ChartMVC: A General Multi-View Contrastive Learning Framework for Robust Chart Question Answering

Bin Wang[1], Jan Hua Li[1]
School of Information Science and Engineering, East China University of Science and Technology
Shanghai, China
{jhli}@ecust.edu.cn

***Abstract*—Chart Question Answering (ChartQA) has recently gained significant attention in visual data understanding. Despite achieving promising results on standard benchmarks, existing vision-language models—ranging from lightweight architectures to recent Large Multimodal Models (LMMs)—often degrade significantly under diverse chart styles and visual disturbances, such as variations in color, font, and layout. To improve robustness, we propose ChartMVC(Chart Multi-View Consistency Contrastive learning), a general Multi-View Contrastive Learning framework designed to enhance visual stability in chart understanding. ChartMVC forces the model to learn semantically consistent representations across diverse visual styles through a queue-based contrastive mechanism. To validate the effectiveness of our framework, we instantiate it using the UniChart architecture. Extensive experiments on the ChartQA benchmark show that our approach significantly outperforms the original baseline in relaxed accuracy metrics. Furthermore, the model exhibits superior robustness against various chart perturbations, demonstrating the effectiveness of the ChartMVC framework for real-world chart understanding tasks.**

***Index Terms*—Chart Question Answering, Contrastive Learning, Robustness, Multi-View Representation**

I. Introduction

Data visualizations, such as bar, line, and scatter plots, are indispensable for modern data analysis, enabling users to rapidly identify patterns and support scientific decision-making [1]. As the volume and complexity of visual data grow, the automated interpretation and understanding of these charts has emerged as a critical yet challenging research frontier [5]–[7]. Specifically, **Chart Question Answering (ChartQA)** [2]–[4] stands as a core task within chart understanding. Unlike general visual question answering, the task of ChartQA demands a unique synergy of capabilities: models must not only parse heterogeneous structural layouts (e.g., axes, legends, and coordinate systems) but also perform high-precision numerical extraction and multi-step reasoning. Consequently, ChartQA represents a core challenge for artificial intelligence, testing a model's ability to both "read" and "calculate" directly from pixels through the deep integration of visual perception and mathematical logic.

To tackle these challenges, the field has undergone a significant paradigm shift. Early approaches primarily relied on OCR-based pipelines, which first converted charts into intermediate tabular formats before applying text-based reasoning. However, these methods often suffered from error propagation during the extraction stage. To overcome this, the community transitioned toward OCR-Free end-to-end architectures. Milestone models such as Pix2Struct [4] introduced visual-language pre-training tasks like screenshot parsing, while MatCha [8] incorporated chart de-rendering and mathematical reasoning to enhance structural comprehension. Building on these, UniChart [9] further advanced this paradigm by proposing a unified vision-language framework. Instead of relying on explicit structural encoders, UniChart achieves comprehensive understanding by training on a diverse set of objectives, ranging from low-level visual element extraction to high-level reasoning and summarization.

More recently, the research focus has shifted towards Multimodal Large Language Models (MLLMs). By leveraging the extensive world knowledge and reasoning capabilities of Large Language Models (LLMs), recent studies have explored visual instruction tuning to empower models with chart literacy. Approaches such as those presented in ChartInstruct [33] and ChartGemma [12] have demonstrated superior performance in complex logical reasoning and zero-shot generalization. These models are typically trained on diverse, large-scale chart-instruction pairs to align visual representations with the LLM's text embedding space, effectively bridging the gap between visual perception and user intent.

However, as illustrated in Fig. 1, current state-of-the-art models often exhibit semantic inconsistency when presented with visually distinct charts that encode the same underlying data. Although these models achieve impressive performance on standard benchmarks, they remain fragile to stylistic and structural variations—such as changes in color schemes, fonts, or axis configurations. This sensitivity suggests that standard encoders often capture style-biased features rather than invariant data semantics. Recent studies like RobustCQA [10] and ChartInsight [11] confirm this, revealing significant performance degradation under visual perturbations and underscoring a persistent gap in the robustness of current models.

To address this challenge, we introduce **ChartMVC** (**Chart** **M**ulti-**V**iew **C**onsistency contrastive learning), a generalizable multi-view contrastive learning framework designed to enhance semantic consistency. The core philosophy of **ChartMVC** is that while visual realizations (e.g., styles, colors) may

