Figure Captioning with Large Generative Models: A Survey

Anonymous ACL submission

Abstract

Figures, such as charts, graphs, and scientific visualizations, are essential for communicating data-driven insights, yet their interpretation often depends on clear and informative captions. Recent advances in large generative models, particularly multimodal large language models (MLLMs), have enabled significant progress in automatic figure captioning. In this paper, we present the first comprehensive survey of figure captioning with large generative models. We introduce an intuitive taxonomy that categorizes these techniques across key dimensions. We formalize the figure captioning task, and provide a systematic overview of the underlying techniques. In addition, we survey benchmarks, datasets, and evaluation methodologies, including both traditional metrics and modern MLLM-as-a-judge approaches. Finally, we propose important open problems and future directions, including personalized caption generation, heterogeneous figure types, and high-fidelity figure understanding. This survey unifies recent progress and provides a foundation for advancing research at the intersection of vision, language, and scientific communication.

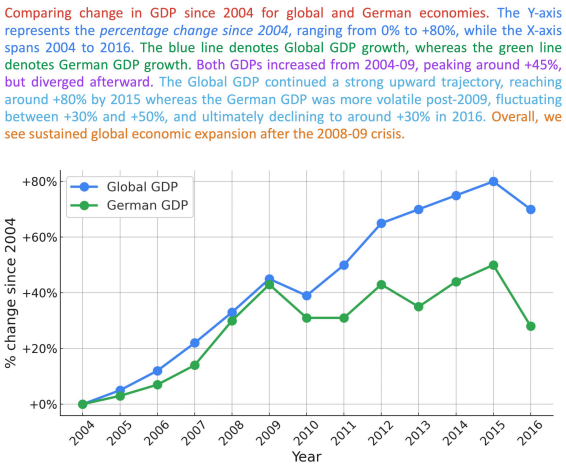

Figure 1: Overview of the figure-captioning task. High-quality captions integrate six criteria, highlighted in color to align with the example caption: **helpfulness** (improves understanding), **OCR** (mentions words from figure), **visual** (mentions visual characteristics, color, etc), **relations** (describe relationships), **stats** (mentions values from the figure), **takeaway** (describe conclusions or insights). See Section 5.2 for further details.

1 Introduction

Figures[1] are critical for communicating complex information across a wide range of domains (Chen et al., 2020; Corio and Lapalme, 1999; Feng and Lapata, 2013; Qian et al., 2021; Mittal et al., 1998; Demir et al., 2008; Zhang et al., 2024; Yan et al., 2025). However, their interpretability relies critically on the presence of clear, accurate, and semantically rich captions (Hsu et al., 2023, 2024). The task of figure captioning aims to generate natural language descriptions that effectively convey the salient patterns, key findings, and contextual insights embedded in figures (Hsu et al., 2021).

Despite rapid progress in figure captioning (Hsu et al., 2021; Yan et al., 2025; Yin et al., 2025), most existing approaches including those based on large generative models, remain limited in their ability to capture key findings and high-level insights conveyed by visualizations. Rather than generating captions that communicate the salient patterns, trends, or takeaways embedded in a figure, current methods predominantly extract and rephrase textual elements such as axis labels, titles, and legends. This reliance on surface-level textual cues overlooks the deeper semantic content and often fails to support effective interpretation or decision-making. As a result, figure captioning remains unsolved (Liu et al., 2025; Mukhopadhyay et al., 2024). One goal of this survey is to succinctly summarize existing figure captioning approaches and highlight the fundamental open problems and remaining challenges. We hope that the findings from this survey helps enable researchers to develop novel approaches that solve these challenges leading to significantly better captions.

Motivated by the increasing importance of this

[1] Figure, chart, visualization, diagram, and graph are often used synonymously in the literature.

	Modalities (Section 3)				Techniques (Section 4)						Figure Types (Section 2.3)							Evaluation (Section 5)		
	Figure Only (§3.1)	Figure-Caption Pairs (§3.2)	Fig-Cap Pairs + Aux. Text (§3.3)	Conversational (§3.4)	In-Context (§4.1)	Fine-Tuning (§4.2)	RL / RLHF (§4.3)	Distillation (§4.4)	Retrieval/Hybrid (§4.5)	Personalization (§4.6)	Bar	Line	Scatter	Pie	Heatmap	Flowchart	Other	Human Evaluation	Traditional Metrics	M(LLM)-as-a-Judge
FigCAP (Chen et al., 2019)	✗	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓	✓	✗	✓	✗	✗	✗	✗	✓	✗
SciCap (Hsu et al., 2021)	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓	✓	✓	✓	✗	✗	✓	✗	✓	✗
DePlot (Liu et al., 2022)	✗	✓	✗	✗	✓	✓	✗	✗	✗	✗	✓	✓	✗	✓	✗	✗	✗	✗	✓	✗
VisText (Tang et al., 2023)	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✓	✗
StructChart (Xia et al., 2023)	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓	✓	✗	✓	✗	✗	✗	✗	✓	✗
ChartLlama (Han et al., 2023)	✗	✓	✗	✗	✓	✓	✗	✗	✗	✗	✓	✓	✓	✓	✓	✗	✓	✗	✓	✗
UniChart (Masry et al., 2023)	✗	✓	✗	✗	✗	✓	✗	✓	✗	✗	✓	✓	✗	✓	✗	✗	✗	✗	✓	✗
FigCaps-HF (Singh et al., 2023)	✗	✓	✗	✗	✗	✗	✓	✗	✗	✓	✓	✓	✗	✓	✗	✗	✗	✓	✓	✗
SciCap+ (Yang et al., 2024)	✗	✓	✓	✗	✗	✓	✗	✗	✗	✗	✓	✓	✓	✓	✗	✓	✓	✓	✓	✗
FlowLearn (Pan et al., 2024)	✗	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✓	✗
ChartAssistant (Meng et al., 2024)	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓	✓	✓	✓	✗	✗	✓	✗	✓	✗
ChartThinker (Liu et al., 2024)	✗	✓	✓	✗	✓	✓	✗	✗	✓	✗	✓	✓	✓	✓	✗	✗	✗	✓	✓	✗
ChartGemma (Masry et al., 2024b)	✓	✗	✗	✗	✗	✓	✗	✓	✗	✗	✓	✓	✓	✓	✗	✗	✓	✓	✓	✗
FigurA11y (Singh et al., 2024)	✗	✗	✓	✓	✓	✗	✗	✗	✗	✗	✓	✓	✓	✓	✗	✓	✓	✓	✗	✗
ChartInstruct (Masry et al., 2024a)	✓	✗	✗	✗	✗	✓	✗	✓	✗	✗	✓	✓	✗	✓	✗	✗	✗	✓	✓	✗
LaMP-Cap (Ng et al., 2025b)	✗	✓	✓	✗	✓	✗	✗	✗	✗	✓	✓	✓	✓	✓	✗	✗	✗	✓	✓	✗
ChartCap (Lim et al., 2025)	✗	✓	✗	✗	✓	✓	✗	✗	✗	✗	✓	✓	✓	✓	✓	✗	✓	✓	✓	✗
ChartMoE (Xu et al., 2025)	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓	✓	✓	✓	✓	✗	✓	✗	✓	✗
Arrow-Guided (Omasa et al., 2025)	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✓	✗	✓
Pluto (Srinivasan et al., 2025)	✗	✗	✓	✓	✓	✗	✗	✗	✗	✗	✓	✓	✓	✗	✗	✗	✓	✓	✗	✗
ChartVLM (Xia et al., 2025)	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓	✓	✓	✓	✓	✗	✓	✗	✓	✓

Table 1: Taxonomy of figure captioning methods categorized by input modalities, techniques, and evaluation metrics.

unsolved problem, we present the first comprehensive survey on figure captioning with large generative models. Our goal is to synthesize existing approaches, datasets, and evaluation methods, and to establish a roadmap for advancing research at the intersection of vision, language, and reasoning.

Summary of Main Contributions: A summary of the key contributions are as follows:

1. **A formalization of figure captioning as a unified multimodal generation problem.** We introduce a formal problem formulation that captures the multimodal nature of figure captioning. This formulation serves as a foundation for understanding how large generative models perceive, interpret, and describe figures.
2. **A survey and taxonomy of techniques for figure captioning.** We present a comprehensive taxonomy that organizes the approaches across two dimensions: input modalities and techniques.
   - **Input Modalities**: Figure-Only, Figure-Caption Pairs, Fig-Cap Pairs + Auxiliary Text, and Conversational Captioning
   - **Techniques**: In-Context, Fine-Tuning, RL/RLHF, Distillation, Retrieval/Hybrid, Personalization

   This taxonomy gives rise to a systematic framework for categorizing the capabilities and limitations of current methods. We provide a detailed summary of this taxonomy in Table 1.
3. **An in-depth evaluation framework for assessing caption quality and figure understanding.** We review evaluation metrics for figure caption-

ing, covering reference based measures (e.g., ROUGE, BLEU), human centered criteria (e.g., takeaway, helpfulness), and emerging MLLM as a judge approaches. The survey highlights their limitations and calls for metrics focused on reasoning and content fidelity.

