originals. As previously discussed, manually rewriting the entire DaTikZ-v3 dataset is not scalable. For this reason, we also analyzed captions generated by multimodal models.

Figures 6.8 and 6.9 present examples from our model comparison. Each figure includes the original image from the DaTikZ dataset alongside versions generated based on descriptions from BLIP, Qwen, LLaVA and GPT4o-mini. Based on our earlier analysis of semantic similarity, we anticipated that Qwen and GPT-4o-mini would produce the most accurate and detailed captions, which was confirmed by the results. Their descriptions generated figures that closely resembled the structure of the original, successfully capturing key elements such as labels, colors, and mathematical symbols. Although some minor details were omitted or simplified, the overall layout and primary components were effectively represented.

BLIP and LLaVA demonstrated weaker performance. LLaVA's captions were generally vague, often relying on broad phrases such as "a variety of mathematical symbols" or "several arrows", without specifying their nature or placement. As a result, the figures produced from these captions did not closely resemble the originals. Qwen and GPT-4o-mini, by comparison, provided far more specific information, including references to symbols (e.g., ∂B) and descriptions of directional flow. This level of detail helped the generation of figures that more closely align with the originals. BLIP's captions were particularly minimal and this led to overly simplistic figures. Interestingly, as illustrated in Figure 6.9, although the BLIP caption does not mention any numbers, the generated figure includes points labeled 1 through 5. This likely occurred because the caption was so brief and generic that the model defaulted to familiar patterns. Nonetheless,

Generated Captions and Their Figures

Original	BLIP	Qwen	LLaVA	GPT4o-mini



Original: The domain Ω with boundary conditions.

BLIP: a diagram of a rectangular box with two lines

Qwen: The image depicts a rectangular domain with labeled boundaries. The rectangle is outlined in black, and its dimensions are indicated by the labels at the top and bottom edges: "$r_{N,\partial}$" on both the top and bottom sides. On the left side of the rectangle, there is a red vertical line segment labeled "$r_i nflow$," and on the right side, there is a blue vertical line segment labeled "$r_o ut$." Inside the rectangle, near the center-left, there is a small circular region with a cyan border and a darker cyan interior. This circle is labeled "∂B" in cyan text. Within this circle, two points are marked: one labeled "x1" in red and the other labeled "x2" in red. An arrow pointing from "x1"to "x2" is drawn within the circle, indicating a direction or flow. The overall layout suggests a mathematical or physical model, possibly related to fluid dynamics or boundary conditions, where the rectangle represents a domain, the lines represent boundaries, and the circle with points and an arrow indicates a specific region of interest within that domain.

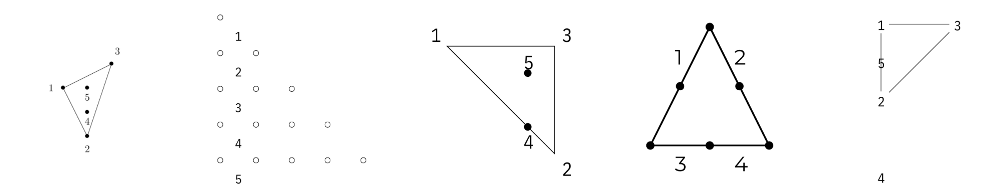
LLaVA_: "The image features a white background with a variety of mathematical symbols and equations. There are several arrows pointing in different directions, indicating the flow of information or the relationships between the equations. In the scene, there are two main equations, one located towards the left side and the other towards the right side. The left equation is accompanied by a few smaller equations, while the right equation is accompanied by a larger equation. The entire image is filled with mathematical symbols, making it a complex and detailed visual representation of mathematical concepts."

GPT4o-mini: The image depicts a rectangular domain with labeled boundaries. The left side is labeled "r inflow" in red, while the right side is labeled "ro" in blue. The top and bottom boundaries are labeled "rN,0" in black. Inside the rectangle, there is a boundary denoted as "∂B" represented by a circle with two points marked, labeled "x1" and "x2." The overall structure conveys a mathematical or physical representation, possibly related to boundary conditions in a fluid dynamics or mathematical analysis context.

Figure 6.8.: Comparison between the original figure and caption, and each model-generated description along with its corresponding regenerated figure. Each figure was generated using TikZ code produced by GPT4o-mini based on the model's caption.

