International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

ChartSynth: Structured Summarization and Comparison of Multi-Chart Scientific Figures

Neelu Verma

Indian Institute of Technology Jodhpur, India
Email: *verma.14[at]iitj.ac.in*

Abstract: ***Multi-panel chart figures are ubiquitous in scientific publications, where multiple coordinated visualizations are used to compare experimental conditions, characterize parameter-dependent behavior, and expose complementary trends. However, existing chart understanding benchmarks and vision-language models predominantly operate on individual charts, limiting their ability to reason over relationships distributed across multiple panels. We introduce multi-chart summarization, a new task that requires jointly interpreting a set of charts and generating a concise summary of their similarities, differences, and cross-panel relationships. To study this task, we present MultiChartSum, a benchmark containing over 10,000 multi-panel scientific figures and 30,000+ chart panels paired with comparative summaries. We further propose ChartSynth, a modular vision-language framework that independently encodes chart panels, performs cross-panel visual-semantic alignment, and generates contrastive summaries through structured representation fusion. Unlike panel-wise captioning approaches, ChartSynth explicitly models inter-panel dependencies to distinguish shared trends from condition-specific changes. On the MultiChartSum test set, ChartSynth achieves 41.2 ROUGE-L, 89.5 BERTScore, and 75.6 Trend-F1, improving over the strongest LLaVA-MultiInput baseline by 5.6, 2.1, and 6.1 points, respectively. Human evaluation further shows that ChartSynth achieves 89.5% fluency, 79.6% insightfulness, and 76.8% comparative accuracy, substantially outperforming the evaluated baselines. These results demonstrate that explicit cross-chart alignment and relational reasoning are critical for generating scientifically meaningful summaries of multi-panel visualizations. MultiChartSum and ChartSynth establish a benchmark and modeling framework for structured multi-chart understanding and comparative vision-language reasoning.***

Keywords: Multi-chart summarization, Chart understanding, Vision-language models, Cross-chart reasoning, Scientific visualization, Multimodal learning, Comparative summariza-tion, ChartSynth

1. Introduction

Scientific publications frequently employ multi-chart figures to present complex experimental results, compare conditions, or illustrate temporal and causal relationships. These figures, often arranged in grid-like subplots (e.g., Figure 1), provide rich visual information that supports claims across diverse domains such as biology, physics, climate science, and machine learning. Understanding and summarizing such figures requires not only recognizing trends within individual charts but also identifying key similarities, differences, and relationships across related panels.

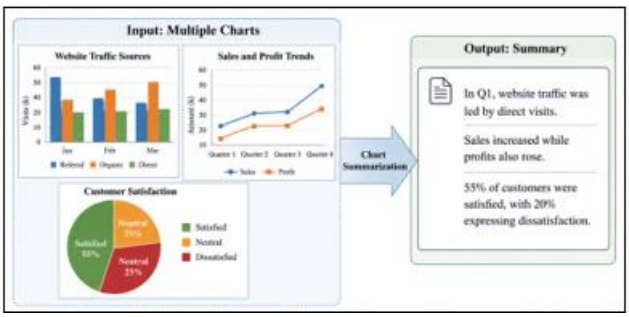

Figure 1: Overview of the multi-chart summarization task. Given multiple heterogeneous charts containing complementary visual information, the goal is to jointly interpret individual trends and cross-chart relationships and generate a concise, coherent, and evidence-grounded textual summary. Unlike single-chart summarization, the task requires identifying shared metrics, aligning conditions, comparing trends, and preserving salient information across charts.

Despite the prevalence of multi-chart figures in scientific literature, existing chart understanding research focuses predominantly on *single-chart tasks*, including question answering, retrieval, captioning, and fact verification. For example, ChartQA [1] introduced a benchmark for question answering over chart images, while subsequent work has explored chart retrieval [2], question answering over bar-chart images [3], and chart-grounded fact verification [4, 5, 6]. Although these approaches have substantially advanced chart understanding, they primarily treat charts as independent visual entities and do not explicitly model relationships among multiple coordinated charts.

This limitation becomes particularly important in scientific figures, where individual panels often provide complementary evidence rather than isolated information. A single panel may reveal a temporal trend, whereas another may show the corresponding trend under a different experimental condition. Meaningful interpretation therefore requires aligning shared axes, variables, legends, conditions, and visual patterns across panels. Simply processing each chart independently can result in redundant descriptions and may fail to identify comparative insights that emerge only when the charts are considered jointly.