4. **A comprehensive overview of benchmarks and datasets for figure captioning.** We curate and categorize existing datasets based on key characteristics like figure type, task objective, and dataset size.
5. **An overview of key open problems and challenges that future work should address.** We identify key remaining challenges and failure cases that are important to solve in the future. For each open problem, we discuss various aspects to the problem, and potential directions for future work.

This survey offers a unified view of recent advances in figure captioning , providing a unified framework that serves as a basis for studying and developing more advanced approaches, datasets, and evaluations. By proposing intuitive taxonomies for these aspects, we aim to chart a path forward for researchers and practitioners building next-generation models that understand, describe, and interact with visual data.

Scope of this Survey. While there has been a few recent surveys on image captioning (Thobhani et al., 2025; Salgotra et al., 2025), there is no survey focused on figure captioning. In this work, we seek to fill this gap and does not attempt to cover the extensive literature on general image captioning.

2 Problem Formulation

In this section, we formalize the problem of captioning figures and formally define the important problem settings.

2.1 Preliminaries & Notation

Let $\mathcal{I}$ denote the space of figures , *e.g.*, tensors in $\mathbb{R}^{H\times W\times C}$. Let $\mathcal{Y} = \bigcup_{T=1}^{T_{\max}} \mathcal{V}^T$ denote the space of captions where $\mathcal{V}$ is a finite vocabulary. Hence, a caption is $Y = (w_1, \ldots, w_T) \in \mathcal{Y}$ with $w_t \in \mathcal{V}$. Further, let $\mathcal{T} = \bigcup_{L=1}^{L_{\max}} \mathcal{V}^L$ be the space of auxiliary texts (*e.g.*, paragraphs), which is optional. A dataset $\mathcal{D}$ is thus a collection of tuples containing subsets of these modalities. We use a loss $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ (*e.g.*, negative log-likelihood surrogate), and a quality/utility function $Q : \mathcal{Y} \times \mathcal{I} \to \mathbb{R}$ (*e.g.*, human-judgment or metric score). When a probabilistic model is used, let $p_\theta(Y \mid \cdot)$ denote a conditional distribution parameterized by θ. At test time, decoding is $\hat{Y} = \arg\max_{Y\in\mathcal{Y}} s_\theta(Y \mid \cdot)$ where s_θ is a scoring function (e.g., $\log p_\theta$). The goal is to produce a caption $\hat{Y} \in \mathcal{Y}$ that maximizes some notion of quality $Q(\hat{Y}, Y, I)$, where Q is a quality/utility function $Q : \mathcal{Y} \times \mathcal{Y} \times \mathcal{I} \to \mathbb{R}$ that may be instantiated by human judgment or an automatic metric. Note that many quality metrics such as BLEU compare two captions only, hence, $Q(\hat{Y}, Y)$ where $Q : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.

2.2 Input Modalities

We explore four primary input modalities, defined by the input data provided to the model. In the most basic **Figure Only** setting, the model must generate a caption $\hat{Y}$ given only an input figure $I \in \mathcal{I}$, often in a zero-shot manner by leveraging a pretrained model. A natural extension is the **Figure-Caption Pairs** setting, which provides a collection of figures and their corresponding ground-truth captions for supervised training. The **Fig-Cap Pairs + Auxiliary Text** setting further enriches the available information, allowing the model to use surrounding paragraphs, document metadata, or related charts to improve captioning ability. Finally, the **Conversational Captions** setting addresses real-world applications where figures are created through a user's dialogue with a system, using the conversational history as context for generation.

2.3 Figure Types

Figures vary widely in form, each emphasizing different data or relationships. We categorize them into common types such as **bar**, **line**, **scatter**, and **pie** charts that represent core statistical patterns, and specialized types like **heatmaps** and **flowcharts** that show higher dimensional or structural relations. An **other** category covers domain specific or unconventional visualizations. This taxonomy captures real world diversity and informs captioning methods, since each type presents distinct reasoning and description challenges.

3 Input Modality

This section reviews prior studies on MLLM-based figure captioning, organized by the input modalities. We highlight representative approaches and provide a structured view of the diverse methodologies explored in this field.

3.1 Figure Only

Given a figure I, caption $\hat{Y}$ is generated with the equation:

$$\hat{Y} = \arg\max_{Y\in\mathcal{Y}} \; p_{\mathcal{M}}(Y \mid I, P), \qquad (1)$$

where $p_{\mathcal{M}}(Y \mid I, P)$ denotes the probability assigned by model $\mathcal{M}$ to caption Y conditioned on figure I and prompt P. In practice, $\hat{Y}$ is computed via approximate decoding methods such as greedy search, beam search, or nucleus sampling.

Adapting pretrained MLLMs for figure captioning is challenging in data scarce settings. Synthetic supervision is often applied either through in-context learning, which inserts example figure-caption pairs into prompts, or through knowledge distillation, where a teacher model generates captions for unlabeled figures and a student model is finetuned on them. ChartInstruct (Masry et al., 2024a) generates a large-scale instruction-following dataset by using MLLMs to create a diverse set of tasks from the underlying data tables of charts, which is then used to finetune models for general-purpose chart comprehension. ChartGemma (Masry et al., 2024b) follows this paradigm by using a high-capacity MLLM to generate a diverse, visual instruction-tuning dataset directly from chart images, which is then used to fine-tune a student model. In addition, this process can also support Reinforcement Learning from Human Feedback (RLHF) (Mnih et al., 2015; Ouyang et al., 2022) and personalization by generating multiple caption candidates, collecting human preferences to build a reward model, and fine tuning the student model to align with human judgment.

Structured guidance methods (Omasa et al., 2025) represent figures as graphs derived from detected objects and text, improving model grounding. Retrieval based approaches instead provide relevant figure caption pairs from a knowledge base during generation. Finally, accessibility oriented studies (Kakehi et al., 2024) explore how summarizing grouped bars can improve comprehension for visually impaired readers while balancing information load and accuracy.

3.2 Figure-Caption Pairs

In this setting, the training data consists of figure-caption pairs

$$\mathcal{D} = \{(I_i, Y_i)\}_{i=1}^{n}, \quad I_i \in \mathcal{I},\ Y_i \in \mathcal{Y},$$

and the objective is to learn parameters θ such that

$$\hat{\theta} \in \arg\max_{\theta} \sum_{i=1}^{n} \log p_\theta(Y_i \mid I_i)$$

At test time, given an unseen figure I, the model produces

$$\hat{Y} = \arg\max_{Y \in \mathcal{Y}} p_{\hat{\theta}}(Y \mid I),$$

where $\hat{Y}$ serves as the best available description of I.

Recent work has focused on improving the quality and representation of paired data. VisText (Tang et al., 2023) enriches chart representations with tables and scene graphs for semantically richer captions. ChartLlama (Han et al., 2023) and UniChart (Masry et al., 2023) use MLLMs for synthetic dataset creation and knowledge distillation from chart data. To combat the issue of hallucination, ChartCap (Lim et al., 2025) mitigates hallucination through dense, type-specific captions, leading to more factually accurate models. FlowLearn (Pan et al., 2024) introduces a dataset of scientific flowcharts with captions and simulated flowcharts with structural annotations, evaluating models through few-shot in-context learning. Another approach is to enhance the model's internal reasoning capabilities. FigCAP (Chen et al., 2019) incorporates an attention mechanism that explicitly models relationships among chart elements, leading to more comparative and analytical captions. Sasaki (2025) introduce structure-aware contrastive learning that leverages diagram code to generate difficult synthetic examples, encouraging models to capture fine-grained information.

To simplify the end-to-end captioning process. DePlot (Liu et al., 2022) introduces a "plot-to-table translation" step that converts charts into linearized data tables before captioning. StructChart (Xia et al., 2023) similarly transforms charts into structured intermediate representations, which serve as inputs for downstream tasks. ChartAssistant (Meng et al., 2024) adopts chart-to-table translation as a key pretraining stage before multitask instruction tuning, improving generalization across chart types. ChartMoE (Xu et al., 2025) replaces the conventional linear projector with a mixture-of-experts design, capturing diverse and specialized patterns in chart understanding. ChartVLM (Xia et al., 2025) further enhances this pipeline with a cascaded decoder that first extracts structural data and then performs captioning or summarization.