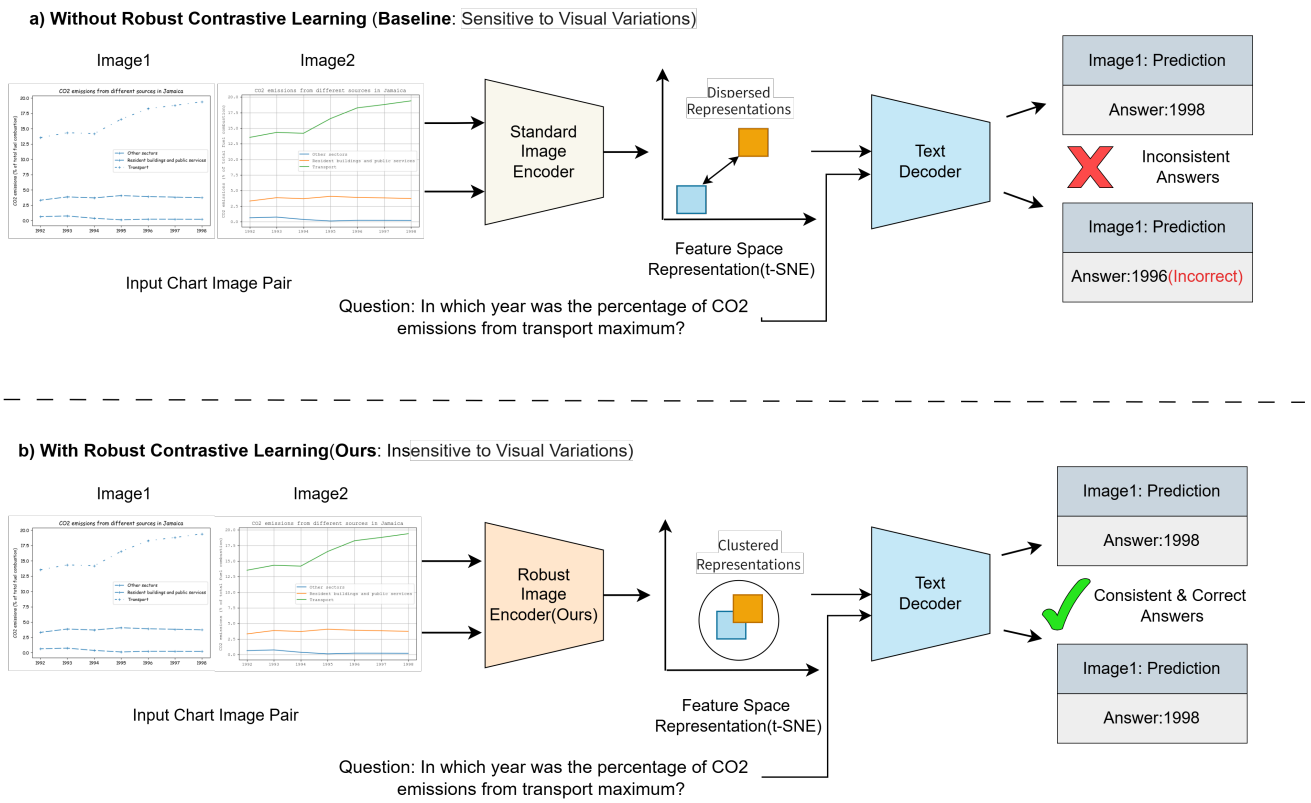


Fig. 1: **Motivation: Overcoming Visual Inconsistency via Feature Alignment.** The upper section (Without Robust Contrastive Learning) depicts how a standard encoder processes two visually distinct charts sharing the same underlying data. Due to sensitivity to visual styles, the encoder maps these semantically equivalent inputs to distant regions in the feature space (dispersed representations), leading to inconsistent predictions. In contrast, the lower section (With Robust Contrastive Learning) demonstrates the effect of our contrastive consistency framework. By explicitly optimizing for invariance, the encoder projects the original and augmented views into tightly clustered representations in the embedding space. This feature alignment ensures that the model ignores stylistic perturbations (e.g., color or font changes) and generates consistent, correct answers.

vary, the intrinsic data semantics remain invariant. Building on this, we design a model-agnostic contrastive pretraining strategy that explicitly encourages the encoder to align representations of visually diverse but semantically equivalent charts. By doing so, the framework forces the model to decouple data semantics from visual styles, learning features that are robust to real-world perturbations.

To enable this alignment, we construct **MV-ChartQA** (**M**ulti-**V**iew **Chart** **Q**uestion **A**nswering), a multi-view dataset organized into "view groups," each containing an original chart and its visually augmented variants (e.g., color, layout) with aligned QA pairs. By leveraging these groups with a feature-queue contrastive mechanism, **ChartMVC** introduces an auxiliary loss alongside standard QA supervision. This formulation forces the encoder to learn semantically consistent representations despite visual shifts, effectively serving as a model-agnostic module to enhance robustness without requiring architectural changes.

To evaluate the effectiveness of our approach, we instantiate the ChartMVC framework on the UniChart architecture. We conduct extensive experiments on the ChartQA benchmark and perform additional robustness analyses under various visual perturbations. The results demonstrate that the framework-enhanced model maintains significantly more stable predictions across diverse chart appearances, validating the effectiveness and generalization capability of ChartMVC in complex real-world scenarios.

Our main contributions are summarized as follows:

- **Multi-View Consistency Framework (ChartMVC):** We propose a model-agnostic contrastive learning framework that explicitly enforces representation consistency across diverse chart views. By maximizing the similarity between semantically equivalent but visually distinct views, ChartMVC enables the encoder to decouple visual styles from intrinsic data semantics, achieving superior robustness.
- **Multi-view Chart Dataset (MV-ChartQA):** We introduce **MV-ChartQA**, a large-scale dataset comprising ∼18,000 view groups derived from PlotQA. By aligning original charts with diverse visual augmentations, it

provides the essential multi-view supervision required for robust contrastive learning.