To bridge this gap, we introduce the novel task of **multi-chart summarization**, which aims to generate structured or free-text summaries that explain shared trends, divergent patterns, and important relationships across a set of related charts. Unlike conventional chart captioning or single-chart summarization, the proposed task requires the model to reason jointly over multiple visual inputs and determine which relationships are most relevant for the final summary. This introduces several challenges, including cross-panel

visual-semantic alignment, identification of corresponding variables and conditions, comparison of numerical and temporal trends, and generation of a coherent narrative without unnecessary repetition.

To address these challenges, we propose **ChartSynth**, a structured vision-language framework for cross-chart understanding and summarization. ChartSynth first encodes each chart panel using a shared vision-language backbone and subsequently performs cross-panel alignment to capture relationships among the visual-semantic representations. The aligned representations are then integrated by a summary generation module to produce concise and comparative textual descriptions. This design explicitly separates individual chart understanding from cross-chart reasoning, enabling the model to preserve panel-specific evidence while also capturing relationships that emerge across the complete figure.

We further introduce MultiChartSum, a benchmark dataset specifically designed for multi-chart summarization (Figure 2). The dataset comprises scientific multi-panel figures paired with comparative summaries generated through a combination of expert annotation and controlled LLM-assisted generation. It covers diverse chart types, layouts, and scientific domains, enabling systematic evaluation of both linguistic quality and cross-chart reasoning capability.

Our experimental evaluation compares ChartSynth with strong vision-language and chart summarization baselines. Results show that explicitly modeling cross-chart relationships improves both summary quality and comparative reasoning. In addition to automatic evaluation using ROUGE-L and BERTScore, we conduct domain-specific trend evaluation and human assessment to measure fluency, insightfulness, and comparative accuracy. The results demonstrate that ChartSynth can produce summaries that are not only linguistically coherent but also better aligned with the relationships represented across multiple chart panels.

The main contributions of this work are summarized as follows:

- We define the novel task of **multi-chart summarization**, which extends conventional chart understanding from individual visualizations to coordinated multi-panel figures.
- We introduce **MultiChartSum**, a bench-mark dataset containing more than 10,000 scientific multi-panel figures paired with comparative summaries for evaluating cross-chart understanding and summarization.
- We propose **ChartSynth**, a structured vision-language framework that explicitly models cross-panel visual-semantic relationships for comparative summary generation.
- We conduct comprehensive experiments using automatic and human-centered evaluation metrics, demonstrating the effective-ness of cross-chart reasoning for generating fluent, informative, and comparative summaries.

2. Related Work

Our work intersects with several strands of research across chart understanding, vision-language modeling, and multimodal summarization. We review the most relevant directions below.

2.1 Chart Understanding and Rea-soning

Early efforts in chart understanding focus on synthetic datasets for question answering (e.g., FigureQA [7], PlotQA [8], ChartQA [1], where models answer questions based on bar or line charts. Verma [3] further explored question answering over bar-chart images using an optimization-based hybrid LST-ResNet architecture. More recent benchmarks such as Chart Retrieval [2], ChartFC [4], and ChartCheck [5] extend chart understanding to fact verification and declarative claim classification. More recently, Verma and Mishra [6] introduced a retrieval-augmented framework for chart-grounded fact verification with self-consistent vision–language reasoning. Despite these advances, existing approaches primarily focus on *individual* charts and do not explicitly address multi-chart understanding, cross-chart comparison, or the synthesis of coherent insights from multiple visualizations.

Several methods target chart-specific parsing, including chart-to-table extraction deplot [9], chart element detection Chart-to-text [10], and chart captioning Vistext [11]. While useful for downstream tasks, these approaches do not model relationships across multiple figures. In contrast, our work aims to generate coherent, high-level summaries that describe how related charts vary under different conditions-a reasoning capability not addressed by prior chart understanding models.

2.2 Vision-Language Summarization

Vision-language summarization has emerged as a key challenge in multimodal learning, aiming to generate coherent textual descriptions from visual input. Prior work has addressed summarization of image sequences VidSum [12], videos videosum [13], and egocentric photo streams egosum [14], but not structured data visualizations like scientific charts.

More closely related are recent efforts in chart captioning and trend description Vis-text, Chartgpt [11, 15], which generate textual summaries from single chart images. These approaches typically describe axes, trends, or visual marks, but do not generalize to multi-panel figures or model inter-chart comparisons.

Recent vision-language models (VLMs) such as BLIP [16], Flamingo [17], and LLaVA [18] have shown impressive capabilities in image-conditioned generation, including some zero-shot chart descriptions. However, these models treat input images independently and do not explicitly perform cross-image comparison or aggregation.

Our work differs by targeting the summarization of *multiple related charts*—a setting that requires reasoning about similarities and differences across panels, understanding vary-ing experimental conditions, and producing a concise summary that captures the collective insight. To our knowledge, this is the first work to propose and benchmark

this task.