Beyond supervised fine-tuning, RLHF optimizes directly for user preference. FigCaps-HF (Singh et al., 2023) trains a reward model on human ratings of caption quality and fine-tunes the generator accordingly, producing more informative and human-aligned outputs. LaMP-Cap (Ng et al., 2025b) introduces multimodal figure profiles that include related images, captions, and contextual text from the same document, enabling personalized captioning that reflects the domain's or authors' style.

3.3 Fig-Cap Pairs + Auxiliary Text

To formally define this setting, the dataset is:

$$\mathcal{D} = \{(I_i, Y_i, T_i)\}_{i=1}^{n}, \quad T_i \in \mathcal{T}$$

where T_i is the auxiliary text corresponding to the figure-caption pair. Using this triplet data naturally gives rise to two general regimes:

- *Auxiliary text available only at training:* In this setting, we do not have the auxiliary text at inference time. One possible approach is to learn a teacher distribution $p_\theta(Y \mid I, T)$ and distill into a student $q_\phi(Y \mid I)$ that operates without T at test time. Intuitively, despite not having the auxiliary text for inference, by leveraging $\mathcal{D} = \{(I_i, Y_i, T_i)\}_{i=1}^{n}$ for training, we can optimize a model that generates better captions that accurately reflect the figure.

- *Auxiliary text available at both training and test:* When auxiliary text is available for both training and inference, then one possibility is to learn a single conditional model $p_\theta(Y \mid I, T)$ and then leverage it to generate $\hat{Y}$ as:

$$\hat{Y} = \arg\max_{Y \in \mathcal{Y}} p_{\hat{\theta}}(Y \mid I, T)$$

In both regimes, auxiliary text serves as a source of semantic grounding, giving rise to captions that more faithfully reflect the intended meaning of the figures.

Empirical evidence supports the value of integrating textual context. Rojas et al. (Rojas and Carranza, 2024) combine CLIP-based visual embeddings with RoBERTa representations of nearby paragraphs to improve scientific chart captioning accuracy. SciCap+ (Yang et al., 2024) expands the SciCap dataset with mention-paragraphs and OCR tokens, enabling models to leverage both textual and visual signals. Hsu et al. (Hsu et al., 2025) show that conditioning on a figure and its surrounding paragraphs allows GPT-4V to produce captions preferred by expert editors over those written by the original authors.

When auxiliary text is available only during training, distillation effectively transfers contextual knowledge to models deployed without text input. A teacher trained on (I, C, T) pairs provides rich semantic supervision, while a student distilled from it learns to generate captions grounded in figure content alone.

Auxiliary text can be incorporated dynamically through retrieval at inference time. LaMP-Cap (Ng et al., 2025b) retrieves related figures, captions, and paragraphs from the same document to form multimodal profiles that personalize caption generation to an author's tone and style. ChartThinker (Liu et al., 2024) extends this approach by coupling retrieval with reasoning chains, showing that retrieved context enhances coherence and logical accuracy in chart summarization.

3.4 Conversational Captioning

In this setting, the input consists of a figure I together with a conversational history $C = \{(u_1, a_1), (u_2, a_2), \ldots, (u_m, a_m)\}$, where u_j and a_j denote the jth user query and model response, respectively. Since the models prior captions naturally appear in C, the conversation history already contains Y_{prev}, the previously generated captions for I. Thus, the next caption $\hat{Y}$ can be generated as:

$$\hat{Y} = \arg\max_{Y \in \mathcal{Y}} p_\theta(Y \mid I, C)$$

As an aside, it is sometimes useful to explicitly distinguish the most recent caption Y_{prev} from the broader conversational context:

$$\hat{Y} = \arg\max_{Y \in \mathcal{Y}} p_\theta(Y \mid I, C, Y_{\text{prev}})$$

These formulations capture settings such as multimodal assistants or figure-grounded chatbots where the caption must not only describe the visual content but also align with the intent of the user expressed in the conversation. Unlike the figure-only setting, conversational captioning requires grounding in both modalities, including the figure I and conversational C (which often also includes the previously generated captions that are being iteratively refined).

Conversational captioning often becomes iterative refinement where captions evolve based on both visual input and user feedback, with Y_{prev} serving as the anchor for revision along with the user-guided instructions. Pluto (Srinivasan et al., 2025) presents a mixed-initiative system combining MLLMs with heuristic parsing to enable auto-completion, verification, and user-guided refinement of figure captions. Similarly, FigurA11y (Singh et al., 2024) demonstrates a system where MLLM provides interactive suggestions to help authors iteratively refine figure captions.

4 Techniques

This section overviews the main techniques used in various ways to solve the figure captioning problem. For a summary of the techniques and the settings applicable for each, see Table 4 in Appendix. As an

aside, the techniques may be viewed as orthogonal design dimensions, for instance, one can apply in-context learning with retrieval augmentation or fine-tuned model followed by RLHF.

4.1 In-context

In-context techniques (Brown et al., 2020; Wei et al., 2022; Dong et al., 2022; Min et al., 2022) leverage a pretrained model $\mathcal{M}$ in its frozen state and guide its predictions by conditioning on exemplars or prompts at inference time, without modifying the model parameters (Liu et al., 2022; Han et al., 2023; Liu et al., 2024; Ng et al., 2025b; Lim et al., 2025; Omasa et al., 2025; Srinivasan et al., 2025). More formally, given a figure I and k figure-caption pairs $\{(I_j, Y_j)\}_{j=1}^{k}$ along with a prompt P, the model infers the following caption:

$$\hat{Y} = \arg\max_{Y \in \mathcal{Y}} p_{\mathcal{M}}(Y \mid I, \{(I_j, Y_j)\}_{j=1}^{k}, P),$$

where the prompt P encodes instructions and formatting. This approach is particularly appealing in settings where training data is scarce, such as zero- or few-shot regimes, and gives rise to rapid adaptation across new domains. Variants of this paradigm include chain-of-thought prompting (Wei et al., 2022), instruction-oriented prompts, and iterative prompting strategies that progressively refine model outputs.

4.2 Fine-tuning

Fine-tuning (Parthasarathy et al., 2024; Hu et al., 2022) adapts a pretrained base model to the figure–caption distribution by updating its parameters on a dataset $\mathcal{D} = \{(I_i, Y_i)\}_{i=1}^{n}$. The learning objective is typically formulated as cross-entropy under teacher forcing, which is equivalent to maximizing the conditional likelihood of the ground-truth captions:

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} \log p_\theta(Y_i \mid I_i).$$

This objective aligns the model parameters with the empirical data distribution. In contrast to in-context techniques that leave model parameters unchanged, fine-tuning directly modifies them. Practical realizations range from full end-to-end parameter updates to parameter-efficient strategies such as LoRA and adapters (Hu et al., 2022), which make this approach feasible even for extremely large models. Fine-tuning can be domain-specific (e.g., scientific figure-captions, or even sub-domain such as CHI figure-captions compared to figure-captions typically found in ML papers) or user-specific (for personalization, see Sec. 4.6).

4.3 RL & RLHF

Reinforcement learning (RL) (Sutton et al., 1998; Mnih et al., 2015; Schulman et al., 2017; Mnih et al., 2013) techniques optimize captions with respect to a quality metric Q (*e.g.*, CIDEr, CLIP similarity, or a human preference score) rather than a likelihood surrogate. Given a generated caption $\hat{Y}$ for figure I and reference caption Y, a scalar reward is computed as $r = Q(\hat{Y}, Y, I)$. Model parameters θ are then updated to maximize expected reward:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\hat{Y} \sim p_\theta(\cdot \mid I)}\big[\nabla_\theta \log p_\theta(\hat{Y} \mid I)\, r\big],$$

where $p_\theta(\hat{Y} \mid I)$ is the models distribution over captions. Intuitively, captions that correctly capture facts (*e.g.*, describe the figure, or describes relationships, or higher-order insights, etc) receive higher r, which increases their probability under p_θ. Unlike cross-entropy fine-tuning that aligns parameters with the empirical data distribution, RL directly optimizes for the downstream metric of interest. When human preferences are available, RLHF (Christiano et al., 2017; Ouyang et al., 2022) replaces Q with a learned reward model $R_\phi(I, \hat{Y})$ that better matches human judgements, therefore leading to better relevance and faithfulness. As an aside, common examples include PPO-based policy optimization and direct preference optimization (DPO-based) that provide stable updates even for large captioning models (Rafailov et al., 2023; Zhang et al., 2021), among many others (Srivastava and Aggarwal, 2025).

