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.

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; Li et al., 2024, 2025a). 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 in figures. As illustrated in Figure 1, high-quality captions may integrate six key dimensions, ranging from basic OCR and visual statistics to high-level takeaways and helpfulness.

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

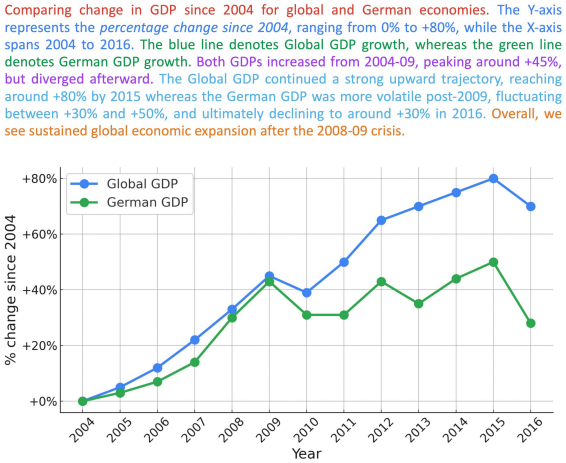

Figure 1: Overview of the figure-captioning task. High-quality captions integrate six criteria, highlighted in color to align with the example caption: **helpfulness** (improves understanding), **OCR** (mentions words from figure), **visual** (mentions visual characteristics, color, etc), **relations**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

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

	Modalities (Section 3)				Techniques (Section 4)						Figure Types (Section 2.2)							Evaluation (Section 5)		
	Figure Only (§3.1)	Figure-Caption Pairs (§3.2)	Fig-Cap + Aux. Text (§3.3)	Conversational (§3.4)	In-Context (§4.1)	Fine-Tuning (§4.2)	RL / RLHF (§4.3)	Distillation (§4.4)	Retrieval (§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., 2025c)	✗	✓	✓	✗	✓	✗	✗	✗	✗	✓	✓	✓	✓	✓	✗	✗	✗	✓	✓	✗
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, figure types, and evaluation metrics.

leading to significantly better captions.

Motivated by the increasing importance of this 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 captioning, 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.

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. Table 2 also shows the trending of figure captioning research.

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. A related survey on chart understanding (Huang et al., 2024a) covers a wider landscape of visualization analysis but does not deeply examine figure captioning as a standalone multimodal problem. Note that figure caption is a self-contained description intended to stand alone and summarize the salient information in the figure, while QA/understanding is conditioned on a specific question and is evaluated for answering that query, and the response is narrow. Our work complements this by presenting, to our knowledge, the first focused and systematic survey of the figure-captioning task, its modeling paradigms, evaluation dimensions, and remaining challenges. Appendix E illustrates the difference between figure captioning and image captioning.

2 Problem Formulation

2.1 Preliminaries & Notation

Let $\mathcal{I}$ be the space of figures, where $I \in \mathbb{R}^{H\times W\times C}$ and $\mathcal{Y} = \bigcup_{T=1}^{T_{\max}} \mathcal{V}^T$ is the space of captions over vocabulary $\mathcal{V}$. A caption is $Y = (w_1, \ldots, w_T) \in \mathcal{Y}$. We use C to denote *additional context* available at generation time, with domain $\mathcal{C}$, such as surrounding text, document metadata, retrieved items, or dialogue history. Captioning is posed as conditional decoding:

$$\hat{Y} = \arg\max_{Y\in\mathcal{Y}} s_\theta(Y \mid I, C), \qquad (1)$$

where $s_\theta(\cdot)$ is a scoring function (e.g., $\log p_\theta$). Evaluation compares $\hat{Y}$ with reference Y via a utility $Q(\hat{Y}, Y, I)$. All notations are shown in Table 4.

2.2 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, there is no additional context, we set $C = \phi$. In practice, $\hat{Y}$ is obtained via approximate decoding, such as greedy, beam, or nucleus sampling. Adapting pretrained MLLMs to figure captioning is difficult under limited labeled data. Prior work therefore relies on synthetic supervision via in-context exemplars or teacher-student distillation. ChartInstruct (Masry et al., 2024a) creates large-scale instruction data from chart tables, and ChartGemma (Masry et al., 2024b) extends this by generating visual instruction-tuning data directly from chart images to fine-tune student models. The same pipeline can support preference alignment and personalization by collecting human feedback, and training preference-aligned models, such as RLHF (Christiano et al., 2017).

Beyond synthetic data, structured guidance methods (Omasa et al., 2025) represent figures as graphs built from detected objects and text to improve grounding. Accessibility focused studies (Kakehi et al., 2024) examine summaries such as grouped bars that balance informativeness, accuracy, and cognitive load for visually impaired readers.

3.2 Figure-Caption Pairs

In this setting, the dataset is $\mathcal{D} = \{(I_i, Y_i)\}_{i=1}^n$ with $I_i \in \mathcal{I}$ and $Y_i \in \mathcal{Y}$. In training time, we set $C = Y_i$ and learn θ by maximizing $\sum_{i=1}^n \log p_\theta(Y_i \mid I_i)$. At test time, $C = \phi$.

Recent work improves paired data quality and supervision. VisText (Tang et al., 2023) augments

Year Range	# Papers	Key Model Architecture	Representative Datasets
pre-2018	23	CNNs, RNNs/LSTMs, and early attention models	DVQA (Kafle et al., 2018), FigureQA (Kahou et al., 2017)
2019	11	Transformer-based encoder–decoder models	FigCap (Chen et al., 2019)
2020	16		PlotQA (Methani et al., 2020)
2021	24		SciCap (Hsu et al., 2021)
2022	40	Vision–language pre-training	DePlot (Liu et al., 2022)
2023	60		VisText (Tang et al., 2023)
2024	63	Large VLMs and instruction-tuned VLMs	SciCap+ (Yang et al., 2024)
2025	72		ChartX (Xia et al., 2025), LaMP-Cap (Ng et al., 2025b)

Table 2: Approximate publication volume and dominant modeling trends by period. Publication counts were estimated through a manual Google Scholar search using a consistent set of survey-relevant keywords, including "figure captioning," "chart captioning," "scientific figure description," "diagram captioning," "plot captioning," and their combinations with representative model architectures. The complete list of papers is provided in Appendix H.

charts with tables and scene graphs, while ChartLlama (Han et al., 2023) and UniChart (Masry et al., 2023) leverage MLLMs for synthetic data creation and distillation. To reduce factual errors, ChartCap (Lim et al., 2025) proposes dense, type specific captions, and FlowLearn (Pan et al., 2024) contributes annotated scientific flowcharts for few shot evaluation.

Other approaches strengthen reasoning and simplify generation. FigCAP (Chen et al., 2019) models relations among chart elements, and structure aware contrastive learning (Sasaki, 2025) uses diagram code to construct hard examples that encourage fine grained understanding. A common pipeline converts charts to structured intermediates before generation, as in DePlot (Liu et al., 2022), StructChart (Xia et al., 2023), ChartAssistant (Meng et al., 2024), and ChartVLM (Xia et al., 2025), while ChartMoE (Xu et al., 2025) improves capacity via specialized experts. Beyond supervised tuning, FigCaps-HF (Singh et al., 2023) applies preference optimization, and LaMP-Cap (Ng et al., 2025c) supports personalized captioning using multimodal figure profiles.

3.3 Figure-Caption 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_θ with $C = T_i$ and distill into a student q_ϕ that operates without C at test time. Intuitively, despite not having the auxiliary text for inference, by leveraging $\mathcal{D}$ 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, we set $C = T_i$. One possibility is to learn a single conditional model $p_\theta(Y \mid I, C)$ and then leverage it to generate $\hat{Y}$.

Auxiliary text provides semantic grounding that improves caption fidelity. Prior work shows consistent gains from leveraging nearby context. Rojas et al. (Rojas and Carranza, 2024) fuse visual features with paragraph representations, SciCap+ (Yang et al., 2024) adds mention paragraphs and OCR tokens to enable joint visual and textual modeling, and Hsu et al. (Hsu et al., 2025) find that conditioning GPT 4V on surrounding paragraphs yields captions preferred by expert editors over author written ones.

When text is available only during training, teacher student distillation transfers contextual knowledge to a text-free captioner. When text is available at inference, retrieval can supply it on demand. LaMP Cap (Ng et al., 2025c) builds multimodal figure profiles for personalized generation, and ChartThinker (Liu et al., 2024) combines retrieval with explicit reasoning to improve coherence and logical accuracy.

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 $\hat{Y}$ is 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:

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

This formulation captures multimodal assistants and figure grounded chat systems, where the goal is not only to describe visual content but also to follow user intent expressed in the dialogue. Compared to figure only captioning, effective generation requires grounding in both modalities and maintaining conversational consistency.