2.3 Visual Comparison and Multi-Image Analysis

Comparing visual content across multiple im-ages has been explored in tasks such as image-text matching Vse++ [19], image retrieval [20], and set-level reasoning VisDiff [21]. However, these approaches often focus on unstructured image collections or visual similarity, rather than structured visualizations with semantic axes and numeric patterns.

Some prior work studies visual analogy rea-soning raven [22], visual entailment SNLI-VE [23], or visual question answering with mul-tiple images Multi-Image VQA [24], but these tasks are either synthetic or not grounded in scientific visualization. In contrast, our task requires comparing charts that encode struc-tured data and extracting trends or contrasts meaningful to scientific analysis.

Moreover, while layout-aware models Lay-outlm [25] and structured document under-standing methods docvqa [26] handle spatially arranged visual information, they do not model the semantics of numeric plots or the relation-ships between experimental conditions. Our approach draws inspiration from these works but focuses on trend comparison and summarization across chart panels.

2.4 Scientific Visualization and Multi-Panel Figures

Multi-panel figures are a hallmark of scientific communication, yet they remain largely unaddressed in multimodal AI research. Some recent efforts have explored figure mining Deep-Figures [27] and table summarization Table-sum [28], but comparable efforts for scientific charts are scarce.

Scientific document parsing tools science parse [29] extract figure metadata or captions, but do not understand the visual semantics of the charts themselves. Our work fills this gap by providing a benchmark and model for under-standing multi-chart scientific figures and summarizing their comparative insights, enabling downstream applications in scientific summa-rization, tutoring, and retrieval.

3. Task Definition

We define the task of **multi-chart summa-rization** as generating a concise and coherent textual summary that captures the similarities, differences, and key trends across a set of related chart images. Each input instance consists of a set of charts grouped as a single multi-panel figure, typically appearing as a composite image in scientific papers.

Formally, given a set of n chart images $C_1, C_2, \ldots, C_n$ that share a common context (e.g., varying a single experimental parameter), the goal is to produce a summary S that describes their collective insights:

$$S = \text{Summarize}(C_1, C_2, \ldots, C_n) \quad (1)$$

Unlike traditional summarization tasks that operate on text or video streams, this task is rooted in numeric visualizations that require quantitative reasoning. The model must not only perceive visual elements such as axes, la-bels, and plot types, but also interpret their semantic relationships and temporal or causal patterns.

3.1 Supported Chart Types and Settings

The task primarily targets structured charts such as bar charts, line graphs, and scatter plots. These are prevalent in scientific publications and often represent quantitative measurements under varying conditions. Each multi-chart fig-ure is assumed to contain subplots that differ by a controlled variable (e.g., hyperparameter settings, treatment conditions) while maintain-ing consistent semantics across axes or legends. We also consider cases where the x- or y-axis is shared across panels, or where global labels indicate the condition being varied.

The summarization task may optionally in-corporate access to figure captions or axis titles, though the core challenge lies in deriving in-sights from the chart images themselves.

3.2 Task Variants

While free-text summary generation is the primary objective, several auxiliary variants can enhance the task formulation and evaluation:

- **Structured Comparison Extraction:** Generate table-like key-value summaries describing key metric values or changes across panels.
- **Trend Classification:** For each panel or variable, classify the direction of change (increasing, decreasing, stable).
- **Difference Highlighting:** Explicitly contrast key differences, such as "Model A outperforms Model B under high noise."
- **Anomaly Detection:** Identify outlier panels that diverge from the dominant trend or contain inconsistent data.

These variants allow for more fine-grained model evaluation and support diverse down-stream applications such as retrieval, tutoring, and QA.

4. Challenges

Multi-chart summarization introduces several unique challenges for vision-language modeling:

- **Cross-Chart Alignment:** Charts may use different visual encodings (e.g., color, shape) or layout conventions that must be normalized for comparison.
- **Semantic Grouping:** The model must infer which charts belong to comparable conditions and which differences are meaningful.
- **Scale Normalization:** Y-axes across charts may differ in range or units, requiring normalization for trend interpretation.
- **Reasoning Across Panels:** Identifying consistent patterns (e.g., a performance metric improving as a parameter increases) requires relational and often causal reasoning.
- **Natural Language Generation:** Sum-maries must be fluent, concise, and non-redundant, avoiding repetition across similar panels.

These challenges motivate the design of our dataset and

model architecture, which are tailored to handle structural variability, trend alignment, and contrastive reasoning across charts.

4.1 Dataset Construction (MultiChartSum)

To enable research on multi-chart summarization, we construct **MultiChartSum**, a large-scale dataset of multi-panel scientific figures paired with human-authored and LLM-generated comparative summaries. This section outlines our data collection, preprocessing, annotation, and quality control processes.