4.4 Distillation & Model Transfer

Distillation and model transfer methods (Hinton et al., 2015; Gou et al., 2021; Xu et al., 2024b) adapt a captioning model by first training a *teacher* that conditions on all available modalities, including the figure $I \in \mathcal{I}$, auxiliary text $T \in \mathcal{T}$, and the ground-truth captions Y_i that are paired with I_i in the dataset. The teacher parameters θ are optimized to maximize the likelihood of Y_i:

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} \log p_\theta(Y_i \mid I_i, T_i).$$

Once trained, its predictive distribution is distilled into a student $q_\phi(Y \mid I)$ with parameters ϕ by minimizing

$$\mathcal{L}_{\text{distill}} = \sum_{i=1}^{n} \mathrm{KL}\Big[p_\theta(\cdot \mid I_i, T_i) \,\|\, q_\phi(\cdot \mid I_i)\Big],$$

where $\mathrm{KL}[\cdot \| \cdot]$ measures divergence between teacher and student outputs. This formulation naturally gives rise to a two-stage process: (i) a teacher trained with full multimodal supervision (I, T, Y) and (ii) a distilled student that preserves the semantic grounding induced by T and Y but requires only I at inference. As a result, the student yields captions that remain faithful to the figure and aligned with the distribution of gold captions while being efficient and deployable in scenarios where auxiliary text is unavailable.

4.5 Retrieval & Hybrid

Retrieval methods serve as a mechanism for augmenting caption generation with non-parametric memory (Liu et al., 2024). These techniques can be leveraged in many different ways and along side most of the previous techniques described to improve performance (hence, hybrid). Given a figure I, a retrieval function $\mathcal{R}(I)$ selects a set of top-k related figure–caption pairs $\{(I_j, Y_j)\}_{j=1}^{k}$ from a corpus, which serve as non-parametric examples providing additional visual and textual context. The model then generates a caption conditioned jointly on the figure and the retrieved parts:

$$\hat{Y} = \arg\max_{Y \in \mathcal{Y}} p_\theta\big(Y \mid I, \mathcal{R}(I)\big).$$

This formulation gives rise to hybrid models in which generation is jointly conditioned on the parametric knowledge captured by p_θ and the non-parametric evidence retrieved by $\mathcal{R}$, yielding captions that combine learned priors with contextually relevant examples. Such methods enable rapid adaptation to new domains and figure styles without full retraining, and they naturally extend to many other types of techniques as well.

4.6 Personalization

Personalization adapts the captioning process to a particular user by conditioning on a representation of user preferences, style, or the users previous interactions (Singh et al., 2023; Ng et al., 2025b). Let U denote a user embedding capturing style cues or historical behavior, and then we have:

$$\hat{Y} = \arg\max_{Y \in \mathcal{Y}} p_\theta(Y \mid I, U),$$

where I is the input figure and $\mathcal{Y}$ is the space of candidate captions. Personalization can be realized through fine-tuning on user-specific data, in-context examples that reflect the user's preferred style, or latent conditioning vectors inferred from the user's previous conversations. Personalizatin of figure-captions for a specific user are very important for many assistive systems, figure-grounded agents, user/persona generated reports, among many others. As an aside, we utilize personalization at the finest granularity; however, it may also be very beneficial to personalize a users experience based on persona's, especially if the user lacks a lot of information that can be used to personalize their experience.

5 Evaluation Metrics

In this section, we discuss evaluation criterion, and the fundamental evaluation methodology that can be used including traditional metrics, human evaluation, and MLLM-as-a-Judge. More details are in Appendix A.

5.1 Traditional Metrics

Traditional figure captioning is typically evaluated with text-similarity metrics such as BLEU (Papineni, 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and CIDEr (Vedantam et al., 2015), which measure N-gram overlap between generated and reference captions. Although efficient, these metrics poorly reflect human judgment and are sensitive to phrasing, length, and reference quality.

5.2 Human Evaluation

Recent studies (Hsu et al., 2025; Ng et al., 2025a; Hsu et al., 2023; Yang et al., 2024; Singh et al., 2023; Liu et al., 2024; Masry et al., 2024b,a; Singh et al., 2024; Ng et al., 2025b; Omasa et al., 2025; Srinivasan et al., 2025) propose richer human centric evaluations emphasizing **informativeness**, **factual correctness**, and **structure** (Masry et al., 2024a,b). Additional criteria such as **hallucination avoidance** (Lim et al., 2025), **reasoning correctness** (Liu et al., 2024), and **helpfulness**, **OCR**, **relation**, **stats**, **takeaway**, and **visual** quality (Hsu et al., 2023; Singh et al., 2023; Ng et al., 2025b) capture deeper semantic and perceptual dimensions. Together, these form a comprehensive framework assessing informational, structural, and interpretive quality.

5.3 MLLM-as-a-Judge

To address the limits of reference based and human evaluations, recent work introduces MLLMs as automated judges (Chen et al., 2024; Gu et al., 2024). Pairwise methods such as AutoJ (Li et al., 2023) and PandaLM (Wang et al., 2023) train open models on preference data for consistent judgments, while score based systems such as GEval (Liu et al.,

2023) and JudgeLM (Zhu et al., 2023) provide aspect wise ratings (e.g., factuality, coherence, completeness) using GPT4 annotations. In figure captioning, GPT4 has proven effective as a zero shot evaluator aligned with expert judgments (Hsu et al., 2023).

6 Datasets

We compile a comprehensive overview of existing datasets that support the development and evaluation of large generative models for figure captioning. Appendix B and Table 2 provide detailed descriptions and dataset comparisons.

7 Open Problems & Challenges

This section discusses open problems and highlights important challenges for future work. Appendix C further ellaborates on these challenges.

7.1 Immersive, Interactive, and Animated Figures

Emerging figure types such as animated, interactive, and immersive visualizations introduce significant challenges for figure captioning. Animated figures require models to reason about motion, temporal continuity, and evolving visual meaning, which most current captioning systems are not equipped to handle. Interactive figures introduce user controlled variability, demanding captions that adapt to shifting contexts and exploration paths. Immersive figures in three dimensional or virtual environments add further complexity due to occlusion, distortion, and viewpoint changes that affect spatial understanding. These challenges point to the need for future work on models that maintain temporal consistency, integrate user context, and reason over three dimensional spatial information.

7.2 Personalized Caption

Caption personalization involves adapting the generation either to the author's style or to the target reader's perspective. The first focuses on reflecting the author's linguistic and interpretive preferences, while the second requires anticipating the reader's background, intent, and potential biases. Future work should explore mechanisms for modeling these dual dimensions of personalization through data design, fine-tuning, and prompt conditioning.

7.3 Figure Editing

A main challenge in figure editing is generating a new figure that correctly reflects the given editing instructions while preserving the original layout and meaning. Models often struggle to modify only the intended parts without introducing visual errors or misleading changes. Future work should focus on maintaining both visual accuracy and data integrity during edits.

7.4 Captioning for Complex Figures

Most works have focused on generating captions for relatively simple chart types such as bar charts or line charts. Captioning for complex figures, such as flow charts or multi-panel figures with heterogeneous chart types, remains challenging. Such complex figures may require more advanced planning-based captioning techniques to ensure correctness and to maximize readability and understandability.

7.5 Attribution-Aware and Fine-Grained Grounding

One important open problem is attribution-aware and fine-grained grounding of captions to the insights, data, legend, and other parts of the figure itself. Achieving such grounding requires accurately linking textual claims to their precise visual or data sources, which is challenging due to overlapping visual elements, implicit references, and varying figure styles. Moreover, existing grounding models often lack the granularity or attribution awareness needed to explain why a particular insight is supported by specific visual evidence.

7.6 Generated Caption Evaluation

Evaluating generated captions requires verifying whether key data-driven insights are both logically supported by the underlying dataset and correctly expressed in the caption. Emerging methods such as MLLM-as-a-Judge, combined with OCR-based text extraction and grounding (Jiang et al., 2019), offer promising automated evaluation strategies for assessing semantic and visual alignment. However, challenges remain in calibrating MLLMs' visual understanding and ensuring stable, human-aligned evaluation across diverse figure types.