Practically, conversational captioning is often an iterative refinement process. Captions evolve through successive turns based on visual evidence and user feedback, with Y_{prev} serving as the anchor for revision and the conversation providing constraints and preferences. Pluto (Srinivasan et al., 2025) illustrates a mixed initiative workflow that combines MLLMs with heuristic parsing to support auto completion, verification, and guided edits, while FigurA11y (Singh et al., 2024) demonstrates interactive suggestions that help authors incrementally improve 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 5 in Appendix. In Appendix F, we also provide the mapping of input availability and figure complexity to model design choices.

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., 2025c; 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 especially attractive in data scarce settings, including zero and few shot regimes, enabling rapid adaptation to new domains. Common variants include chain of thought prompting (Wei et al., 2022), instruction style prompts, and iterative prompting that progressively refines outputs. While in context learning provides a flexible, training-free way to leverage pretrained MLLMs, performance is highly sensitive to prompt design and often yields weaker grounding than approaches that adapt model parameters.

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 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 fits model parameters to the empirical training distribution. Unlike in-context methods that keep parameters fixed, fine-tuning updates them directly, ranging from full end-to-end optimization to parameter-efficient alternatives such as LoRA and adapters (Hu et al., 2022). Fine-tuning can target specific domains or individual users (see Sec. 4.6), enabling strong task adaptation and typically more stable, faithful captions. However, it requires labeled data and can overfit or reduce generality when the fine-tuning distribution is narrow.

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, 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 better capture factual content and relationships receive higher reward r, increasing their likelihood under p_θ. Unlike cross entropy fine tuning, which matches the empirical training distribution, reinforcement learn-

ing optimizes the target metric directly. With human preferences, RLHF (Christiano et al., 2017; Ouyang et al., 2022) replaces Q with a learned reward model $R_\phi(I, \hat{Y})$ to better reflect human judgment, improving relevance and faithfulness. Common optimization choices include PPO (Schulman et al., 2017) and DPO (Rafailov et al., 2023), which have been applied to large captioning models (Rafailov et al., 2023; Zhang et al., 2021), among others (Srivastava and Aggarwal, 2025). While this paradigm targets human preferred qualities such as informativeness and factuality, it depends on reliable reward signals and careful tuning, and can be costly and unstable in practice.

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

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

where $\mathrm{KL}[\cdot \parallel \cdot]$ 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. Although distillation effectively compresses richer teacher supervision into smaller models, the student may inherit teacher biases or lose fine-grained reasoning that is not preserved during transfer.

4.5 Retrieval

Retrieval augments caption generation with non-parametric memory (Liu et al., 2024) and is commonly used as a component in hybrid pipelines. Given a figure I, a retrieval function $\mathcal{R}(I)$ selects top-k related figure–caption pairs $\{(I_j, Y_j)\}_{j=1}^{k}$ from a corpus, providing additional visual and textual evidence beyond the parametric knowledge in p_θ. The caption is then generated by conditioning jointly on the input figure and the retrieved examples:

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

This formulation combines learned priors with contextually relevant exemplars, often improving grounding and enabling rapid adaptation to new domains or figure styles without full retraining. However, performance is highly sensitive to retrieval quality, and the added retrieval and prompting pipeline increases system complexity and can reduce interpretability.

4.6 Personalization

Personalization adapts caption generation to an individual user by conditioning on user preferences, style, or interaction history (Singh et al., 2023; Ng et al., 2025c). Let U denote a user representation (e.g., an embedding from prior behavior or style cues), and generate

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

where I is the input figure. In practice, U can be implemented via user-specific fine-tuning, in-context exemplars reflecting the user's style, or latent conditioning inferred from prior conversations. Personalization is especially important for assistive systems and figure-grounded agents, and can also operate at the level of broader personas when per-user data are scarce. While it improves stylistic alignment and domain conventions, it depends on reliable user signals and may reduce robustness when examples are limited or narrowly distributed.

4.7 Hybrid Approaches

Retrieval-Augmented In-Context Learning. Methods such as LaMP-Cap (Ng et al., 2025b) and ChartThinker (Liu et al., 2024) retrieve figures, captions, or paragraphs and insert them as in-context exemplars, enabling adaptation to author-specific style or domain structure. These systems leverage non-parametric memory alongside in-context reasoning.

Distillation for Fine-Tuning. UniChart (Masry et al., 2023), ChartInstruct (Masry et al., 2024a), and ChartGemma (Masry et al., 2024b) first use a teacher model to generate synthetic captions or structured supervision, then fine-tune a student model on the enriched data. This couples teacher

supervision with parametric fine-tuning.

Generation with Intermediate Representations. Approaches such as DePlot (Liu et al., 2022), StructChart (Xia et al., 2023), ChartAssistant (Meng et al., 2024), and ChartVLM (Xia et al., 2025) produce an intermediate structured representation, such as tables, and then condition caption generation on it. This blends structured perception, intermediate reasoning, and generative modeling.

RLHF with Synthetic Supervision. Works like FigCaps-HF (Singh et al., 2025) apply RLHF on captions that are distilled or synthetically generated, combining supervised fine-tuning, teacher-based distillation, and reward-aligned optimization.

Retrieval-Enhanced Personalization. LaMP-Cap (Ng et al., 2025b) retrieves user-specific historical captions and feeds them into the prompt or encoder, blending retrieval with personalization and in-context learning.

In Table 1, we assign each method to the category corresponding to its dominant operational component, while this hybrid section clarifies overlaps for completeness.

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., 2025c; Omasa et al., 2025; Srinivasan et al., 2025) advocate human centric evaluations that emphasize **informativeness**, **factual correctness**, and **structure** (Masry et al., 2024a,b), along with 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., 2025c).

A practical concern is score reliability across annotators and evaluation setups. Only a few works report agreement statistics: SciCap-EVAL (Hsu et al., 2023) finds a Kendall correlation of 0.401 between GPT-4 zero-shot ratings and PhD students' rankings of helpfulness, with moderate agreement between two expert annotators on a subset. Recent multi-LLM collaborative captioning further reports low to moderate Fleiss' kappa (0.15–0.38) for best or worst caption selection, and the SciCap Challenge similarly observes modest Kendall values across judges (Kim et al., 2025), underscoring the difficulty of consensus.

In contrast, some chart captioning and visualization studies (Ashok, 2025) report high inter rater agreement under narrow, well specified coding schemes with trained annotators, suggesting stricter protocols can improve reliability at the cost of evaluation breadth. Overall, most figure captioning papers report at most small scale checks, and no common standard exists.

5.3 MLLM-as-a-Judge

To address the limits of reference based and human evaluations, recent work introduces MLLMs as 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 models on preference data for consistent judgments, while score based systems such as GEval (Liu et al., 2023b) 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 summarize key figure captioning datasets in Table 3, comparing scale, visualization types, and task focus. Existing resources span from synthetic captioning benchmarks, such as FigCAP (Chen et al., 2019) to large-scale real-world corpora, such as SciCAP (Hsu et al., 2021). Recent extensions add richer supervision and new settings. SciCAP+ (Yang et al., 2024) incorporates OCR tokens and mention paragraphs for knowledge-augmented generation, while LaMP-Cap (Ng et al., 2025b) introduces personalized captioning conditioned on user writing style. Furthermore, Chart-to-Text (Kantharaj et al., 2022) and ChartSumm (Rahman et al., 2023) target insight-oriented summarization, and reasoning benchmarks such as FigureQA (Kahou et al., 2017), DVQA (Kafle et al., 2018), and ChartX (Xia et al., 2025) evaluate com-

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., 2025c)	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 (Rahman et al., 2023)	84k images / 84k 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 3: Summary of key datasets used for training and evaluating MLLMs. Each dataset varies in image/text scale, covered chart types, and core tasks.

positional and higher-order understanding across diverse chart types. Detailed dataset descriptions are deferred to Appendix B.

7 Open Problems & Challenges

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

Complex Figure Types: Emerging figure types such as animated, interactive, and immersive visualizations introduce significant challenges for figure captioning. Animated figures require temporal reasoning over motion and evolving semantics, while interactive figures demand captions that adapt to user-driven context and exploration. Immersive 3D or virtual environment figures add occlusion and viewpoint variation that complicate spatial understanding. These settings motivate models that ensure temporal consistency, incorporate user context, and reason over 3D structure.

Personalized Caption: Caption personalization adapts generation to either the author's style or the target reader's perspective. The former reflects the author's linguistic and interpretive preferences, while the latter accounts for the reader's background, intent, and potential biases. Future work should model both dimensions via data design, fine-tuning, and prompt conditioning.

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.

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.

Attribution-Aware Grounding: One important open problem is attribution-aware and fine-grained grounding of captions to the insights, legend, and other parts of the figure. 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 fine-grained, attribution-aware evidence to justify insights.

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*.

Tejas Tumakuru Ashok. 2025. Llm-powered chart captioning and insight generation: Towards trustworthy visualization narratives. *Authorea Preprints*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report.

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.