4.2 Source Collection

We curate multi-chart figures from open-access scientific papers across diverse domains such as biomedical science, computer vision, climate studies, and social sciences. Figures are extracted using PDF parsing tools (e.g., GROBID, PDFFigures 2.0) from arXiv, PubMed Central, and selected conference proceedings. We retain only figures with at least two structured chart panels (e.g., bar, line, or scatter plots).

4.3 Panel Segmentation and Filtering

Each composite figure is decomposed into individual chart panels using layout segmentation and whitespace heuristics. To filter out tables and non-visual elements, we apply a trained classifier that distinguishes between chart types and other figure content. Visual OCR (via Tesseract) is used to extract axis titles, labels, and legends for metadata enrichment.

4.4 Summary Annotation

For each set of chart panels, we produce a summary that captures trends, differences, and salient observations. We employ a hybrid approach:

- **Human-Written Summaries:** Expert annotators summarize a subset of figure sets, focusing on accurate cross-chart comparison.
- **LLM-Generated Summaries:** We use GPT-4 with careful prompting to generate draft summaries, followed by human review and correction.

Each summary is designed to include comparative language, highlight key trends, and maintain scientific clarity.

4.5 Dataset Statistics

The final dataset includes over 10,000 multi-panel figures and 30,000+ chart panels. Average summary length is 3–5 sentences (approx. 50 words). Figures span more than 25 scientific fields and include varied layouts, chart types, and data scales. Each example includes:

- The composite figure and segmented chart panels.
- Extracted text metadata (e.g., axis titles, legends).

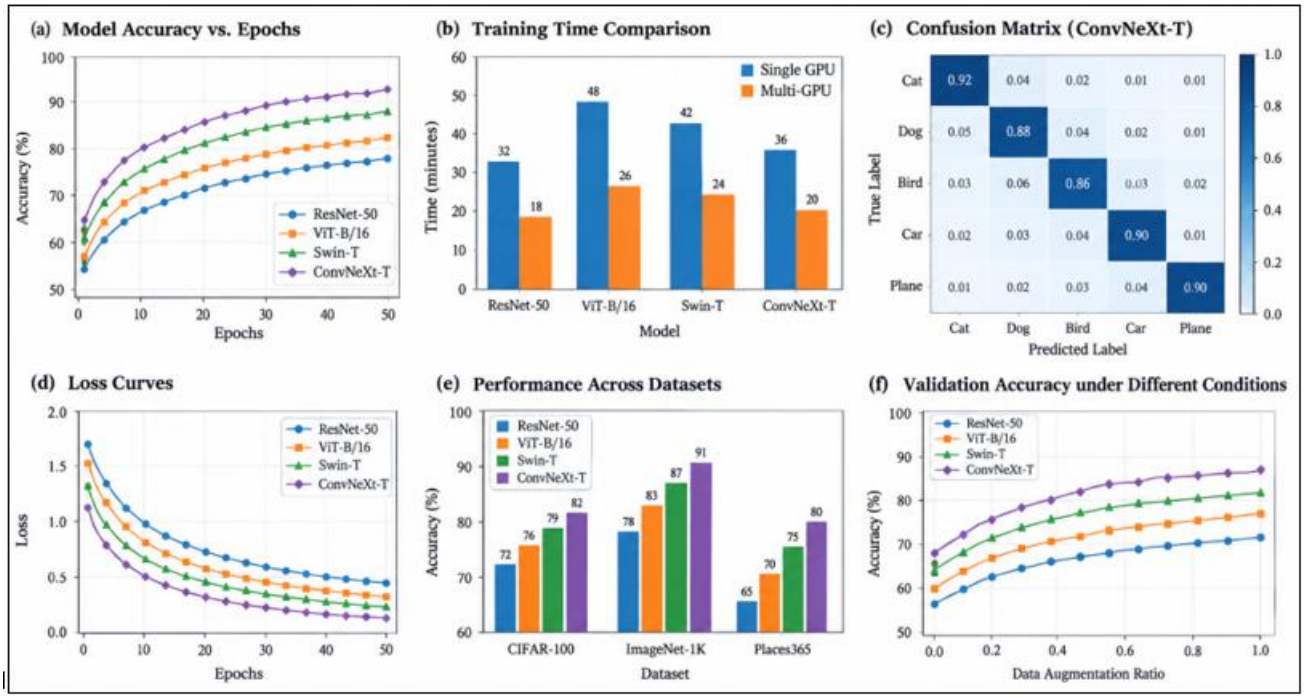

Figure 2: Construction of a single MultiChartSum dataset instance. A composite scientific figure is first collected from a research paper and decomposed into individual chart panels using layout segmentation. Each panel is enriched with extracted metadata (axes, legends, and labels), and the complete figure is paired with a human-verified comparative summary describing shared trends and cross-panel differences. This structured representation forms one training example for ChartSynth.