8 Conclusions

This survey provides a comprehensive survey on figure captioning with large generative models through a unified framework along four input settings and six techniques. We reviewed an overview of existing methods to generate figure captions, summarized available datasets and benchmarks, discussed the evaluation metrics. Finally, we highlighted open challenges that point toward future research on complex and reliable figure captioning.

9 Limitations

While this survey aims to provide a comprehensive overview of figure captioning, it has several limitations. First, our main taxonomy is centered around the available input provided (figure only, figure-captions, and so on), then we categorize the works further by the general technique leveraged. While this taxonomy is intuitive, we could have also discussed works by other factors such as the categories of charts and other facets to the problem. Nevertheless, we did summarize all the works surveyed along with a few additional facets in Table 1 focused on the chart types supported and evaluation used.

10 Ethics

This survey aims to advance understanding of figure captioning in scientific contexts, and we aim to support researchers in developing robust and accurate figure captioning methods. All datasets referenced are publicly available and used solely for research purposes under their respective licenses. No personally identifiable or sensitive information was collected or used. Therefore, we do not anticipate any ethical issues emerging from this survey.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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

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

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Marc Corio and Guy Lapalme. 1999. Generation of texts for information graphics. In *Proceedings of the 7th European Workshop on Natural Language Generation EWNLG'99*, pages 49–58.

Seniz Demir, Sandra Carberry, and Kathleen F McCoy. 2008. Generating textual summaries of bar charts. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 7–15.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Yansong Feng and Mirella Lapata. 2013. Automatic caption generation for news images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4):797–812.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International journal of computer vision*, 129(6):1789–1819.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

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

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Ting-Yao Hsu, C Lee Giles, and Ting-Hao'Kenneth' Huang. 2021. Scicap: Generating captions for scientific figures. *arXiv preprint arXiv:2110.11624*.

Ting-Yao Hsu, Chieh-Yang Huang, Shih-Hong Huang, Ryan Rossi, Sungchul Kim, Tong Yu, C Lee Giles, and Ting-Hao Kenneth Huang. 2024. 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*, pages 1–9.

Ting-Yao Hsu, Chieh-Yang Huang, Ryan Rossi, Sungchul Kim, C Lee Giles, and Ting-Hao K

Huang. 2023. Gpt-4 as an effective zero-shot evaluator for scientific figure captions. *arXiv preprint arXiv:2310.15405*.

Ting-Yao E. Hsu, Yi-Li Hsu, Shaurya Rohatgi, Chieh-Yang Huang, Ho Yin Sam Ng, Ryan Rossi, Sungchul Kim, Tong Yu, Lun-Wei Ku, C. Lee Giles, and Ting-Hao K. Huang. 2025. Do large multimodal models solve caption generation for scientific figures? lessons learned from scicap challenge 2023.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, Pengchuan Zhang, Zhe Gan, Jana Diesner, and Jianfeng Gao. 2019. Tiger: Text-to-image grounding for image caption evaluation. *arXiv preprint arXiv:1909.02050*.

Kushal Kafle, Scott Cohen, Brian Price, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *CVPR*.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*.

Banri Kakehi, Masakazu Iwamura, Kazunori Minatani, and Koichi Kise. 2024. Grouping effect for bar graph summarization for people with visual impairments. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–6.

Shankar Kantharaj, Rixie Tiffany Ko Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. 2022. Chart-to-text: A large-scale benchmark for chart summarization.

Zeba Karishma, Shaurya Rohatgi, Kavya Shrinivas Puranik, Jian Wu, and C. Lee Giles. 2023. Acl-fig: A dataset for scientific figure classification.

Daniel A Keim. 2002. Information visualization and visual data mining. *IEEE transactions on Visualization and Computer Graphics*, 8(1):1–8.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023. Generative judge for evaluating alignment. *arXiv preprint arXiv:2310.05470*.

Junyoung Lim, Jaewoo Ahn, and Gunhee Kim. 2025. Chartcap: Mitigating hallucination of dense chart captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Fangyu Liu, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. 2022. Deplot: One-shot visual language reasoning by plot-to-table translation. *arXiv preprint arXiv:2212.10505*.

Junteng Liu, Weihao Zeng, Xiwen Zhang, Yijun Wang, Zifei Shan, and Junxian He. 2025. On the perception bottleneck of vlms for chart understanding. *arXiv preprint arXiv:2503.18435*.

Mengsha Liu, Daoyuan Chen, Yaliang Li, Guian Fang, and Ying Shen. 2024. Chartthinker: A contextual chain-of-thought approach to optimized chart summarization. *arXiv preprint arXiv:2403.11236*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. 2023. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. *arXiv preprint arXiv:2305.14761*.

Ahmed Masry, Mehrad Shahmohammadi, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. 2024a. Chartinstruct: Instruction tuning for chart comprehension and reasoning. *arXiv preprint arXiv:2403.09028*.

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

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

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Vibhu O Mittal, Johanna D Moore, Giuseppe Carenini, and Steven Roth. 1998. Describing complex charts in natural language: A caption generation system. *Computational Linguistics*, 24(3):431–467.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533.

Srija Mukhopadhyay, Adnan Qidwai, Aparna Garimella, Pritika Ramu, Vivek Gupta, and Dan Roth. 2024. Unraveling the truth: Do vlms really understand charts? a deep dive into consistency and robustness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16696–16717.

Ho Yin Ng, Ting-Yao Hsu, Jiyoo Min, Sungchul Kim, Ryan A. Rossi, Tong Yu, Hyunggu Jung, and Ting-Hao 'Kenneth' Huang. 2025a. Understanding how paper writers use ai-generated captions in figure caption writing. In *AI for Research and Scalable, Efficient Systems*, pages 173–192, Singapore. Springer Nature Singapore.

Ho Yin'Sam' Ng, Ting-Yao Hsu, Aashish Anantha Ramakrishnan, Branislav Kveton, Nedim Lipka, Franck Dernoncourt, Dongwon Lee, Tong Yu, Sungchul Kim, Ryan A Rossi, et al. 2025b. Lamp-cap: Personalized figure caption generation with multimodal figure profiles. *arXiv preprint arXiv:2506.06561*.

Takamitsu Omasa, Ryo Koshihara, and Masumi Morishige. 2025. Arrow-guided vlm: Enhancing flowchart understanding via arrow direction encoding. *arXiv preprint arXiv:2505.07864*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Huitong Pan, Qi Zhang, Cornelia Caragea, Eduard Dragut, and Longin Jan Latecki. 2024. Flowlearn: Evaluating large vision-language models on flowchart understanding. *arXiv preprint arXiv:2407.05183*.

Kishore Papineni. 2002. Machine translation evaluation: N-grams to the rescue. In *LREC*.

Venkatesh Balavadhani Parthasarathy, Ahtsham Zafar, Aafaq Khan, and Arsalan Shahid. 2024. The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities. *arXiv preprint arXiv:2408.13296*.

Xin Qian, Eunyee Koh, Fan Du, Sungchul Kim, Joel Chan, Ryan A Rossi, Sana Malik, and Tak Yeon Lee. 2021. Generating accurate caption units for figure captioning. In *Proceedings of the Web Conference 2021*, pages 2792–2804.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Jonathan Roberts, Kai Han, Neil Houlsby, and Samuel Albanie. 2024. Scifibench: Benchmarking large multimodal models for scientific figure interpretation. *Advances in Neural Information Processing Systems*, 37:18695–18728.

Mateo Alejandro Rojas and Rafael Carranza. 2024. Enhancing scientific figure captioning through cross-modal learning. *arXiv preprint arXiv:2406.17047*.

Garima Salgotra, Pawanesh Abrol, and Arvind Selwal. 2025. A survey on automatic image captioning approaches: Contemporary trends and future perspectives. *Archives of Computational Methods in Engineering*, 32(3):1459–1497.

Hiroshi Sasaki. 2025. Structure-aware contrastive learning for diagram understanding of multimodal models. *arXiv preprint arXiv:2509.01959*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

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

Nikhil Singh, Lucy Lu Wang, and Jonathan Bragg. 2024. Figura11y: Ai assistance for writing scientific alt text. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI '24, page 886–906, New York, NY, USA. Association for Computing Machinery.

Arjun Srinivasan, Vidya Setlur, and Arvind Satyanarayan. 2025. Pluto: Authoring semantically aligned text and charts for data-driven communication. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 1123–1140.

Saksham Sahai Srivastava and Vaneet Aggarwal. 2025. A technical survey of reinforcement learning techniques for large language models. *arXiv preprint arXiv:2507.04136*.