Kung-Hsiang Huang, Hou Pong Chan, May Fung, Haoyi Qiu, Mingyang Zhou, Shafiq Joty, Shih-Fu Chang, and Heng Ji. 2024a. From pixels to insights: A survey on automatic chart understanding in the era of large foundation models. *IEEE Transactions on Knowledge and Data Engineering*, 37(5):2550–2568.

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

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.

Jaeyoung Kim, Jongho Lee, Hong-Jun Choi, Ting-Yao Hsu, Chieh-Yang Huang, Sungchul Kim, Ryan Rossi, Tong Yu, Clyde Lee Giles, Ting-Hao 'Kenneth' Huang, et al. 2025. Multi-llm collaborative caption generation in scientific documents. In *International Workshop on AI for Transportation*, pages 142–160. Springer.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2023. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR.

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*.

Li Li, Wei Ji, Yiming Wu, Mengze Li, You Qin, Lina Wei, and Roger Zimmermann. 2024. Panoptic scene graph generation with semantics-prototype learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 3145–3153.

Shawn Li, Huixian Gong, Hao Dong, Tiankai Yang, Zhengzhong Tu, and Yue Zhao. 2025a. Dpu: Dynamic prototype updating for multimodal out-of-distribution detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10193–10202.

Shawn Li, Jiashu Qu, Yuxiao Zhou, Yuehan Qin, Tiankai Yang, and Yue Zhao. 2025b. Treble counterfactual vlms: A causal approach to hallucination. *arXiv preprint arXiv:2503.06169*.

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*.

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

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. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, et al. 2025. Chartqapro: A more diverse and challenging benchmark for chart question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19123–19151.

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.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the ieee/cvf winter conference on applications of computer vision*, pages 1527–1536.

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, and Ting-Hao 'Kenneth' Huang. 2025b. Lamp-cap: Personalized figure caption generation with multimodal figure profiles.

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. 2025c. 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.

Raian Rahman, Rizvi Hasan, Abdullah Al Farhad, Md Tahmid Rahman Laskar, Md Hamjajul Ashmafee, and Abu Raihan Mostofa Kamal. 2023. Chartsumm: A comprehensive benchmark for automatic chart summarization of long and short summaries. In *Canadian AI*.

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*.

Ashish Singh, Ashutosh Singh, Prateek Agarwal, Zixuan Huang, Arpita Singh, Tong Yu, Sungchul Kim, Victor Bursztyn, Nesreen K. Ahmed, Puneet Mathur, Erik Learned-Miller, Franck Dernoncourt, and Ryan A. Rossi. 2025. Figcaps-hf: A figure-to-caption generative framework and benchmark with human feedback.

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.

Hao Tan, Chen-Tse Tsai, Yujie He, and Mohit Bansal. 2022. Scientific chart summarization: Datasets and improved text modeling. In *SDU@ AAAI*.

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., 2025c) 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., 2025c), **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; Li et al., 2025b). 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., 2023b) 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.

A.4 Best-Practice Evaluation

To make the evaluation section more actionable, we recommend a default protocol that we believe is appropriate for the Figure Captioning task. We suggest reporting a small set of reference-based metrics. For example, CIDEr together with ROUGE-L

and METEOR or BLEU, can be included only for comparison with older papers. We caution against using these scores in isolation because captions with the same meaning can differ substantially in wording. For this reason, we also recommend including one semantic metric, such as BERTScore, to better reflect paraphrastic equivalence. Since the dominant errors of large generative captioners are often errors of grounding rather than fluency, we further recommend adding a faithfulness component whenever the figure contains dense text or quantitative content by checking consistency between the generated caption and extracted visual text/entities, and for charts, explicitly verify statements about extrema, ordering, and trend direction.

Finally, we recommend complementing automatic scores with a rubric-based judge evaluation and validating it on a small human-rated subset with agreement statistics. The same protocol can be applied across settings, with the emphasis shifted as needed.

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 3.

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. (2025c) 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 pro-

posed 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. Automatically 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 systems by grounding language understanding in

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

Table 5: 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	✓					

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.

E Relation to General Image Captioning

Figure captioning shares a strong methodological foundation with general image captioning, as both tasks aim to translate visual information into natural language. Consequently, the majority of figure captioning systems leverage architectures originally developed for natural images, including the use of visual encoders and the adoption of encoder and decoder paradigms for sequence generation. Evaluation methodologies have also been directly adapted; metrics such as CIDEr and text image grounding techniques were pioneered in the general domain before being applied to chart evaluation. However, the direct application of general image captioning techniques often yields suboptimal results due to the unique semantic density of scientific figures (Hsu et al., 2021; Tan et al., 2022; Huang et al., 2024b). While general image captioning typically focuses on perceptual description, such as identifying objects, actions, and scenes, figure captioning requires logical reasoning to extract salient patterns and data driven insights. Generic models often fail to capture the specialized context of figures, such as correctly interpreting axis scales, legend mappings, or specific scientific nomenclature. Unlike natural scenes where approximations are often acceptable, scientific captioning demands high fidelity because a failure to recognize a specific data trend or axis label can lead to hallucinations that fundamentally misrepresent the underlying data. Furthermore, effective figure captioning demands a significantly higher degree of OCR and statistical integration to ground textual claims in specific data points, which is a requirement rarely present in natural scene description. Where a general caption might suffice with visual descriptors, a figure caption must precisely articulate relationships and quantitative shifts. Consequently, recent trends in the field have had to diverge from general image captioning by emphasizing attribution aware grounding and the curation of knowledge augmented datasets to bridge the gap between visual perception and scientific interpretation.

F Method Selection

When **numerical fidelity**, like exact values, rankings, trends, or comparisons, is the primary objective, we will highlight that **chart-to-structure** approaches are typically most suitable, since extracting an explicit representation provides a verifiable substrate for generation and reduces hallucinated quantities, such as DePlot (Liu et al., 2022), StructChart (Xia et al., 2023), ChartAssistant (Meng et al., 2024), ChartVLM (Xia et al., 2025).

When the figure exhibits **strong compositional structure, such as flowcharts or diagrams with arrows**, we will note that structure-guided modeling and targeted reasoning augmentations are often better aligned with the underlying semantics than purely pixel-to-text generation, as in FlowLearn (Pan et al., 2024).

When captions must **align with document context**. For example, scientific figures whose meaning depends on surrounding claims, experimental setup, or terminology. We will emphasize that **auxiliary-text** and **retrieval/hybrid** approaches are more appropriate because crucial intent is often not recoverable from the figure alone. Like in SciCap+ (Yang et al., 2024), ChartThinker (Liu et al., 2024); LaMP-Cap (Ng et al., 2025b) for profile/context conditioning.

When the task is iterative authoring or accessibility, conversational assistants are a natural fit, enabling verification and incremental refinement rather than one-shot captioning, such as Pluto (Srinivasan et al., 2025) and FigurA11y (Singh et al., 2024).

G Practical Trade-offs Across Techniques

G.1 Techniques

Across the surveyed literature, differences between figure captioning systems are often best understood through the degree to which generation is constrained by evidence in the figure and, when relevant, by external context. This perspective helps explain recurring trade-offs among prompting-only methods, fine-tuning, structure-assisted pipelines, and retrieval- or tool-augmented approaches, particularly with respect to factual faithfulness and hallucination.

Prompting-only approaches, including in-context learning, remain attractive due to their low implementation overhead and their ability to adapt quickly across domains and figure styles without additional training. Their limitations are also consistent. When the prompt does not explicitly enforce grounding, generated captions can reflect plausible priors rather than visually supported content, leading to unsupported entities, incorrect mappings between visual encodings and series, or imprecise numerical statements. More structured prompts that encourage a staged "extraction then summary" behavior can improve reliability, but reported outcomes are often sensitive to prompt formatting, context length, and the underlying base model, and may not transfer uniformly across datasets or chart types.