- A gold comparative summary.
- Chart-level annotations for selected samples (e.g., trend direction, highlight regions).

MultiChartSum offers the first large-scale testbed for structured multi-chart understanding, enabling both supervised training and rigorous benchmarking of vision-language models.

5. Method: ChartSynth

ChartSynth is a modular vision-language framework designed to perform cross-chart reasoning and generate comparative summaries for multi-panel figures. The system operates in three main stages: (1) chart encoding, (2) cross-chart alignment and representation fusion, and (3) summary generation.

5.1 Chart Encoding

Each chart panel is independently encoded using a vision-language backbone. We use a pre-trained VLM (e.g., BLIP-2 or LLaVA) to extract multimodal embeddings from the chart image and optionally associated metadata (e.g., axis titles, figure captions).

Given an input chart C_i, the model extracts a visual-textual embedding $v_i = \text{VLM}(C_i)$. This embedding captures chart semantics including plot type, axis content, and visual trends.

5.2 Cross-Chart Alignment and Fu-sion

To enable comparative reasoning, we project all $v_1, \ldots, v_n$ into a shared space and model inter-chart relationships using a transformer-based fusion module. This module captures similarities, differences, and contextual shifts across panels.

We apply a self-attention mechanism over the n chart embeddings to construct a fused summary representation z:

$$z = \text{CrossChartTransformer}(v_1, v_2, \ldots, v_n) \quad (2)$$

The transformer is trained to emphasize dimensions indicative of comparative insight, such as axes alignment or trend inversion.

5.3 Summary Generation

The fused representation z is passed to a language decoder (e.g., T5 or GPT-style model) to generate a textual summary:

$$S = \text{Decoder}(z) \quad (3)$$

We explore both fine-tuning and instruction prompting paradigms for the decoder. The decoder is trained with summaries from Multi-ChartSum, optimized using cross-entropy loss and optionally contrastive objectives.

5.4 Training Objectives

In addition to standard supervised generation loss, we incorporate two auxiliary objectives:

- **Contrastive Insight Loss:** Encourages the model to distinguish between semantically divergent chart sets.
- **Trend Direction Consistency:** Penalizes contradictory generation when visual trend direction (e.g., increasing vs decreasing) is misrepresented.

These components ensure the model captures not only surface-level appearance but also the high-level reasoning required for effective summarization.

In the next section, we describe experimental settings and evaluate ChartSynth against strong baselines.

6. Experiments and Results

We evaluate ChartSynth on the MultiChartSum benchmark and compare it to a suite of strong baselines across multiple evaluation metrics. Our experiments aim to assess the effectiveness of cross-chart reasoning, examine the benefits of structured summarization, and analyze component-wise contributions of the model.

6.1 Experimental Setup

Dataset Splits. We partition MultiChartSum into 70% training, 15% validation, and 15% test sets. Each split is stratified to maintain diversity in chart types and scientific domains.

Baselines. We compare ChartSynth against the following models:

- **SingleChart-GPT:** Applies a large language model (e.g., GPT-4) to each panel independently, concatenating outputs.
- **BLIP-2 (Image-Level):** Generates a single caption per panel and merges them post hoc.
- **LLaVA-MultiInput:** Encodes the full composite image and uses a prompt-based decoder.
- **ChartBART:** A fine-tuned multimodal encoder-decoder trained on chart-caption pairs.

Metrics. We evaluate summaries using:

- **ROUGE-L:** Measures overlap of longest common subsequences.
- **BERTScore:** Computes semantic similarity based on contextual embeddings.
- **Trend-F1:** A domain-specific metric assessing whether key trends are correctly summarized.
- **Human Evaluation:** Experts rate fluency, informativeness, and inter-panel com-parison.

6.2 Main Results

Table 1 presents the quantitative comparison on the MultiChartSum test set. ChartSynth achieves the best performance across all three automatic evaluation metrics, obtaining **41.2 ROUGE-L**, **89.5 BERTScore**, and **75.6**

Trend-F1. Compared with the strongest baseline, LLaVA-MultiInput, ChartSynth improves ROUGE-L by **5.6 points**, BERTScore by **2.1 points**, and Trend-F1 by **6.1 points**. The larger improvement in Trend-F1 is particularly notable, suggesting that explicitly modeling relationships across chart panels improves the preservation of trend-level information rather than merely increasing lexical or semantic similarity.