Richard S Sutton, Andrew G Barto, et al. 1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

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

Alaa Thobhani, Beiji Zou, Xiaoyan Kui, Amr Abdussalam, Muhammad Asim, Sajid Shah, and Mohammed Elaffendi. 2025. A survey on enhancing image captioning with advanced strategies and techniques. *Computer Modeling in Engineering & Sciences (CMES)*, 142(3).

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, et al. 2023. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *arXiv preprint arXiv:2306.05087*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Renqiu Xia, Bo Zhang, Haoyang Peng, Hancheng Ye, Xiangchao Yan, Peng Ye, Botian Shi, Yu Qiao, and Junchi Yan. 2023. Structchart: Perception, structuring, reasoning for visual chart understanding. *arXiv preprint arXiv:2309.11268*.

Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Peng Ye, Min Dou, Botian Shi, Junchi Yan, and Yu Qiao. 2025. Chartx chartvlm: A versatile benchmark and foundation model for complicated chart reasoning.

Peixin Xu, Yujuan Ding, and Wenqi Fan. 2024a. Chartadapter: Large vision-language model for chart summarization.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024b. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*.

Zhengzhuo Xu, Bowen Qu, Yiyan Qi, SiNan Du, Chengjin Xu, Chun Yuan, and Jian Guo. 2025. Chartmoe: Mixture of diversely aligned expert connector for chart understanding. In *The Thirteenth International Conference on Learning Representations*.

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

Zhishen Yang, Raj Dabre, Hideki Tanaka, and Naoaki Okazaki. 2024. Scicap+: A knowledge augmented dataset to study the challenges of scientific figure captioning. *Journal of Natural Language Processing*, 31(3):1140–1165.

Ho Yin, Ting-Yao Hsu, Jiyoo Min, Sungchul Kim, Ryan A Rossi, Tong Yu, Hyunggu Jung, Ting-Hao'Kenneth' Huang, et al. 2025. Understanding how paper writers use ai-generated captions in figure caption writing. *arXiv preprint arXiv:2501.06317*.

Le Zhang, Yanshuo Zhang, Xin Zhao, and Zexiao Zou. 2021. Image captioning via proximal policy optimization. *Image and Vision Computing*, 108:104126.

Zhehao Zhang, Weicheng Ma, and Soroush Vosoughi. 2024. Is GPT-4V (ision) all you need for automating academic data visualization? exploring vision-language models' capability in reproducing academic charts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8271–8288, Miami, Florida, USA. Association for Computational Linguistics.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. Judgelm: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631*.

A Additional Details on Evaluations

A.1 Traditional metrics

Early figure captioning systems were evaluated using metrics borrowed from machine translation and summarization. In general, these metrics evaluate the generated caption by comparing it with human-generated reference captions using n-gram analysis. Specifically, **BLEU** (Papineni, 2002) measures n-gram precision against reference captions. **ROUGE** (Lin, 2004) focuses on recall, comparing overlapping sequences between candidate and reference texts. **METEOR** (Banerjee and Lavie, 2005) introduced the harmonic mean of precision/recall with stemming, synonym matching, and paraphrase tables for improved flexibility. In the general image-captioning literature, **CIDEr** (Vedantam et al., 2015) was proposed specifically for captions, using TF–IDF weighted n-gram matching to reward consensus across multiple human references. While these methods offer computational efficiency and consistency in evaluation, they suffer from multiple drawbacks: **(1) low agreement with human judgment** when captions are phrased differently than references and **(2) sensitivity to sentence length and detail**, where descriptive captions that include additional correct details not present in references are unreasonably penalized and finally **(3) over-dependence on good reference captions** wherein, effectiveness of these methods is limited by the quality of the reference captions, which can often be inadequately written.

A.2 Human Evaluation

Human evaluation methods can be broadly categorized based on following high-level criteria:
Evaluation Form: A common approach followed by most works (Yang et al., 2024; Singh et al., 2023; Hsu et al., 2023; Liu et al., 2024; Masry et al., 2024b,a) is to define specific qualitative metrics for evaluation and record human evaluator ratings on a Likert scale. These scores are then aggregated across samples and subjects to generate absolute ratings. Contrary to this, some methods rely on ranking generated captions (Hsu et al., 2023; Lim et al., 2025; Ng et al., 2025b) based on specific qualitative properties. Finally, in (Omasa et al., 2025), human evaluation focuses on single-item judgment (binary) where correctness is determined by comparing the predicted caption with ground-truth caption.

Qualitative Metrics: (Masry et al., 2024a,b) defines **Informativeness**, **Factual correctness** and **Structure** as qualitative criteria for human evaluation. They are defined as:

1. **Informativeness:** *How much useful and relevant information from the chart does the summary cover? Does it effectively convey the main trends and insights?*

2. **Factual correctness:** *How accurate is the summary in reflecting the information presented in the chart?*

3. **Structure:** *How well-structured is the summary? Does it include an introduction, a body with key insights, and a conclusion?*

Apart from informativeness, (Lim et al., 2025) also defines **Hallucinations** as another key criterion, which is formulated as: *Does the caption avoid information that cannot be inferred from the chart?* Other key metrics includes: **Reasoning Correctness** (Liu et al., 2024), **Helpfulness** (Hsu et al., 2023; Singh et al., 2023; Ng et al., 2025b), **OCR (Optical Character Recognition)** (Hsu et al., 2023; Singh et al., 2023), **Relation**, **Stats** (Hsu et al., 2023), **Takeaway** and **Visual** (Hsu et al., 2023; Singh et al., 2023). They are defined as follows:

1. **Helpfulness:** *Does this caption help you understand the figure?*

2. **OCR (Optical Character Recognition):** *Does this caption mention any words or phrases that appear in the figure?*(Examples include the figure title, X or Y axis titles, legends, names of models, methods, subjects, etc.)

3. **Relation:** *Does this caption describe a relationship among two or more elements or subjects in the figure?* (For example, "A is lower than B," "A is higher than B," or "A is the lowest/highest in the figure.")

4. **Stats:** *Does this caption mention any statistics or numbers from the figure?* (For example, "20% of..." or "The value of .. is 0.33...".)

5. **Takeaway:** *Does this caption describe the high-level takeaways, conclusions, or insights the figure tries to convey?*

6. **Visual:** *Does this caption mention any visual characteristics of the figure?* (Examples include color, shape, direction, size, position, or opacity of any elements in the figure.)

While human evaluation schemes provides an extremely precise and reliable metric for evaluating generated figure captions with respect to reader

preferences, it is costly and time-consuming. Human evaluation is primarily constrained by the need for domain expertise and a thorough understanding of the entire scholarly article by the human evaluators. This severely limits its scalability and makes solely relying on human evaluation impractical for large-scale evaluation. Thus while it is recommended to conduct human evaluation, it should also be accompanied with automatic evaluation methods.

A.3 MLLM-as-a-Judge

To address the shortcomings of human and traditional reference based evaluation, recent work has begun treating MLLMs as reference-free evaluators (Chen et al., 2024; Gu et al., 2024). These MLLM-based judges can provide scalable, context-sensitive assessments that approximate expert judgment without needing human-authored references. In particular, pairwise comparison has proven especially effective: LLM judgments on paired outputs tend to align closely with human preferences.For example, systems like Auto-J (Li et al., 2023) and PandaLM (Wang et al., 2023) adopt this approach by fine-tuning open-source MLLMs on pairwise preference data. In practice they train on data labeled by strong models (e.g. GPT-3.5/GPT-4 (Achiam et al., 2023)) and generate an evaluator that chooses the better of two responses, yielding judgments that are far more consistent and robust than classical metric scores.

Complementing pairwise methods, score-based frameworks have emerged to rate multiple quality dimensions. For instance, G-Eval (Liu et al., 2023) and JudgeLM (Zhu et al., 2023) use Likert-style ratings to score aspects like factuality, coherence, and completeness. These systems are trained on datasets annotated by powerful LLMs (such as GPT-4 (Achiam et al., 2023)), which helps their scores better mimic human evaluators. Notably, (Hsu et al., 2023) demonstrated the approach in the figure-caption domain: they built a dataset of scientific figure captions and showed that GPT-4 can serve as a zero-shot caption-quality judge. In their work, GPT-4 was prompted (with the figure and context) to score each caption on a 1–6 scale based on how well it aids reader understanding – all without any human-written reference caption. GPT-4's scores even correlated well with expert rankings, illustrating a cost-effective, reference-free evaluation of captions.

B Details on Benchmarks and Datasets

In this section, we summarize the existing datasets that enable the development and evaluation of large generative models for figure captioning and reasoning tasks. We summarize the datasets and their properties in Table 2.