Fine-tuning tends to improve stability in faithfulness when supervision matches the target distri-

Method	Technique	Params	ChartQA RA
Specialized chart understanding models			
Pix2Struct-Base (Lee et al., 2023)	Pretrain + FT	282M	56.0
Pix2Struct-Large (Lee et al., 2023)	Pretrain + FT	1.3B	58.6
MatCha (Liu et al., 2023a)	Pretrain + FT	282M	64.2
UniChart (Masry et al., 2023)	Distillation + FT	201M	66.24
ChartInstruct-LLaMA2 (Masry et al., 2024a)	Instruction FT	7B	66.64
ChartLlama (Han et al., 2023)	LLM FT on synthetic data	13B	69.66
DePlot + LLM (Liu et al., 2022)	Intermediate rep. + ICL	–	79.3
ChartAssistant (Meng et al., 2024)	Chart-to-table + multitask FT	13B	79.90
ChartGemma (Masry et al., 2024b)	Visual instruction tuning	3B	80.16
ChartMoE (Xu et al., 2025)	MoE experts	8B	**84.64**
General-purpose frontier models			
GPT-4	ICL, closed-source	–	85.7
Gemini 1.5	ICL, closed-source	–	87.2
Qwen2.5-VL (Bai et al., 2025)	ICL, open-source	72B	89.5
Claude Sonnet 3.5	ICL, closed-source	–	**90.5**

Table 6: ChartQA Relaxed Accuracy (RA@0.05, averaged over the augmented and human splits) for representative chart understanding methods and recent frontier models. Specialized open-source rows report selected configurations from the cited papers or later comparison tables. Pix2Struct-Base, MatCha, UniChart, ChartLlama, ChartInstruct, ChartAssistant, and ChartGemma follow the comparison protocols reported in Masry et al. (2023) and Masry et al. (2024b), while Pix2Struct-Large reports the best result from Lee et al. (2023). ChartMoE reports Acc.@0.05 in its own paper. The frontier model results are taken from Masry et al. (2025). The DePlot + LLM row uses FlanPaLM and Codex with self-consistency over CoT and PoT prompting.

bution and evaluation criteria. In settings where the dominant figure types are well represented, fine-tuned models more consistently follow domain conventions for describing trends, comparisons, and salient elements, and they are generally less prone to introducing unsupported claims. This comes at the cost of additional data curation and training infrastructure, as well as reduced flexibility when figure distributions shift. The benefits can also be uneven for long-tail figure types, where limited coverage in training data restricts generalization.

Structure-assisted approaches occupy a complementary point in this design space by constraining generation through intermediate representations. Methods that extract tables, series, or relational structure before caption generation often yield improved numeric and relational faithfulness because the caption is conditioned on an explicit representation rather than only on raw pixels. These approaches are particularly aligned with datasets that reward precise values, ranks, and comparisons. At the same time, they introduce distinct failure modes, since extraction errors can propagate into the generated caption, and they generally require additional engineering. The advantage is that the intermediate representation can make errors more interpretable and debugging more tractable than purely end-to-end generation.

Retrieval-augmented and tool-augmented approaches are most beneficial when captions must reflect information beyond the figure itself, such as surrounding scientific narrative, definitions, or experimental context. When retrieval is accurate and the generation is explicitly anchored to retrieved evidence, these methods can improve both relevance and factuality. However, they can be sensitive to retrieval noise and to mismatches between retrieved text and the figure content, which may introduce inconsistencies if not carefully controlled. Taken together, the surveyed findings suggest a consistent trend. As datasets place greater emphasis on precise grounding, such as numeric fidelity, correct visual-to-text alignment, and avoidance of unsupported inferences, stronger constraints through fine-tuning and structured intermediates become increasingly advantageous over prompt-only solutions. Prompting-only methods remain compelling for broad coverage and rapid adaptation, particularly when captions are evaluated primarily for fluent high-level description rather than strict factual precision.

G.2 Empirical Comparisons

The discussion above stays qualitative. Numbers in the literature tell a slightly more pointed story, and pulling them onto a single axis makes a few trade-offs concrete. Table 6 collects representative ChartQA Relaxed Accuracy (RA) scores across the surveyed methods, with the closed- and open-source frontier shown for context. Three patterns

recur.

Distillation can beat scale. UniChart (Masry et al., 2023), at 201M parameters and distilled from LLM-generated chart summaries, beats MatCha (Liu et al., 2023a) on both ChartQA and on Chart-to-Text Pew, and runs more than 11× faster. ChartGemma (Masry et al., 2024b) does the same thing one tier up, holding state-of-the-art on five chart benchmarks while staying substantially smaller than the 13B ChartLlama (Masry et al., 2024b). Across the 1B–13B range, well-curated chart-grounded instruction data and a strong teacher have mattered more for captioning quality than raw parameter count. That is not what one would predict from general-purpose MLLM benchmarks, and it argues against treating scale as the primary axis of progress in this subfield.

High BLEU is compatible with non-factual captions. Huang et al. (2024b) find that 82.06% of captions produced by state-of-the-art models, GPT-4V included, contain at least one factual error on the CHOCOLATE benchmark, even though several of those models post competitive BLEU and ROUGE scores. Lim et al. (2025) push in the other direction. A Phi-3.5-Vision-4B model fine-tuned on ChartCap's dense, type-specific captions is preferred by human raters over Claude 3.5 Sonnet on informativeness, accuracy, and hallucination reduction, despite being roughly an order of magnitude smaller. The two findings reinforce a point Appendix A.1 already makes about n-gram overlap, and they suggest that closing the faithfulness gap has so far come from carefully constructed training data rather than from scaling.

Per-technique deltas over a fine-tuning baseline. A few papers report direct head-to-head numbers that give a rough sense of which technique class moves the needle most.

- **RLHF over SFT.** FigCaps-HF (Singh et al., 2023) report mean gains of +35.7% ROUGE, +16.9% BLEU, and +9% METEOR over a BLIP fine-tune when RLHF is applied on top. This is the largest single-technique boost in the surveyed work for figure-only inputs.
- **Auxiliary text and profile conditioning.** LaMP-Cap (Ng et al., 2025c) lifts GPT-4o from BLEU-4 0.063 (no profile) to 0.115 (full profile) on personalized scientific captioning, an ≈82% relative jump, by adding a multimodal figure profile drawn from the same paper. This is the largest input-side effect reported anywhere in the survey.
- **Intermediate structured representations.** DePlot+LLM (Liu et al., 2022) in a one-shot setting matches or beats fine-tuned MatCha on human-sourced ChartQA queries. Explicit table extraction can substitute for tens of thousands of fine-tuning examples when the task is numerically demanding, though extraction errors propagate to the caption when they occur.

A rough practical ordering follows. If auxiliary text or an author profile is available at inference time, conditioning on it usually pays the most for the least training cost. For numerically demanding tasks, a structured intermediate representation is the next most reliable lever. RLHF on top of an existing fine-tune adds a further but typically smaller gain, and is most useful when reasonably calibrated preference data is on hand.

Comparability caveats. We aggregated each number from the configuration the original paper reports as its best, and these are not always directly comparable. Chart-to-Text preprocessing pipelines differ slightly across implementations, ChartQA RA tolerances vary in the 5–10% range, and a few methods report only on subsets we have approximated by their dominant figure type. The table is best read as a trend summary rather than a leaderboard. A clean re-evaluation under a single protocol, ideally one that includes a faithfulness component in the spirit of Huang et al. (2024b) and Lim et al. (2025), would be a useful next step for the community.

H Complete Paper List for Table 2

Table 2 reports approximate publication volumes and dominant modeling trends by period. The counts were estimated using a consistent set of survey-relevant Google Scholar queries, including "figure captioning," "chart captioning," "scientific figure description," "diagram captioning," "plot captioning," and combinations of these terms with representative model architectures. This appendix provides the complete list of papers used to support those counts. Papers are grouped by the Table 2 year ranges and ordered chronologically.

H.1 pre-2018

[1] Corio et al., Integrated Generation of Graphics and Text: A Corpus Study, 1998.

[2] Mittal et al., Describing Complex Charts in Natural Language: A Caption Generation System, 1998.

[3] Corio et al., Generation of Texts for Information Graphics, 1999.

[4] Ramloll et al., Constructing Sonified Haptic Line Graphs for the Blind Student: First Steps, 2000.

[5] Brewster et al., Visualization Tools for Blind People Using Multiple Modalities, 2002.

[6] Brown et al., Browsing Modes for Exploring Sonified Line Graphs, 2002.

[7] Brown et al., Drawing by Ear: Interpreting Sonified Line Graphs, 2003.

[8] Elzer et al., A Probabilistic Framework for Recognizing Intention in Information Graphics, 2005.

[9] McGookin et al., SoundBar: Exploiting Multiple Views in Multimodal Graph Browsing, 2006.

[10] Demir et al., Generating Textual Summaries of Bar Charts, 2008.

[11] Elzer et al., Accessible Bar Charts for Visually Impaired Users, 2008.

[12] Demir et al., A Discourse-Aware Graphical Presentation System for Information Graphics, 2010.

[13] Elzer et al., The Automated Understanding of Simple Bar Charts, 2011.

[14] Demir et al., Summarizing Information Graphics Textually, 2012.

[15] Al-Zaidy et al., Automatic Summary Generation for Scientific Data Charts, 2016.

[16] Clark et al., Semantic Scholar: A Search Engine for Scientific Literature, 2016.

[17] Siegel et al., FigureSeer: Parsing Result-Figures in Research Papers, 2016.

[18] Poco et al., Reverse-Engineering Visualizations: Recovering Visual Encodings from Chart Images, 2017.

[19] Cliche et al., Scatteract: Automated Extraction of Data from Scatter Plots, 2017.

[20] Johnson et al., CLEVR Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning (widely used in 2019), 2017.

[21] Thiyam et al., Evaluation of Convolutional Neural Network Architectures for Chart Image Classification, 2018.

[22] Balaji et al., Chart-Text: A Fully Automated Chart Image Descriptor, 2018.