Compared with single-chart approaches, including SingleChart-GPT and BLIP-2, ChartSynth provides consistent gains across all metrics. This indicates that independently generating descriptions for individual panels is insufficient for multi-chart summarization, as the task requires reasoning over relationships that emerge only when multiple panels are considered jointly. ChartBART and LLaVA-MultiInput provide stronger results by jointly incorporating multimodal information, but remain below ChartSynth, highlighting the benefit of explicit cross-chart

alignment and relational fusion.

Human Evaluation. Automatic metrics alone may not fully capture whether a gener-ated summary correctly expresses relationships between multiple charts. We therefore conduct human evaluation using a 5-point rating scale, where scores of 4 or higher are considered satisfactory. As shown in Table 2, ChartSynth obtains the highest scores across all three criteria, achieving **89.5%** fluency, **79.6%** insightfulness, and **76.8%** comparative ac-curacy. Compared with LLaVA-MultiInput, ChartSynth improves comparative accuracy by **13.1 percentage points** and insightfulness by **11.7 percentage points**. These gains indicate that the improvement is not limited to surface-level textual quality; human evaluators also identify stronger cross-panel insights and more accurate comparisons in ChartSynth-generated summaries.

6.3 Ablation Studies

We conduct ablation experiments to quantify the contribution of the major components of ChartSynth. Specifically, we evaluate the effects of cross-chart attention, joint decoding, and the contrastive insight objective.

Effect of Cross-Chart Attention. Removing the cross-chart transformer results in a **6.1-point decrease in BERTScore**, demonstrating the importance of explicitly modeling depen-dencies between chart panels. Without cross-panel attention, the model has limited access to relationships between corresponding visual and semantic features and tends to produce more panel-specific descriptions.

Effect of Joint Decoding. Replacing the shared summary decoder with independent decoders for individual charts reduces Trend-F1 by 9.0 points. This substantial degradation indicates that independently

Table 1: Comparison on the MultiChartSum test set. Best results are shown in bold

Model	ROUGE-L	BERT Score	Trend-F1
SingleChart-GPT	29.8	83.1	61.2
BLIP-2	31.4	84.6	63.0
ChartBART	33.2	85.9	66.4
LLaVA-MultiInput	35.6	87.4	69.5
ChartSynth (ours)	41.2	89.5	75.6

Table 2: Human evaluation results. Values indicate the percentage of summaries rated 4 or higher on a 5-point scale. Best results are shown in bold

Model	Fluency	Insight	Comp. Acc.
ChartBART	78.4	62.1	58.9
LLaVA-MultiInput	81.2	67.9	63.7
ChartSynth (ours)	**89.5**	**79.6**	**76.8**

generated panel descriptions do not adequately capture relationships that emerge at the figure level. Joint decoding allows the model to condition the generated summary on information from multiple panels simultaneously.

Effect of Contrastive Insight Loss. Removing the contrastive insight objective results in more repetitive and less focused summaries. The degradation indicates that the auxiliary objective encourages the model to distinguish between visually similar chart sets that convey different scientific conclusions and improves the discriminative quality of generated summaries.

Overall, the ablation results support the central design of ChartSynth: effective multi-chart summarization requires explicit cross-panel interaction rather than simple aggregation of independently generated chart descriptions.

6.4 Quantitative Analysis

Figure 3 presents representative qualitative comparisons between ChartSynth and competing approaches. The examples demonstrate that the principal advantage of ChartSynth lies in its ability to synthesize information across panels rather than merely describe each chart independently.

For example, in a multi-panel figure comparing model performance under increasing noise levels, ChartSynth identifies the contrasting behavior of the evaluated models and generates a relationship-aware summary such as “Model A maintains performance under noise, while Model B degrades rapidly.” In contrast, panel-wise baselines tend to describe the individual curves separately and may fail to explicitly state the cross-panel contrast. Similar behavior is observed for examples involving trend consistency, condition-specific changes, and complementary evidence across panels.

These qualitative results complement the quantitative improvements in Trend-F1 and human comparative accuracy, suggesting that the gains of ChartSynth arise from its ability to reason over relationships distributed across multiple visualizations rather than from improved caption fluency alone.

6.5 Generalization Across Domains

To evaluate the robustness of ChartSynth beyond the training distribution, we conduct a cross-domain evaluation using scientific figures from domains not included in the training data, including economics and chemistry. Without additional fine-tuning, ChartSynth achieves a BERTScore of **87.0**, indicating that the learned cross-chart representation can transfer to previously unseen scientific domains.

This result suggests that the proposed framework captures relationships between visual structures and semantic trends that are not restricted to a particular scientific domain. Nevertheless, the performance gap relative to the in-domain evaluation indicates that domain-specific terminology, visualization conventions, and scale variations remain challenging.