There are many different figure-captioning benchmark datasets. FigCAP (Chen et al., 2019) builds upon FigureQA images to provide both high-level (FigCAP-H) and detailed (FigCAP-D) descriptions for five chart types. It enhances linguistic diversity through paraphrasing templates and introduces logically structured long-form captions, posing greater challenges for multimodal models. SciCAP (Hsu et al., 2021) is a dataset of over 416,000 scientific charts extracted from arXiv papers. Its multi-granularity captioning design and strict train-validation-test split establish a new benchmark for understanding visual content. SciCap+ (Yang et al., 2024) is an enhanced dataset for scientific figure captioning, extending SciCap with mention-paragraphs and OCR tokens to improve model comprehension. It supports multimodal tasks like automated caption generation and scientific summarization.

One recent work by Ng et al. (2025b) focused on personalized caption generation focused on generation captions of figures that align with the style of a specific user. In particular, they introduce a benchmark called LaMP-Cap that leverages the large corpus of SciCAP in the style of a specific user extends SciCAP with personalized context by including up to three background charts and descriptions per target chart. It enables research on context-aware, personalized multimodal generation by aligning captions with individual writing styles.

A few other benchmarks have focused on a slightly different setting of figure-captioning that focused on generating a summary from an input figure (Kantharaj et al., 2022; Xu et al., 2024a) and also on generating captions from a large variety of traditional and more complex chart types (Karishma et al., 2023). One such work called chart-to-text (Kantharaj et al., 2022) introduced a large-scale benchmark for the chart summarization task. The dataset consists of two subsets with a total of 44k charts, covering a wide range of topics and chart types. It is designed to generate natural language summaries to help people infer key information from charts. This work also introduces several neural network-based models as baselines for tackling the chart-to-text generation task. Similarly, another more recent work called ChartSumm (Xu et al., 2024a) introduced a benchmark consisting

of over 190,000 samples across diverse chart types. By providing semantically aligned data, it provides semantically aligned summaries to support multilingual and cross-domain chart understanding, aiding the development of large-scale multimodal summarization models. Finally, one recent benchmark called ACL-Fig (Karishma et al., 2023) focused on many different diverse and complex figure types (flowcharts and beyond). This benchmark includes 112,052 scientific figures extracted from 56,000 ACL papers, including a manually labeled subset (1,671 figures, 19 classes) called ACL-Fig-PILOT, designed for figure classification and captioning.

Figure understanding is also a closely related problem that can also be leveraged to improve figure captioning in more advanced techniques, since figure understanding is crucial to generate the best caption for a specific figure. One such dataset is the FigureQA benchmark (Kahou et al., 2017) which is a large-scale synthetic benchmark designed to test visual reasoning on chart images. It includes over 1.3 million yes/no questions grounded in more than 100,000 charts. Differentiated color schemes and auxiliary annotations (e.g., bounding boxes) prevent shortcut learning, making it a key benchmark for visual reasoning on structured figures. DVQA (Kafle et al., 2018) is a visual question answering dataset focused on bar chart understanding. It contains 300,000 synthetic images and 3.48 million question-answer pairs, covering three tasks: structure understanding, data retrieval, and logical reasoning. The charts are generated using Matplotlib with diverse styles and are supplemented with detailed metadata such as bar bounding boxes and underlying table data. More recently, SciFIBench (Roberts et al., 2024) evaluates the scientific figure interpretation of MLLMs using adversarially-filtered multiple-choice questions that require matching figures to their correct captions.

A few other recent works have focused on other aspects of this problem. ChartX (Xia et al., 2025) is a multimodal benchmark featuring 18 chart types and 22 scientific domains. Each sample contains an image, CSV data, Python code, and textual description. Its dual-layer tasks—perception and cognition—enable comprehensive evaluation of chart understanding in LLMs. SCICAP-EVAL (Hsu et al., 2023) is an evaluation dataset with 600 scientific charts and 3,600 captions (human- and model-generated). It introduces a 1–6 readability score and expert annotations to assess caption quality, offering a robust framework for comparing automatic captioning methods.

C Additional Details on Open Challenges

C.1 Animated Figure

Animated figures evolve automatically over time, allowing temporal changes in data to be communicated through motion. Such dynamics are typically author-driven, as seen in animations that visualize evolving phenomena or transitions across datasets. Data visualization often employs animation to represent datasets that unfold along an additional temporal dimension. Examples include heatmaps of eye-tracking data, where gaze density changes across time, or airflow simulations visualized over geographic maps. Similarly, scientific visualizations such as CT (Computer Tomography) or MRI (Magnetic Resonance Imaging) scans sequentially present slices of 3D objects to convey volumetric structures. As in computer vision, where models analyze object tracking, segmentation, and motion patterns, animated figures could similarly incorporate such analytical reasoning into captions to reflect temporal transitions more accurately. Building on video understanding capabilities, extending multimodal models to animated figures could enable more comprehensive and temporally aware captioning of visualization content.

C.2 Interactive Figure

Interactive figures, in contrast, invite reader-driven exploration by enabling users to manipulate what is shown or emphasized. Through techniques such as filtering, zooming, or brushing and linking, readers can dynamically shift focus and reveal different aspects of a dataset (Keim, 2002). These interactions alter how relationships among data points are represented, meaning that the same dataset may produce multiple narratives depending on user intent and engagement. Understanding how such representational flexibility affects caption generation calls for models that integrate awareness of interaction context. Rather than generating isolated captions for each state, models could incorporate exploration history or user actions to produce captions that align with the reader's interpretive trajectory, bridging the gap between static description and dynamic sense-making.

C.3 Immersive Figure

Data visualization has also been actively studied across different modalities. A representative example is rendering 3D data visualization in VR and AR environments rather than on traditional 2D displays. Captioning a snapshot of such a 3D figure, or generating captions that incorporate auxiliary information from immersive interactions such as free

exploration or guided tours, could be an interesting research direction.

Because data points and compositional elements exist in 3D space, issues such as occlusion and distortion depending on camera settings often arise. Captioning methods that can account for these visual characteristics of immersive environments would be valuable. Moreover, problems commonly studied in computer vision for video, such as object segmentation, classification, and tracking, could be extended to this setting.

C.4 Personalized Caption

Caption personalization can be considered from two complementary perspectives. First, personalization for transferring the author's writing style into the model. This includes adapting to the author's preferred style of interpretation and explanation, adherence to auxiliary texts, or alignment with the persona of the author as the narrator. Research in this direction could explore how author-specific characteristics can bias or guide the caption generation process. Second, personalization for the target reader of the generated caption. When the author has a particular audience persona in mind, the generation pipeline should accommodate the author's intentions and design considerations. However, authors may often have only limited knowledge of their readers, or may want to reduce uncertainty by leveraging the model's ability to anticipate reader needs. This involves considerations such as what kind of experience the reader might have, what responses they might show, whether visual perception could lead to biased interpretations, or whether misleading representations arise due to the reader's mental model. If the model is not already aware of these factors, benchmarks and strategies based on fine-tuning or prompt engineering could be proposed to help address them.

C.5 Accurate Evaluation

For evaluating the generated captions, it is often necessary to require the inclusion of a specific argument or data-driven takeaway, and to verify whether the generated caption successfully contains this insight. To address this, an evaluation framework should automatically assess two aspects: (1) whether the generated insight is logically supported by the underlying dataset, and (2) whether this insight is indeed entailed in the generated caption. Developing evaluation metrics through NLP-based approaches, or conducting a comprehensive investigation into whether stable and reliable evaluation is feasible, would be valuable directions. MLLM-as-a-Judge is emerging as a promising solution, as it can rapidly apply diverse evaluator perspectives to analyze the relationship between a provided figure and a generated caption. Combined with robust text extraction methods such as OCR, MLLMs can help ensure accurate chart information retrieval while also supporting evaluation dimensions that may otherwise remain subjective. However, concerns remain that MLLMs' ways of visually interpreting figures are not always aligned with human perception, highlighting the need to calibrate these models and to training them for a broader space of figures. In addition, evaluation strategies from general image captioning can be adapted. For example, one could regenerate a figure from the information contained in a generated caption and then assess the degree of visual or semantic alignment with the original figure. Another approach is to decompose the image and caption into visual elements and analyze their correspondence. For example, text–image grounding methods like TIGEr (Jiang et al., 2019) evaluate which parts of a caption are matched to which regions of an image. In the case of figures, visual regions like axes, legends, or data marks can often be defined algorithmically, making this approach especially feasible for chart caption evaluation.