Generated Captions and Their Figures

Original **BLIP** **Qwen** **LLaVA** **gpt4o-mini**



Original: (Left) The polytope P_A for Example **??**; (Right) The completable region and curves defined by $f(u, v) = 0$ and $g(u, v) = 0$

BLIP: a diagram showing the number of dots in a triangle

Qwen: The image depicts a geometric figure consisting of three vertices connected by straight lines to form a triangle. The vertices are labeled with numbers: 1 at the top left, 2 at the bottom right, and 3 at the top right. Inside the triangle, there are two additional points labeled 4 and 5, positioned closer to vertex 2. The points are marked with small black dots. The lines connecting the vertices are solid and black, forming the perimeter of the triangle. There are no arrows or colors used in the image; it is entirely monochromatic with black lines and dots on a white background.

LLaVA_: The image is a black and white drawing of a triangle with a dot at each corner. The triangle is divided into four sections, each with a number label. The numbers are arranged in a clockwise order, starting from the top left corner and moving counterclockwise around the triangle. The triangle is composed of three main lines, with the top line connecting the corners of the triangle. The bottom line is slightly curved, connecting the bottom corners of the triangle. The middle line is also slightly curved, connecting the middle corners of the triangle. The dots are positioned at the intersection of each line, creating a well-defined geometric shape."

gpt4o-mini: The image depicts a simple graph consisting of six nodes and several edges. At the top, there is a node labeled "3" positioned near the right side. Directly below "3," there is another node labeled "5." To the left of both nodes, there's a node labeled "1," which is connected by a diagonal line to node "2" located below it, and also connected to node "3." Node "2" is situated at the lowest point of the diagram. From "2," a vertical line connects downward to another node labeled "4." This creates a vertical structure where nodes "2" and "4" are directly aligned. The edges visually create a triangular shape with "3" at the top vertex, while "1" and "2" form the base. The overall structure has a simple, clear arrangement of connected nodes.

Figure 6.9.: Comparison between the original figure and caption, and each model-generated description along with its corresponding regenerated figure. Each figure was generated using TikZ code produced by GPT4o-mini based on the model's caption.

the outputs from BLIP generally did not align well with the original images. These findings indicate that detailed, model-generated captions, particularly those from Qwen and GPT-4o-mini, can serve as an approach for scaling figure description tasks.

6.4. Large Scale

To evaluate the scalability and practical feasibility of enhancing caption quality in the DaTikZ dataset, we conducted a large-scale captioning experiment using Qwen, identified in prior evaluations as the best-performing language model. Given its superior performance in terms of semantic similarity to human-authored descriptions, Qwen was selected to generate revised captions for a substantially larger subset of the dataset. Specifically, we annotated 2,000 figures extracted from the DaTikZ-v3 release, employing Qwen to automatically process each image and generate a corresponding description.

A key outcome of this experiment was the demonstration of both the efficiency and scalability of using Qwen for large-scale caption refinement. The generation process for all 2,000 figures was completed in approximately three hours, indicating that high-quality captioning with large language models is not only accurate but also computationally efficient. This approach presents several notable advantages. First, the ability to generate improved captions with minimal human intervention suggests that it is feasible to enhance large-scale datasets such as DaTikZ-v3. This enhancement has the potential to improve the performance of downstream models, particularly in tasks such as text-to-figure generation and figure classification. Furthermore, employing a single, consistent model like Qwen supports stylistic and structural uniformity across captions. While human-authored descriptions remain the benchmark for accuracy and completeness, our results indicate that LLMs can serve as effective tools for augmenting and accelerating the captioning process, especially in domain-specific applications involving scientific figures.

7. Conclusions

In this thesis, a structured human evaluation framework was introduced to assess the quality of captions in scientific figure datasets. The focus was on a representative sample of 200 figures from the DaTikZ-v3 dataset (Belouadi et al. 2024). By applying a multi-dimensional framework, which evaluates functional role, structural elements and completeness, this study revealed that many captions in the dataset lack descriptiveness, are too brief and often do not accurately present the visual content of the figure itself. This shows in an average caption rating of 1.9 across the sample. To validate the reliability of the human evaluation framework, inter-annotator agreement was measured. Objective categories such as figure type and OCR achieved high agreement, demonstrating that the framework offers a reliable and reproducible basis for caption assessment.

To address these shortcomings, the thesis proposed as visually grounded, context-independent approach to caption re-writing. Instead of relying on surrounding text or external context, each caption was rewritten solely based on the information visually present in the figure. This resulted in longer, more informative captions that more accurately reflected the figures content. Semantic similarity analyses using Sentence-BERT embeddings confirmed the consistency of the rewritten captions across different annotators, with high cosine similarity scores indicating strong alignment in visual interpretation.

To test the practical impact of caption improvement, both the original and the revised captions were used to generate TikZ figures using GPT4o-mini. The generated figures were then compared to the ground-truth images in DaTikZ-v3 using both human judgments and automated metrics. The analysis demonstrated that figures generated from the revised captions aligned more closely in structure and appearance with the originals, validating the hypothesis that improved caption quality leads to better figure generation.