[23] Poco et al., Extracting and Retargeting Color Mappings from Bitmap Images of Visualizations (IEEE TVCG 2020), 2018.

H.2 2019

[24] Davila et al., ICDAR 2019 Competition on Harvesting Raw Tables from Infographics (CHART-Infographics), 2019.

[25] Chagas et al., Chart Classification Using Simplified VGG Model (IWSSIP 2019), 2019.

[26] Puduppully et al., Data-to-text Generation with Content Selection and Planning (AAAI 2019), 2019.

[27] Puduppully et al., Data-to-text Generation with Entity Modeling (ACL 2019), 2019.

[28] Clark et al., Looking Beyond Text: Extracting Figures, Tables and Captions from Computer Science Papers, 2019.

[29] Hudson et al., GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering (CVPR 2019), 2019.

[30] Srinivasan et al., Collecting and Characterizing Natural Language Utterances for Specifying Data Visualizations (CHI 2019), 2019.

[31] Setlur et al., Inferencing Underspecified Natural Language Utterances in Visual Analysis (IUI 2019), 2019.

[32] Zhong et al., PubLayNet: Largest Dataset Ever for Document Layout Analysis (ICDAR 2019), 2019.

[33] Choi et al., Visualizing for the Non-Visual: Enabling the Visually Impaired to Use Visualization (EuroVis 2020), 2019.

[34] Chen et al., Figure Captioning with Reasoning and Sequence-Level Training, 2019.

H.3 2020

[35] Chaudhry et al., LEAF-QA: Locate, Encode \& Attend for Figure Question Answering, 2020.

[36] Kim et al., Answering Questions about Charts and Generating Visual Explanations, 2020.

[37] Li et al., DocBank: A Benchmark Dataset for Document Layout Analysis, 2020.

[38] Methani et al., PlotQA: Reasoning over Scientific Plots (WACV 2020), 2020.

[39] Singh et al., STL-CQA: Structure-based Transformers with Localization and Encoding for Chart Question Answering (EMNLP 2020), 2020.

[40] Kafle et al., Answering Questions about Data Visualizations using Efficient Bimodal Fusion (WACV 2020), 2020.

[41] Obeid et al., Chart-to-Text: Generating Natural Language Descriptions for Charts by Adapting the Transformer Model (INLG 2020), 2020.

[42] Spreafico et al., Neural Captioning for Charts (ASE Workshop 2020), 2020.

[43] Rebuffel et al., A Hierarchical Model for Data-to-Text Generation (ECIR 2020), 2020.

[44] Luo et al., Synthesizing Natural Language to Visualization (NL2VIS) Benchmarks from NL2SQL Benchmarks, 2020.

[45] Chen et al., Figure Captioning with Relation Maps for Reasoning, 2020.

[46] Xu et al., LayoutLM: Pre-training of Text and Layout for Document Image Understanding (KDD 2020), 2020.

[47] Zhong et al., Image-based Table Recognition: Data, Model, and Evaluation (NeurIPS 2020 D\&B), 2020.

[48] Chen et al., UNITER: UNiversal Image-TExt Representation Learning (ECCV 2020), 2020.

[49] Lu et al., 12-in-1: Multi-Task Vision and Language Representation Learning (CVPR 2020), 2020.

[50] Kim et al., Accessible Visualization: Design Space, Opportunities, and Challenges, 2020.

H.4 2021

[51] Gehrmann et al., The GEM Benchmark: Natural Language Generation, its Evaluation and Metrics (GEM Workshop 2020), 2021.

[52] Puduppully et al., Data-to-text Generation with Macro Planning (TACL 2020, submitted), 2021.

[53] Levy et al., Challenges in Chart Image Classification (DocEng 2021), 2021.

[54] Luo et al., ChartOCR: Data Extraction from Charts Images via a Deep Hybrid Framework (WACV 2021), 2021.

[55] Ma et al., Towards Robust and Scalable Chart Data Extraction from Complex Documents (WWW 2021 Companion), 2021.

[56] Masry et al., Integrating Image and Data Table for Chart Question Answering and Summarization, 2021.

[57] Hsu et al., SciCap: Generating Captions for Scientific Figures (EMNLP 2021 Findings), 2021.

[58] Suadaa et al., Towards Table-to-Text Generation with Numerical Reasoning (ACL 2021), 2021.

[59] Chen et al., Finqa: A Dataset of Numerical Reasoning over Financial Data (EMNLP 2021), 2021.

[60] Nan et al., FeTaQA: Free-form Table Question Answering (TACL 2021), 2021.

[61] Nan et al., DART: Open-Domain Structured Data Record to Text Generation (NAACL 2021), 2021.

[62] Luo et al., Synthesizing Natural Language to Visualization (NL2VIS) Benchmarks from NL2SQL, 2021.

[63] Wu et al., AI4VIS: Survey on Artificial Intelligence Approaches for Data Visualization (IEEE TVCG 2021), 2021.

[64] Lundgard et al., Accessible Visualization via Natural Language Descriptions: A Four-Level Model of Semantic Content (IEEE TVCG 2021), 2021.

[65] Xu et al., LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding (ACL 2021), 2021.

[66] Xu et al., LayoutXLM: Multimodal Pre-training for Multilingual Visually-rich Document Understanding, 2021.

[67] Appalaraju et al., DocFormer: End-to-End Transformer for Document Understanding (ICCV 2021), 2021.

[68] Tang et al., StructuralLM: Structural Pre-training for Form Understanding (ACL 2021), 2021.

[69] Liu et al., Neural Collaborative Reasoning for Table Structure Recognition (WWW 2021), 2021.

[70] Cho et al., Unifying Vision-and-Language Tasks via Text Generation (ICML 2021), 2021.

[71] Li et al., ALIGN: Scaling Up Visual and Vision-Language Representation Learning with Noisy Text Supervision (ICML 2021), 2021.

[72] Tian et al., ICDAR 2021 Competition on Document Visual Question Answering (ICDAR 2021), 2021.

[73] Shen et al., Towards Natural Language-Based Visualization Authoring (IEEE TVCG 2021), 2021.

[74] Talmor et al., MultiModalQA: Complex Question Answering over Text, Tables and Images (ICLR 2021), 2021.

H.5 2022

[75] Jung et al., Communicating Visualizations without Visuals: Investigation of Visualization Alternative Text for People with Visual Impairments, 2022.

[76] Kim et al., Donut: Document Understanding Transformer (OCR-Free, submitted Nov 2021), 2022.

[77] Smock et al., PubTables-1M: Towards Comprehensive Table Extraction from Unstructured Documents, 2022.

[78] Nassar et al., TableFormer: Table Structure Understanding with Transformers, 2022.

[79] Wang et al., VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts, 2022.

[80] Masry et al., ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning (ACL 2022 Findings), 2022.

[81] Kantharaj et al., OpenCQA: Open-ended Question Answering with Charts (EMNLP 2022), 2022.

[82] Kantharaj et al., Chart-to-Text: A Large-Scale Benchmark for Chart Summarization (ACL 2022), 2022.

[83] Obeid et al., Chart-to-Text: Generating Natural Language Descriptions for Charts (Applied Intelligence 2022), 2022.

[84] Davila et al., ICPR 2022 Competition on Chart Understanding (CHART-Info), 2022.

[85] Kim et al., Donut: Document Understanding Transformer without OCR (ECCV 2022), 2022.

[86] Huang et al., LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking (ACM MM 2022), 2022.

[87] Li et al., DiT: Self-supervised Pre-training for Document Image Transformer (ACM MM 2022), 2022.

[88] Tang et al., Unifying Vision, Text, and Layout for Universal Document Processing (UDOP, submitted 2022), 2022.

[89] Davis et al., End-to-end Document Recognition and Understanding with Dessurt (ECCV 2022 Workshop), 2022.

[90] Li et al., TSRFormer: Table Structure Recognition with Transformers (ACM MM 2022), 2022.

[91] Zhao et al., Towards Table-to-Text Generation with Pretrained Language Model: A Table Structure Understanding Approach (EMNLP 2022), 2022.

[92] Kasner et al., Neural Pipeline for Zero-Shot Data-to-Text Generation (ACL 2022), 2022.

[93] Gehrmann et al., GEMv2: Multilingual NLG Benchmarking in a Single Line of Code (EMNLP 2022), 2022.

[94] Luo et al., Natural Language to Visualization by Neural Machine Translation (IEEE TVCG 2022), 2022.

[95] Tang et al., ChartSeer: Interactive Steering Exploratory Visual Analysis with Machine Intelligence (IEEE TVCG 2022), 2022.

[96] Elavsky et al., How Accessible is My Visualization? Evaluating Visualization Accessibility with Chartability (EuroVis 2022), 2022.

[97] Li et al., BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation (ICML 2022), 2022.

[98] Wang et al., Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks (BEiT-3, submitted 2022), 2022.