D Applications

Figure captioning serves as a foundational capability across a wide range of domains, enabling more effective interpretation, interaction, and automation of visual content (Yan et al., 2025). In particular, it gives rise to countless applications, from making figures more accessible to better knowledge extraction and beyond. In this section, we highlight a few important application areas where figure captioning has become increasingly relevant.

D.1 Accessibility

One important application of figure captioning is for improving the accessibility of charts (Yan et al., 2025). By converting figures and diagrams into natural language, it enables screen readers and other assistive tools to convey visual information that would otherwise be lost. This capability gives rise to scalable alt-text generation, reducing manual annotation effort and supporting compliance with accessibility standards such as Web Content Accessibility Guidelines (WCAG). It also naturally lends itself to adaptive interfaces, including educational platforms and interactive dashboards, where dynamic captioning improves usability and inclusivity. Further, figure captioning plays a critical role in enhancing document accessibility. Auto-

Dataset Name	Data Size (Images / Text)	Visualization Types	Core Task
FigCAP (Chen et al., 2019)	100k images / 1.2M captions	Line, Dot-Line, Bar, Pie	Caption Generation
SciCAP (Hsu et al., 2021)	416k plots / 416k captions	Line, Scatter, Histogram, Bar	Scientific Captioning
SciCAP+ (Yang et al., 2024)	414k charts / 394k captions	Scientific figures (enhanced)	Knowledge-augmented captioning
LaMP-Cap (Ng et al., 2025b)	307k images / 110k captions	Line, Bar, Scatter, Histogram	Personalized Captioning
Chart-to-Text (Kantharaj et al., 2022)	44,096 images / 44,096 captions	Line, Bar, Pie, etc.	Chart-to-Text generation
ChartSumm (Xu et al., 2024a)	190k images / 1 90k summaries	Line, Bar, Pie, Scatter, Histogram	Chart Summarization
ACL-Fig (Karishma et al., 2023)	112k figures / 1.67k labeled	19 academic figure types	Classification,QA,Captioning
FigureQA (Kahou et al., 2017)	100k images / 1.32M questions	Line, Dot-Line, Bar, Pie	Visual QA (Yes/No)
DVQA (Kafle et al., 2018)	300k images / 3.48M QA pairs	Vertical Bar, Horizontal Bar	Structure, Data, Reasoning
SciFIBench (Roberts et al., 2024)	- / 2k QA pairs	Line, Dor-Line, Bar, Line	Figure Interpretation
ChartX (Xia et al., 2025)	48k charts / 48k multimodal sets	Heatmap, Candlestick, Bar, Line, etc.	Multi-task Evaluation
SCICAP-EVAL (Hsu et al., 2023)	600 images / 3.6k captions	Line, Bar, Scatter (CV, NLP, HCI)	Caption Quality Evaluation

Table 2: Summary of key chart understanding datasets used for training and evaluating MLLMs. Each dataset varies in image/text scale, covered chart types, and core tasks.

Symbol	Description
$\mathcal{I}$	Space of figures, e.g., tensors in $\mathbb{R}^{H\times W\times C}$
$\mathcal{Y}$	Space of all possible captions composed of tokens from the vocabulary
$\mathcal{V}$	Vocabulary of textual tokens used in captions
$Y=(w_1,\ldots,w_T)$	Caption consisting of T tokens $w_t\in\mathcal{V}$, t is the time variable
$\mathcal{T}$	Space of auxiliary texts, such as paragraphs or surrounding context
$\mathcal{D}$	Dataset of tuples containing figures, captions, and possibly auxiliary text
$\ell(\cdot,\cdot)$	Loss function between predicted and true captions (e.g., negative log-likelihood)
$Q(\cdot)$	Quality or utility function for evaluating captions (e.g., BLEU, human judgment)
$p_\theta(Y\mid\cdot)$	Conditional probability of caption Y given input(s) under model parameters θ
$s_\theta(Y\mid\cdot)$	Scoring function (e.g., $\log p_\theta$) used during decoding
$\hat{Y}$	Predicted caption
$\hat{\mathbb{Y}}$	Set of predicted captions
I	Input figure to be captioned
T	Auxiliary text associated with the figure
C	Conversational history containing user and model utterances
U	User or persona embedding used for personalized captioning
$Q(\hat{Y},Y,I)$	Caption evaluation metric comparing generated caption $\hat{Y}$ with reference Y and figure I
$p_{\mathcal{M}}(Y\mid I,P)$	Probability assigned by a pretrained model $\mathcal{M}$ given figure I and prompt P
P	Text prompt or instruction used during inference or in-context learning
k	Number of in-context examples or retrieved figure–caption pairs
$\mathcal{R}(I)$	Retrieval function returning top-k relevant figure–caption pairs for figure I
$q_\phi(Y\mid I)$	Student model's conditional distribution after distillation
$p_\theta(Y\mid I,T)$	Teacher model's distribution given both figure and auxiliary text
r	Reward in RL or RLHF computed from metric or human preference
$R_\phi(I,\hat{Y})$	Learned reward model approximating human judgments
$\hat{\theta},\hat{\phi}$	Optimal or fine-tuned model parameters
$\mathbb{X},\mathbb{Y},\mathbb{S},\mathbb{G}$	Abstract spaces used for figures, captions, sentences, or graphs

Table 3: List of mathematical notations and their meanings used in the paper.

matically generated descriptions of figures serve as alt-text in compliance with web accessibility standards, supporting screen readers for visually impaired users. This capability ensures that charts are equally accessible to all users, regardless of visual ability.

D.2 Scientific Document Understanding

A primary application of figure captioning lies in the automated understanding of scientific documents. Generating accurate and context-aware captions enables more effective document summarization, improves figure indexing and retrieval across large-scale repositories (e.g., arXiv, PubMed), and supports a variety of downstream tasks such as citation analysis and scientific QA. Furthermore, captioned figures assist reviewers and editors in quickly grasping the intent and insight behind visualizations.

D.3 Knowledge Extraction and Structured Data Generation

Figure captioning enables the transformation of unstructured visual content into machine-readable form. Captions can be leveraged to extract entities and relations, populate multimodal knowledge graphs, and align figure data with textual evidence. In data-centric domains, this facilitates semantic indexing, metadata enrichment, and automatic report generation from charts and tables.

D.4 Cross-modal Retrieval and Search

High-quality figure captions serve as a semantic bridge between visual figure and textual modalities. They enable figure-to-text and text-to-figure retrieval, improving figure search and recommendation systems. In addition, figure captions enhance the performance of multimodal QA and reasoning

Table 4: Taxonomy of input modalities and corresponding technique categories for figure captioning.

Input Modality	In-Context (§4.1)	Fine-Tuning (§4.2)	RL / RLHF (§4.3)	Distillation (§4.4)	Retrieval/Hybrid (§4.5)	Personalization (§4.6)
Figure Only	✓	✓		✓		
Figure–Caption Pairs	✓	✓	✓	✓	✓	✓
Auxiliary Text	✓	✓			✓	✓
Conversational Captioning	✓					

systems by grounding language understanding in figure evidence. This can greatly improve the quality of such systems.

D.5 Educational Technology

In education, figure captioning improves learner engagement and comprehension by providing natural language explanations of complex figures and diagrams. Captioned figures can be used in intelligent tutoring systems, interactive textbooks, and personalized learning environments to provide on-demand clarification, contextual feedback, and improved learning. These capabilities are particularly impactful in STEM education where figures are central to instruction.

D.6 News and Media Automation

In journalism and media, figure captioning enables scalable content generation and summarization for data-driven stories. Automatically generated captions can describe charts, graphs, and infographics used in news articles, dashboards, and social media. This capability supports automated writing pipelines and improves reader comprehension by contextualizing figures with relevant narrative. Furthermore, it can also help to personalize the caption to improve the ability of the user to understand and comprehend the figure in an easy and accessible fashion that is personalized specifically for that user.

D.7 Human-AI Collaboration

Figure captioning gives rise to new forms of collaborative workflows between humans and AI. For instance, (Yin et al., 2025) explores how writers leverage AI-generated captions for writing figure captions. In creative and analytical settings, captions can assist users by automatically generating textual narratives for figures, suggesting titles or annotations, and supporting data storytelling. These capabilities streamline content creation and support interactive authoring tools that adapt to user intent and context.

D.8 Digital Archives and Libraries

For historical and technical archives, figure captioning provides a mechanism to enrich content with missing or outdated annotations. This is particularly valuable in digitization workflows, where legacy documents often contain figures with degraded or absent descriptions. Captioning facilitates improved organization, retrieval, and semantic analysis of figure content in archival collections.