Recognizing the scalability challenge of manual caption re-writing, this thesis also explored whether large-scale vision language models could generate comparable captions. Captions produced by BLIP-2, LLaVA-1.5, Qwen2.5-VL and GPT4o-mini were evaluated for semantic similarity to human-written versions. GPT4o-mini and Qwen2.5-VL produced captions that came semantically closest to human-generated description and resulted in strong figure generation. This shows that LLM-based caption generation is scalable for large datasets.

While the findings of this thesis are promising, the study is also subject to several limitations. First, the dataset scope of only 200 samples. While appropriate for a thesis of this scope, its not large enough to make generalizations across all of DaTikZ-v3 or other scientific datasets. This also resulted in only small sample calculations for Kappa and cosine similarity between humans. Second, due to computational constraints, only small- to mid-scale models were evalauted. Larger models, which could potentially perform better, were not included. Another limitation is the TikZ generation process that relied exclusively on GPT4o-mini for orignal, re-written and model captions. This creates a potential bias, especially for vague or minimal captions.

Building on the findings and acknowledging these limitations, future research could use larger datasets. Expanding the sample size would allow for more comprehensive understanding of caption quality. Future work could investigate a hybrid caption generation approach, in which LLMs produce candidate captions that are refined by human annotators. This could balance quality with scalability. Alternative generation approaches could be tested to more fairly compare caption types without model-specific biases.

Bibliography

Anderson, P., B. Fernando, M. Johnson, and S. Gould (2016). Spice: Semantic propositional image caption evaluation.

Bai, S., K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin (2025). Qwen2.5-vl technical report.

Belouadi, J., E. Ilg, M. Keuper, H. Tanaka, M. Utiyama, R. Dabre, S. Eger, and S. P. Ponzetto (2025). Tikzero: Zero-shot text-guided graphics program synthesis.

Belouadi, J., A. Lauscher, and S. Eger (2024). Automatikz: Text-guided synthesis of scientific vector graphics with tikz.

Chan, D., S. Petryk, J. E. Gonzalez, T. Darrell, and J. Canny (2023). Clair: Evaluating image captions with large language models.

Cheng, K., W. Song, J. Fan, Z. Ma, Q. Sun, F. Xu, C. Yan, N. Chen, J. Zhang, and J. Chen (2025). Caparena: Benchmarking and analyzing detailed image captioning in the llm era.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 20*(1), 37–46.

Dalal, N. and B. Triggs (2005). Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Volume 1, pp. 886–893 vol. 1.

Fardo, F. A., V. H. Conforto, F. C. de Oliveira, and P. S. Rodrigues (2016). A formal evaluation of psnr as quality measurement parameter for image segmentation algorithms.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin 76*(5), 378–382.

Gonzalez, R. C. and R. E. Woods (2002). *Digital Image Processing* (3rd ed.). Pearson Education.

Hessel, J., A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi (2022). Clipscore: A reference-free evaluation metric for image captioning.

Hsu, T.-Y., C. L. Giles, and T.-H. K. Huang (2021). Scicap: Generating captions for scientific figures.

Hsu, T.-Y., C.-Y. Huang, S.-H. Huang, R. Rossi, S. Kim, T. Yu, C. L. Giles, and T.-H. K. Huang (2024, May). Scicapenter: Supporting caption composition for scientific figures with machine-generated captions and ratings. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI '24, pp. 1–9. ACM.

Huang, L., W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu (2025, January). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems 43*(2), 1–55.

Kasai, J., K. Sakaguchi, L. Dunagan, J. Morrison, R. Le Bras, Y. Choi, and N. A. Smith (2022, July). Transparent human evaluation for image captioning. In M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz (Eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States, pp. 3464–3478. Association for Computational Linguistics.

Landis, J. R. and G. G. Koch (1977). The measurement of observer agreement for categorical data. *Biometrics 33*(1), 159–174.

Lavie, A. and A. Agarwal (2007, June). METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In C. Callison-Burch, P. Koehn, C. S. Fordyce, and C. Monz (Eds.), *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic, pp. 228–231. Association for Computational Linguistics.

Li, J., D. Li, S. Savarese, and S. Hoi (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.

Liu, H., C. Li, Q. Wu, and Y. J. Lee (2023). Visual instruction tuning.