6.6 Error Analysis

Despite its overall performance, ChartSynth exhibits several recurring failure modes. First, inconsistent axis scales across panels can lead to incorrect magnitude comparisons, particularly when panels use different ranges or units. Second, missed legends or ambiguous variable names can cause incorrect associations between visual elements and their corresponding entities. Third, highly noisy or densely

populated plots can result in overgeneralized summaries

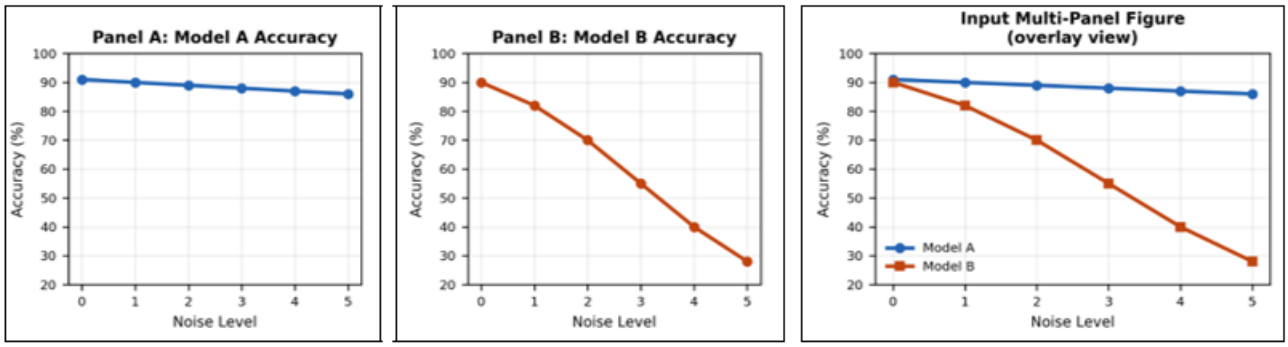

Figure 3: Qualitative comparison between ChartSynth and a panel-wise baseline (LLaVA- MultiInput) on a multi-chart summarization example. Given a two-panel input figure showing accuracy versus noise level for Model A and Model B, ChartSynth explicitly identifies the cross-panel contrast, generating the relationship-aware summary "Model A maintains stable performance as noise increases, while Model B degrades rapidly under the same conditions." The panel-wise baseline instead describes each curve independently and fails to state the cross-panel contrast, illustrating the benefit of explicit cross-chart reasoning for comparative summarization that capture the dominant trend while overlooking localized changes.

These failure cases indicate several directions for future improvement, including explicit axis normalization, stronger legend and entity grounding, and uncertainty-aware reasoning for noisy visualizations. Importantly, these limitations also highlight that multi-chart summarization requires not only language generation but reliable extraction and alignment of structured visual evidence across panels.

7. Conclusion and Future Work

7.1 Conclusion

In this work, we introduced the task of **multi-chart summarization**, addressing the gap between real-world scientific communication and current chart understanding methods. We presented **ChartSynth**, a structured vision-language framework capable of cross-panel reasoning and comparative summary generation. To facilitate development and benchmarking, we introduced **MultiChartSum**, a large-scale dataset of scientific multi-panel figures paired with comparative summaries.

Extensive experiments show that ChartSynth significantly outperforms strong baselines in both automatic and human evaluations, confirming the value of structured cross-chart alignment. Our method captures subtle trends, highlights meaningful differences, and generalizes well to unseen domains.

7.2 Future Work

Several avenues remain open for future exploration:

- **Interactive Summarization:** Incorporating user queries or focus regions to guide the summarization process.
- **Multilingual Support:** Extending the framework to summarize figures from on-English publications.
- **Temporal or Hierarchical Figures:** Handling sequences of chart panels or nested figure structures.
- **Cross-modal Retrieval:** Using chart summaries for document or figure-level retrieval tasks.
- **Explainable Summarization:** Generating rationales or visual justifications for claims in summaries.

We hope our work serves as a foundation for future progress in multimodal scientific understanding and opens new directions in chart-based vision-language research.

Declarations

Use of AI Technology
AI-assisted technologies were used during the preparation of this manuscript for language editing, grammar correction, refinement of wording, and improvement of presentation. The authors reviewed and verified all AI-assisted content and remained fully responsible for the accuracy, originality, and integrity of the manuscript. AI technologies were not used to generate or fabricate experimental results, data, or scientific conclusions.

Conflicts of Interest
All authors declare that they have no conflicts of interest.

Informed Consent
This study did not involve human participants; therefore, informed consent was not applicable.

Data Availability
The datasets and materials used in this study are described in the manuscript. The MultiChartSum benchmark and associated resources will be made available subject to the applicable data and usage restrictions.