[99] Yu et al., CoCa: Contrastive Captioners are Image-Text Foundation Models (ICML 2022, submitted), 2022.

[100] Chen et al., PaLI: A Jointly-Scaled Multilingual Language-Image Model, 2022.

[101] Alayrac et al., Flamingo: A Visual Language Model for Few-Shot Learning (NeurIPS 2022), 2022.

[102] Zheng et al., MultiHiertt: Numerical Reasoning over Multi Hierarchical Tabular and Textual Data (ACL 2022), 2022.

[103] Chen et al., HiTab: A Hierarchical Table Dataset for Question Answering and Natural Language Generation (ACL 2022), 2022.

[104] Mathew et al., InfographicVQA (WACV 2022), 2022.

[105] Kafle et al., Chart Understanding and Reasoning (Chapter in Computer Vision for Industrial Applications), 2022.

[106] Radford et al., Robust Speech Recognition via Large-Scale Weak Supervision — Whisper (applied to chart audio, 2022), 2022.

[107] Ouyang et al., Training Language Models to Follow Instructions with Human Feedback — InstructGPT (NeurIPS 2022), 2022.

[108] Choi et al., Retrieving and Reading Multimodal Documents for Open-domain Question Answering (ACL 2022 Findings), 2022.

[109] Xie et al., UnifiedSKG: Unifying and Multi-Tasking Structured Knowledge Grounding with Text-to-Text Language Models (EMNLP 2022), 2022.

[110] Zheng et al., DocLayNet: A Large Human-Annotated Dataset for Document-Layout Analysis (KDD 2022), 2022.

[111] Li et al., GraPE: A Retrieve-Then-Reason Framework for Answering Complex Questions over Document Tables, 2022.

[112] Liu et al., DePlot: One-shot Visual Language Reasoning by Plot-to-Table Translation (ACL 2023 Findings), 2022.

[113] Davila et al., ICPR 2022 Competition on Understanding Chart Images (CHART-Info, published 2023), 2022.

[114] Kim et al., OCR-Free Document Understanding Transformer — Donut (Extended journal version, IEEE TPAMI 2023), 2022.

H.6 2023

[115] Yan et al., Context-Aware Chart Element Detection (ICDAR 2021), 2023.

[116] Ye et al., DPText-DETR: Towards Better Scene Text Detection with Dynamic Points in Transformer (AAAI 2023, submitted 2022), 2023.

[117] Masry et al., UniChart: A Universal Vision-language Pretrained Model for Chart Comprehension and Reasoning (EMNLP 2023), 2023.

[118] Xu et al., ChartBench: A Benchmark for Complex Visual Reasoning in Charts, 2023.

[119] Li et al., SciGraphQA: A Large-Scale Synthetic Multi-Turn Question-Answering Dataset for Scientific Graphs, 2023.

[120] Zhou et al., CHARTREADER: A Unified Framework for Chart Derendering and Comprehension without Heuristic Rules, 2023.

[121] Liu et al., MatCha: Enhancing Visual Language Pretraining with Math Reasoning and Chart Derendering (ACL 2023), 2023.

[122] Lee et al., Pix2Struct: Screenshot Parsing as Pretraining for Visual Language Understanding (ICML 2023), 2023.

[123] Han et al., ChartLlama: A Multimodal LLM for Chart Understanding and Generation, 2023.

[124] Zhou et al., ChartT5: Enhanced Chart Understanding via Visual Language Pre-training on Plot Table Pairs (ACL 2023 Findings), 2023.

[125] Tang et al., VisText: A Benchmark for Semantically Rich Chart Captioning (ACL 2023, Outstanding Paper Award), 2023.

[126] Rahman et al., ChartSumm: A Comprehensive Benchmark for Automatic Chart Summarization of Long and Short Summaries (Canadian AI 2023), 2023.

[127] Hsu et al., SciCap+: A Knowledge Augmented Dataset for Scientific Figure Captioning, 2023.

[128] Chen et al., Generating Figure Captions for Scientific Documents with Multimodal Models (INLG 2023), 2023.

[129] Prasad et al., LineEX: Data Extraction from Scientific Line Charts (WACV 2023), 2023.

[130] Liu et al., LineFormer: Line Chart Data Extraction using Instance Segmentation, 2023.

[131] Liu et al., LLaVA: Visual Instruction Tuning (NeurIPS 2023), 2023.

[132] Liu et al., LLaVA-1.5: Improved Baselines with Visual Instruction Tuning, 2023.

[133] Dai et al., InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning (NeurIPS 2023), 2023.

[134] Li et al., BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and LLMs (ICML 2023), 2023.

[135] Zhu et al., MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models, 2023.

[136] Ye et al., mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality, 2023.

[137] Bai et al., Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond, 2023.

[138] Chen et al., InternVL: Scaling Up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks, 2023.

[139] OpenAI, GPT-4V(ision) System Card, 2023.

[140] Blecher et al., Nougat: Neural Optical Understanding for Academic Documents, 2023.

[141] Tang et al., Unifying Vision, Text, and Layout for Universal Document Processing — UDOP (CVPR 2023), 2023.

[142] Huang et al., Improving Table Structure Recognition With Visual-Alignment Sequential Coordinate Modeling — VAST (CVPR 2023), 2023.

[143] Ye et al., mPLUG-DocOwl: Modularized Multimodal Large Language Model for Document Understanding, 2023.

[144] Zhang et al., TabPedia: Towards Comprehensive Visual Table Understanding with Concept Synergy, 2023.

[145] Zhao et al., LOFT: Finding Literal Connections for Faithful Table-to-Text Generation (ACL 2023 Findings), 2023.

[146] Kasner et al., Tabgenie: A Toolkit for Table-to-Text Generation (ACL 2023 System Demos), 2023.

[147] Zhao et al., Large Language Models are Effective Table-to-Text Generators, Evaluators, and Feedback Providers, 2023.

[148] Luo et al., NL2VIS Benchmarks: Synthesizing Natural Language to Visualization (IEEE TVCG 2023), 2023.

[149] Kim et al., Chart Reader: Accessible Visualization Experiences Designed with Screen Reader Users, 2023.

[150] Dhamdhere et al., A Survey and Approach to Chart Classification, 2023.

[151] Bajic et al., Chart Question Answering: State of the Art and Future Directions (CGF Survey), 2023.

[152] Fu et al., MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models, 2023.

[153] Liu et al., MMBench: Is Your Multi-modal Model an All-around Player?, 2023.

[154] Li et al., SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension, 2023.

[155] Yu et al., MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities, 2023.

[156] Singh et al., DocILE: Document Information Localization and Extraction Benchmark (ICDAR 2023), 2023.

[157] Touvron et al., LLaMA: Open and Efficient Foundation Language Models, 2023.

[158] Touvron et al., LLaMA-2: Open Foundation and Fine-Tuned Chat Models, 2023.

[159] Chiang et al., Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90\% ChatGPT Quality, 2023.

[160] Driess et al., PaLM-E: An Embodied Multimodal Language Model (ICML 2023), 2023.

[161] Xue et al., ChartSpark: LLM-Driven Interactive Visualization with Contextual Narratives, 2023.

[162] Dibia et al., LIDA: A Tool for Automatic Generation of Grammar-Agnostic Visualizations and Infographics using Large Language Models, 2023.

[163] Luo et al., ChartDETR: A Multi-shape Detection Network for Visual Chart Understanding (extended 2023), 2023.

[164] Wang et al., CogVLM: Visual Expert for Pretrained Language Models, 2023.

[165] Lin et al., SPHINX: The Joint Mixing of Weights, Tasks, and Visual Embeddings for Multi-modal Large Language Models, 2023.

[166] Li et al., Monkey: Image Resolution and Text Label Are Important Things for Large Multimodal Models, 2023.

[167] Dong et al., InternLM-XComposer: A Vision-Language Large Model for Advanced Text-image Comprehension and Composition, 2023.

[168] Gehrmann et al., Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text, 2023.

[169] Cheng et al., Binding Language Models in Symbolic Languages (ICLR 2023, Binder), 2023.

[170] Lv et al., DocPedia: Unleashing the Power of Large Multimodal Model in the Frequency Domain for Versatile Document Understanding, 2023.

[171] Zhao et al., ReasTAP: Injecting Table Reasoning Skills During Pre-training via Synthetic Reasoning Examples (EMNLP 2023), 2023.

[172] Chen et al., Large Language Models for Automated Data Science: Introducing CAAFE for Context-Aware Automated Feature Engineering (NeurIPS 2023), 2023.

[173] Li et al., Evaluating Object Hallucination in Large Vision-Language Models (EMNLP 2023), 2023.

[174] Tannert et al., FlowchartQA: The First Large-Scale Benchmark for Reasoning over Flowcharts, 2023.

H.7 2024

[175] Masry et al., ChartSumm: A Comprehensive Benchmark for Automatic Chart Summarization (submitted 2022), 2024.