Narins, L., A. T. Scott, A. Gautam, A. Kulkarni, M. Castanon, B. Kao, S. Ihorn, Y.-T. Siu, J. M. Mason, A. M. Blum, and I. Yoon (2023). Validated image caption rating dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu (2002, July). Bleu: a method for automatic evaluation of machine translation. In P. Isabelle, E. Charniak, and D. Lin (Eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, pp. 311–318. Association for Computational Linguistics.

Radford, A., J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever (2021). Learning transferable visual models from natural language supervision.

Radford, A., K. Narasimhan, T. Salimans, and I. Sutskever (2018). Improving language understanding by generative pre-training. OpenAI.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever (2019). Language models are unsupervised multitask learners. OpenAI.

Reimers, N. and I. Gurevych (2019, November). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In K. Inui, J. Jiang, V. Ng, and X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, pp. 3982–3992. Association for Computational Linguistics.

Sarto, S., M. Cornia, and R. Cucchiara (2025). Image captioning evaluation in the age of multimodal llms: Challenges and future perspectives.

Tang, B. J., A. Boggust, and A. Satyanarayan (2023). Vistext: A benchmark for semantically rich chart captioning.

Tufte, E. R. (2001). *The Visual Display of Quantitative Information* (2nd ed.). Cheshire, CT: Graphics Press.

Vedantam, R., C. L. Zitnick, and D. Parikh (2015). Cider: Consensus-based image description evaluation.

Wang, Z., A. Bovik, H. Sheikh, and E. Simoncelli (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing 13*(4), 600–612.

Yang, Z., R. Dabre, H. Tanaka, and N. Okazaki (2023). Scicap+: A knowledge augmented dataset to study the challenges of scientific figure captioning.

Zhang, L., L. Zhang, X. Mou, and D. Zhang (2011). Fsim: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing 20*(8), 2378–2386.

Zhang, R., P. Isola, A. A. Efros, E. Shechtman, and O. Wang (2018). The unreasonable effectiveness of deep features as a perceptual metric.

Zhang, Y., R. Zhang, J. Gu, Y. Zhou, N. Lipka, D. Yang, and T. Sun (2024). Llavar: Enhanced visual instruction tuning for text-rich image understanding.

A. Additional Experimental Results

Sample	Metric	Old	New
1	LPIPS ↓	NaN	0.0965
	SSIM ↑	NaN	0.9504
	HOG ↑	NaN	0.3701
2	LPIPS ↓	0.1972	**0.0965**
	SSIM ↑	0.9043	**0.9504**
	HOG ↑	0.0927	**0.1041**
3	LPIPS ↓	0.2434	**0.0917**
	SSIM ↑	0.9240	**0.9494**
	HOG ↑	0.0630	**0.1141**
4	LPIPS ↓	0.4542	**0.2153**
	SSIM ↑	0.6599	**0.8241**
	HOG ↑	0.2915	**0.4509**
5	LPIPS ↓	0.4216	**0.2920**
	SSIM ↑	0.6425	**0.7032**
	HOG ↑	0.1073	**0.5062**
6	LPIPS ↓	0.1554	**0.0508**
	SSIM ↑	0.6425	**0.9705**
	HOG ↑	0.1073	**0.3899**
7	LPIPS ↓	0.1554	**0.0508**
	SSIM ↑	0.9230	**0.9705**
	HOG ↑	0.0223	**0.3899**
8	LPIPS ↓	0.1928	**0.1717**
	SSIM ↑	0.8816	**0.9143**
	HOG ↑	**0.1423**	0.0387
9	LPIPS ↓	0.3516	**0.1104**
	SSIM ↑	0.7104	**0.8951**
	HOG ↑	0.1747	**0.4418**
10	LPIPS ↓	**0.1171**	0.1292
	SSIM ↑	0.9265	**0.9369**
	HOG ↑	**0.1005**	0.0992

Table A.1.: Image–figure similarity (n = 10). Lower LPIPS and higher SSIM / HOG values indicate closer alignment between the generated figure and the ground-truth scientific figure. Best value per row is **bold**.

Ehrenwörtliche Erklärung

Ich versichere, dass ich die beiliegende Bachelor-, Master-, Seminar-, oder Projektarbeit ohne Hilfe Dritter und ohne Benutzung anderer als der angegebenen Quellen und in der untenstehenden Tabelle angegebenen Hilfsmittel angefertigt und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen. Ich bin mir bewusst, dass eine falsche Erklärung rechtliche Folgen haben wird.

Declaration of Used AI Tools

Tool	Purpose	Where?	Useful?
ChatGPT	Rephrasing	Throughout	-
DeepL	Translation	Throughout	++
ChatGPT	LaTeX Layout	Figures	++
GPT4o	Code drafting	Charts	++

Jonas Gnauck
Mannheim, den 02.06 2024