References

[1] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In Findings of the Association for Computational Linguistics: ACL, pages 2263–2279. Association for Computational Linguistics, 2022.
[2] Neelu Verma, Anik De, and Anand Mishra. Bridging language to visuals: towards natural language query-to-chart image retrieval. International Journal of Multimedia Information Retrieval, 13(3):32, 2024.10
[3] Neelu Verma. Lst-resnet: Optimization- based hybrid network architecture for an effective question and

answering system with bar chart images. International Journal of Scientific Research and Technology, 3(09):243–265, 2026.

[4] Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. Reading and reasoning over chart images for evidence-based automated fact-checking. In Findings of the European Chapter of the Association for Computational Linguistics, pages 399–414. Association for Computational Linguistics, 2023.

[5] Mubashara Akhtar, Nikesh Subedi, Vivek Gupta, Sahar Tahmasebi, Oana Cocarascu, and Elena Simperl. Chartcheck: Explainable fact-checking over real-world chart images. In Findings of the Association for Computational Linguistics: ACL 2024, pages 13921–13937, Bangkok, Thailand, 2024. Association for Computational Linguistics.

[6] Neelu Verma and Anand Mishra. Retrieval-augmented chart-grounded fact verification with self-consistent vision–language reasoning. International Journal of Data Science and Analytics, 22(1):300, 2026.

[7] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. In the 6th International Conference on Learning Representations. Open-Review.net, 2018.

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

[9] Fangyu Liu, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. Deplot: One-shot visual language reasoning by plot-to-table translation. In Findings of the Association for Computational Linguistics, pages 10381–10399. Association for Computational Linguistics,2023.

[10] Shankar Kantharaj, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. Chart-to- text: A large-scale benchmark for chart summarization. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4005–4023, 2022.

[11] Benny Tang, Angie Boggust, and Arvind Satyanarayan. 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, 2023.

[12] Nishit Anand, Rupesh Kumar Koshariya, and Varsha Garg. Vidsum-video summarization using deep learning. In 2023 Second International Conference on Informatics (ICI), pages 1–6. IEEE, 2023.

[13] Luis C Garcia-Peraza-Herrera, Sebastien Ourselin, and Tom Vercauteren. Videosum: A python library for surgical video summarization. arXiv preprint arXiv:2303.10173, 2023.

[14] Jia Xu, Lopamudra Mukherjee, Yin Li, Jamieson Warner, James M Rehg, and Vikas Singh. Gaze-enabled egocentric video summarization via constrained sub-modular maximization. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2235–2244, 2015.

[15] Yuan Tian, Weiwei Cui, Dazhen Deng, Xinjing Yi, Yurun Yang, Haidong Zhang, and Yingcai Wu. Chartgpt: Leveraging llms to generate charts from abstract natural language. IEEE Transactions on Visualization and Computer Graphics, 31(3):1731–1745, 2024.

[16] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In the International conference on machine learning, pages 12888–12900. PMLR, 2022.

[17] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems, 35:23716– 23736, 2022.

[18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. Advances in neural information processing systems, 36:34892–34916, 2023.

[19] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. arXiv preprint arXiv:1707.05612, 2017.

[20] Yunchao Wei, Yao Zhao, Canyi Lu, Shikui Wei, Luoqi Liu, Zhenfeng Zhu, and Shuicheng Yan. Cross-modal retrieval with cnn visual features: A new baseline. IEEE transactions on cybernetics, 47(2):449–460, 2016.

[21] Lisa Dunlap, Yuhui Zhang, Xiaohan Wang, Ruiqi Zhong, Trevor Darrell, Jacob Steinhardt, Joseph E Gonzalez, and Serena Yeung-Levy. Describing differences in image sets with natural language. In Proceedings of the IEEE/CVF Conference onComputer Vision and Pattern Recognition,pages 24199–24208, 2024.

[22] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 5317–5327,2019.

[23] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. arXiv preprint arXiv:1901.06706, 2019.

[24] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In Proceedings of the 57th annual meeting of the association for computational linguistics, pages 6418–6428, 2019.

[25] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. In Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, pages 1192–1200, 2020.

[26] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In Proceedings of the IEEE/CVF winter conference on applications of computer vision, pages

2200–2209, 2021.

[27] Noah Siegel, Nicholas Lourie, Russell Power, and Waleed Ammar. Extracting scientific figures with distantly supervised neural networks. In Proceedings of the 18th ACM/IEEE joint conference on digital libraries, pages 223–232, 2018.

[28] Shuo Zhang, Zhuyun Dai, Krisztian Balog, and Jamie Callan. Summarizing and exploring tabular data in conversational search. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 1537–1540, 2020.

[29] Allen Institute for Artificial Intelligence. Science parse. https://github.com/allenai/science-parse, 2015.