[176] Meng et al., ChartAssisstant: A Universal Chart Multimodal Language Model via Chart-to-Table Pre-training and Multitask Instruction Tuning, 2024.

[177] Yang et al., Enhancing Scientific Figure Captioning Through Cross-modal Learning, 2024.

[178] Li et al., Text2Analysis: A Benchmark of Table Question Answering with Advanced Data Analysis and Unclear Queries, 2024.

[179] Wei et al., TinyChart: Efficient Chart Understanding with Visual Token Merging and Program-of-Thoughts Learning, 2024.

[180] Wei et al., From Pixels to Insights: A Survey on Automatic Chart Understanding with Foundation Models, 2024.

[181] Chao et al., The Solution for the ICCV 2023 1st Scientific Figure Captioning Challenge, 2024.

[182] Hu et al., mPLUG-PaperOwl: Scientific Diagram Analysis with the Multimodal Large Language Model, 2024.

[183] Liu et al., Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning (ICLR 2024, submitted 2023), 2024.

[184] Islam et al., Are Large Vision Language Models up to the Challenge of Chart Comprehension and Reasoning? (EMNLP 2024 Findings), 2024.

[185] Razeghi et al., Plot Twist: Multimodal Models Don't Comprehend Simple Chart Details (EMNLP 2024 Findings), 2024.

[186] Mukhopadhyay et al., Unraveling the Truth: Do LLMs Really Understand Charts? A Deep Dive into Consistency and Robustness (EMNLP 2024 Findings), 2024.

[187] Wu et al., ChartInsights: Evaluating Multimodal Large Language Models for Low-Level Chart Question Answering (EMNLP 2024 Findings), 2024.

[188] Wang et al., CharXiv: Charting Gaps in Realistic Chart Understanding in Multimodal LLMs (NeurIPS 2024), 2024.

[189] Xia et al., ChartX \& ChartVLM: A Versatile Benchmark and Foundation Model for Complicated Chart Reasoning, 2024.

[190] Zhang et al., Is GPT-4V(ision) All You Need for Automating Academic Data Visualization? AcademiaChart (EMNLP 2024 Findings), 2024.

[191] Shen et al., SCI-CQA: Rethinking Comprehensive Benchmark for Chart Understanding from Scientific Literature, 2024.

[192] Davila et al., CHART-Info 2024: A Dataset for Chart Analysis and Recognition (ICDAR 2024), 2024.

[193] Zhang et al., TinyChart: Efficient Chart Understanding with Program-of-Thoughts Learning and Visual Token Merging (EMNLP 2024), 2024.

[194] Meng et al., ChartAssistant: A Universal Chart Multimodal Language Model via Chart-to-Table Pre-training (ACL 2024 Findings), 2024.

[195] Masry et al., ChartInstruct: Instruction Tuning for Chart Comprehension and Reasoning (ACL 2024 Findings), 2024.

[196] Liu et al., MMC: Advancing Multimodal Chart Understanding with Large-scale Instruction Tuning (NAACL 2024), 2024.

[197] Xu et al., OneChart: Purify the Chart-Structural Extraction via One Auxiliary Token Sequence, 2024.

[198] Fan et al., On Pre-training of Multimodal Language Models Customized for Chart Understanding — CHOPINLLM (NeurIPS 2024), 2024.

[199] Xia et al., StructChart: Perception, Structuring, Reasoning for Visual Chart Understanding (CVPR 2024), 2024.

[200] Liu et al., ChartThinker: A Contextual Chain-of-Thought Approach to Optimized Chart Summarization (LREC-COLING 2024), 2024.

[201] Rahman et al., A Comparative Study of Language Models for Chart Summarization, 2024.

[202] Mustafa et al., ChartEye: A Deep Learning Framework for Chart Information Extraction, 2024.

[203] Zhao et al., Advancing Multimodal Large Language Models in Chart Question Answering with Visualization-Referenced Instruction Tuning (IEEE VIS 2024), 2024.

[204] Li et al., Synthesize Step-by-Step: Tools, Templates and LLMs as Data Generators for Reasoning about Charts (CVPR 2024), 2024.

[205] Yang et al., ChartMimic: Evaluating LMM's Cross-Modal Generation from Charts, 2024.

[206] Jang et al., Let the Chart Spark: Embedding Semantic Context into Chart with Text-to-Image Generative Model (EMNLP 2024), 2024.

[207] Chen et al., Interactive Data Visualization Generation from Natural Language (KDD 2024), 2024.

[208] Akhtar et al., ChartCheck: An Evidence-based Fact-Checking Dataset over Real-World Chart Images (ACL 2024 Findings), 2024.

[209] Huang et al., Do LVLMs Understand Charts? Analyzing and Correcting Factual Errors in Chart Captioning (ACL 2024 Findings), 2024.

[210] Cao et al., Figuring out Figures: Using Textual References to Caption Scientific Figures, 2024.

[211] Huang et al., SciCapenter: Supporting Caption Composition for Scientific Figures with Machine-Generated Captions and Metadata, 2024.

[212] Hsu et al., FigurA11y: AI Assistance for Making Scientific Figures Accessible, 2024.

[213] Chen et al., On the Effects of Data Scale on UI and Chart Understanding, 2024.

[214] Pan et al., FlowLearn: Evaluating Large Vision-Language Models on Flowchart Understanding (ECAI 2024), 2024.

[215] Laurençon et al., What Matters When Building Vision-Language Models? Idefics2 (NeurIPS 2024), 2024.

[216] Yue et al., MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark (CVPR 2024), 2024.

[217] Li et al., Multimodal ArXiv: A Dataset for Improving Scientific Comprehension of Large Vision-Language Models (ACL 2024), 2024.

[218] Zhang et al., Multimodal Self-Instruct: Synthetic Abstract Image and Visual Reasoning Instruction Using LLM, 2024.

[219] Ye et al., mPLUG-DocOwl2: High-resolution Compressing for OCR-free Multi-page Document Understanding, 2024.

[220] Zhang et al., UniTabNet: Bridging Vision and Language Models for Enhanced Table Structure Recognition (EMNLP 2024 Findings), 2024.

[221] He et al., OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Problems (ACL 2024), 2024.

[222] Wang et al., Mementos: A Comprehensive Benchmark for Multimodal LLM Reasoning over Image Sequences (ACL 2024), 2024.

[223] Zheng et al., Advancing Chart Question Answering with Robust Chart Component Recognition (WACV 2025, published late 2024), 2024.

[224] Liu et al., START: Spatial and Textual Learning for Chart Understanding, 2024.

[225] Zhao et al., Difficult Task Yes but Simple Task No: Unveiling the Laziness in Multimodal LLMs (EMNLP 2024 Findings), 2024.

[226] Koh et al., VisualWebArena: Evaluating Multimodal Agents on Realistic Visual Web Tasks (ACL 2024), 2024.

[227] Xue et al., ChartSpark: An LLM-Driven Interactive Visualization with Contextual Narratives for Data Analysis (IEEE TVCG 2024), 2024.

[228] Chen et al., Unmasking Deceptive Visuals: Benchmarking MLLMs on Misleading Chart Question Answering, 2024.

[229] Xu et al., ChartAdapter: Large Vision-Language Model for Chart Summarization, 2024.

[230] Moured et al., AltChart: Enhancing VLM-based Chart Summarization Through Multi-Pretext Tasks, 2024.

[231] Cheng et al., ChartReformer: Natural Language-Driven Chart Image Editing (AAAI 2025), 2024.

[232] Zhang et al., Multimodal Self-Instruct: Synthetic Abstract Image and Visual Reasoning Instruction (EMNLP 2024), 2024.

[233] Li et al., Synthesize Step-by-Step: Tools, Templates and LLMs as Data Generators for Chart VQA (CVPR 2024), 2024.

[234] Lu et al., MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts (ICLR 2024), 2024.

[235] Pramanick et al., SPIQA: A Dataset for Multimodal Question Answering on Scientific Papers (NeurIPS 2024), 2024.

[236] Jiang et al., Automated Data-Driven Storytelling with Visualizations and Texts (EMNLP 2024), 2024.

[237] Sah et al., NL4DV-LLM: Generating Analytic Specifications for Data Visualization from Natural Language Queries Using LLMs (IEEE VIS 2024), 2024.

H.8 2025

[238] Oprea et al., Rethinking Chart Understanding Using Multimodal Large Language Models, 2025.

[239] Wei et al., ChartMind: A Comprehensive Benchmark for Complex Real-world Multimodal Chart QA, 2025.

[240] Xu et al., ChartPoint: Guiding MLLMs with Grounding Reflection for Chart Reasoning, 2025.

[241] Li et al., ChartAnchor: Chart Grounding with Structural-Semantic Fidelity and Data Recovery, 2025.

[242] Hegde et al., ChartQA-X: Generating Explanations for Visual Chart Reasoning, 2025.

[243] Kaur et al., ChartAgent: A Multimodal Agent for Visually Grounded Reasoning in Complex Chart Question Answering, 2025.

[244] Huang et al., From Pixels to Insights: A Survey on Automatic Chart Understanding in the Era of Large Foundation Models, 2025.

[245] Huang et al., Five Years of SciCap: What We Learned and Future Directions for Scientific Figure Captioning, 2025.

[246] Masry et al., ChartQAPro: A More Diverse and Challenging Benchmark for Chart Question Answering, 2025.

[247] Huang et al., EvoChart: A Benchmark and a Self-Training Approach Towards Real-World Chart Understanding, 2025.

[248] Huang et al., VProChart: Answering Chart Question through Visual Perception Alignment Agent (AAAI 2025), 2025.

[249] Kim et al., SIMPLOT: Enhancing Chart Question Answering by Distilling Essentials (NAACL 2025 Findings), 2025.

[250] Chen et al., Chart-HQA: A Benchmark for Hypothetical Question Answering in Charts, 2025.

[251] Xie et al., InfoChartQA: A Benchmark for Multimodal Question Answering on Infographic Charts, 2025.

[252] Hutchinson et al., Chart Question Answering from Real-World Analytical Narratives (ACL 2025 SRW), 2025.

[253] Tang et al., ChartMuseum: Testing Visual Reasoning Capabilities of Large Vision-Language Models (NeurIPS 2025 D\&B), 2025.

[254] Tang et al., Chart2Code: A Hierarchical Benchmark for Multimodal Models, 2025.

[255] Kim et al., Chart-to-Experience: Benchmarking Multimodal LLMs for Predicting Experiential Impact of Charts (PacificVis 2025), 2025.

[256] Masry et al., BigCharts-R1: Enhanced Chart Reasoning with Visual Reinforcement Finetuning, 2025.

[257] Chen et al., Chart-R1: Chain-of-Thought Supervision and Reinforcement for Advanced Chart Reasoner, 2025.

[258] He et al., Distill Visual Chart Reasoning Ability from LLMs to MLLMs (EMNLP 2025 Findings), 2025.

[259] Zhao et al., ChartM3: A Multi-Stage Code-Driven Pipeline for Multi-Dimensional Visual Reasoning in Chart Comprehension (EMNLP 2025 Findings), 2025.

[260] Zhang et al., Do MLLMs Really Understand the Charts? ChartVRBench \& ChartReasoner, 2025.

[261] Sinha et al., Chart-RVR: Reinforcement Learning with Verifiable Rewards for Explainable Chart Reasoning, 2025.

[262] Shen et al., RECODE: Reasoning Through Code Generation for Visual Question Answering, 2025.

[263] Xu et al., ChartMoE: Mixture of Diversely Aligned Expert Connector for Chart Understanding (ICLR 2025 Oral), 2025.

[264] Masry et al., ChartGemma: Visual Instruction-tuning for Chart Reasoning in the Wild (COLING 2025), 2025.

[265] Yang et al., Effective Training Data Synthesis for Improving MLLM Chart Understanding (ICCV 2025), 2025.

[266] Ji et al., Socratic Chart: Cooperating Multiple Agents for Robust SVG Chart Understanding, 2025.

[267] Wang et al., Omni-Chart-600K: A Comprehensive Dataset of Chart Types for Chart Understanding (NAACL 2025 Findings), 2025.

[268] Hu et al., mPLUG-DocOwl 1.5: Unified Structure Learning for OCR-free Document Understanding (AAAI 2025), 2025.

[269] Zhao et al., ChartCoder: Advancing Multimodal Large Language Model for Chart-to-Code Generation (ACL 2025), 2025.

[270] Rodriguez et al., Rendering-Aware Reinforcement Learning for Vector Graphics Generation, 2025.

[271] Lim et al., ChartCap: Mitigating Hallucination of Dense Chart Captioning, 2025.

[272] Chen et al., V-CoT: End-to-End Chart Summarization via Visual Chain-of-Thought, 2025.

[273] Zhao et al., From Charts to Fair Narratives: Uncovering and Mitigating Geo-Economic Biases in Chart-to-Text (EMNLP 2025), 2025.

[274] Yue et al., F2TEval: Human-Aligned Multi-Dimensional Evaluation for Figure-to-Text Task (EMNLP 2025), 2025.

[275] Vogel et al., RefChartQA: Grounding Visual Answer on Chart Images through Instruction Tuning (ICDAR 2025), 2025.

[276] Bansal et al., ChartAB: A Benchmark for Chart Grounding \& Dense Alignment, 2025.

[277] Liu et al., On the Perception Bottleneck of VLMs for Chart Understanding (EMNLP 2025 Findings), 2025.

[278] Huang et al., Why Vision Language Models Struggle with Visual Arithmetic? Towards Enhanced Chart and Geometry Understanding (ACL 2025 Findings), 2025.

[279] Salamatian et al., ChartGaze: Enhancing Chart Understanding in LVLMs with Eye-Tracking Guided Attention Refinement (EMNLP 2025), 2025.

[280] Su et al., ClimateViz: A Benchmark for Statistical Reasoning and Fact Verification on Scientific Charts (EMNLP 2025), 2025.

[281] Zhao et al., Evaluating MLLMs' Capability via Chart Editing (ACL 2025 Findings), 2025.

[282] Ni et al., VisCoder: Fine-Tuning LLMs for Executable Python Visualization Code Generation (EMNLP 2025), 2025.

[283] Jain et al., Doc2Chart: Intent-Driven Zero-Shot Chart Generation from Documents (EMNLP 2025), 2025.

[284] Rahman et al., Text2Vis: A Challenging and Diverse Benchmark for Generating Multimodal Visualizations from Text (EMNLP 2025), 2025.

[285] Kim et al., Multi-LLM Collaborative Caption Generation in Scientific Documents (AAAI 2025 Workshop), 2025.

[286] Singh et al., FigCaps-HF: A Figure-to-Caption Generative Framework and Benchmark with Human Feedback (RANLP 2025), 2025.

[287] Kim et al., Personalized Scientific Figure Caption Generation: Author-Specific Writing Style Transfer, 2025.

[288] Timklaypachara et al., Leveraging Author-Specific Context for Scientific Figure Caption Generation: 3rd SciCap Challenge, 2025.

[289] Ng et al., LaMP-Cap: Personalized Figure Caption Generation With Multimodal Figure Profiles (EMNLP 2025 Findings), 2025.

[290] Srivastava et al., Enhancing Financial VQA in Vision Language Models using Intermediate Structured Representations, 2025.

[291] Chen et al., VisAnatomy: An SVG Chart Corpus with Fine-Grained Semantic Labels (IEEE VIS 2025 / TVCG), 2025.

[292] Xu et al., ChartBench: A Comprehensive Evaluation Benchmark for Chart Understanding, 2025.

[293] Gao et al., Interleaved-Modal Chain-of-Thought (CVPR 2025), 2025.

[294] Li et al., Graph-to-Vision: Multi-graph Understanding and Reasoning using Vision-Language Models, 2025.

[295] Van-Dinh et al., ViInfographicVQA: A Benchmark for VQA on Vietnamese Infographics, 2025.

[296] Li et al., SciGraphQA: A Large-Scale Synthetic Multi-Turn QA Dataset for Scientific Graphs (NAACL 2025), 2025.

[297] Chen et al., What Makes for Good Image Captions?, 2025.

[298] Ho et al., Table-Text Alignment: Explaining Claim Verification Against Tables in Scientific Papers (EMNLP 2025), 2025.

[299] Zhang et al., DocAssistant: Key-region Reading and Step-wise Reasoning for Robust Document VQA (EMNLP 2025), 2025.

[300] Jiang et al., TabDSR: Decompose, Sanitize, and Reason for Complex Numerical Reasoning in Tabular Data (EMNLP 2025), 2025.

[301] Li et al., Multimodal Document-level Triple Extraction via Dynamic Graph Enhancement (EMNLP 2025), 2025.

[302] Shlapentokh-Rothman et al., Visual Program Distillation with Template-Based Augmentation (EMNLP 2025), 2025.

[303] Xu et al., PreGenie: An Agentic Framework for High-quality Visual Presentation Generation (EMNLP 2025), 2025.

[304] Smock et al., PubTables-v2: A New Large-scale Dataset for Full-page and Multi-page Table Extraction, 2025.

[305] Yu et al., MACT: Visual Document Understanding via Multi-Agent Collaboration with Test-Time Scaling, 2025.

[306] Liu et al., Look as You Think: Visual RL for Verifiable Document RAG, 2025.

[307] Deng et al., Self-Improvement in Multimodal Large Language Models: A Survey, 2025.

[308] Wang et al., Train a Unified Multimodal Data Quality Classifier with Synthetic Data (EMNLP 2025), 2025.

[309] Cheng et al., On Domain-Adaptive Post-Training for Multimodal Large Language Models (EMNLP 2025 Findings), 2025.