- **Robust and lightweight model performance:** We validate the effectiveness of our framework by applying it to the UniChart baseline. The resulting model consistently outperforms the original version across all subsets of the ChartQA benchmark (e.g., +2.08% on the Human subset) and exhibits superior stability under visual perturbations, demonstrating that ChartMVC effectively enhances model robustness while maintaining a compact architecture.

II. Related Work

A. Chart Understanding and Question Answering Models

Chart comprehension and question answering (ChartQA) have evolved significantly, transitioning from rule-based systems to end-to-end neural architectures. Early OCR-free models, such as Pix2Struct [4] and MatCha [8], advanced this field by parsing chart layouts and performing mathematical reasoning directly from images. However, these models often focused heavily on layout understanding or table generation, limiting their capabilities in complex visual reasoning and diverse text generation.

To address these limitations, chart-specific vision-language models like UniChart [9] were introduced. Building on the Donut encoder-decoder architecture, UniChart explicitly aligns visual components (e.g., bars, lines) with textual elements (e.g., legends, labels) and underlying data values. While UniChart achieves state-of-the-art results on standard benchmarks, lightweight models of this class remain sensitive to visual perturbations due to limited training data diversity and the lack of explicit robustness-oriented pretraining objectives.

Recently, the field has seen a shift towards specialized Multimodal Large Language Models (MLLMs) and Program-of-Thoughts (PoT) strategies [32]. Models such as ChartLlama [13],ChartGemma [12], and TinyChart [14] leverage large-scale instruction tuning and code generation to enhance reasoning. For instance, ChartLlama utilizes synthetic data generated by GPT-4 to create a diverse instruction-tuning corpus, while TinyChart integrates a visual token merging module and PoT learning to perform numerical calculations via Python programs. Similarly, ChartInstruct [33] and ChartAssistant [34] demonstrate that fine-tuning on diverse chart instructions significantly improves generalization across downstream tasks.

Parallel to these specialized models, general-purpose MLLMs like GPT-4V and LLaVA have shown impressive zero-shot capabilities in chart analysis. However, despite these advancements in reasoning and generalization, recent studies indicate that both lightweight models and large-scale MLLMs still struggle with visual robustness. They often yield inconsistent predictions when facing stylistic variations or low-level visual perturbations. This persistent limitation highlights the critical need for learning paradigms that can explicitly enforce semantic consistency and disentangle invariant data features from diverse visual renderings.

B. Robustness of Vision-Language Models in ChartQA

The robustness of vision-language models (VLMs) has recently become an important research focus [15], [16]. Prior studies [17] have shown that these models are highly sensitive to visual perturbations, adversarial noise, and stylistic variations, raising concerns about their reliability in real-world applications. In chart understanding tasks, even subtle changes in color schemes, scaling, or layout can lead to substantial prediction shifts. Early work examined the robustness of models such as DePlot and MatCha under visual and structural perturbations, but the analyses mainly focused on limited types of modifications and question categories, lacking a systematic robustness evaluation across diverse chart elements.

More recently, the community has moved towards systematic benchmarking to probe model limitations. Representative studies such as RobustCQA [10] and ChartInsights [11] evaluate the consistency of Large Multimodal Models (LMMs) across perturbed representations and low-level visual noises. In parallel, ChartAB [38] pioneers the evaluation of dense grounding and cross-chart alignment, revealing that models exhibit significant perception biases and hallucination when subjected to visual attribute perturbations. Similarly, ChartX [37] introduces a versatile evaluation covering 18 chart types across 22 disciplines, highlighting the necessity of diverse data for general reasoning. Furthermore, CharXiv [35] extends this rigorous testing to the scientific domain. Collectively, these studies reveal that regardless of model scale—from lightweight architectures to GPT-4V—state-of-the-art models suffer from notable performance degradation when facing complex or visually perturbed real-world charts, underscoring a persistent gap in robustness.

However, these data-centric approaches primarily address inter-chart diversity (generalizing across different chart types) rather than intra-chart consistency (robustness to stylistic variations of the same chart). Consequently, even models trained on such large-scale datasets remain fragile to low-level visual perturbations (e.g., color or font shifts), highlighting the need for our ChartMVC framework which explicitly enforces feature-space consistency.

Complementing these efforts, our research advances this line of work by integrating multi-view augmentation directly into a contrastive learning framework. While previous methods largely focus on data diversity, our ChartMVC framework explicitly optimizes for feature-space consistency. By forcing the encoder to map visually distinct but semantically equivalent views to aligned representations, our approach offers a rigorous mechanism for learning perturbation-invariant features, effectively bridging the gap between data augmentation and representation learning.

III. Method Overview

We present ChartMVC, a generalizable contrastive learning framework designed to enhance the robustness of chart understanding models. As illustrated in Fig. 2, our framework is model-agnostic and can be applied to various vision-language architectures, ranging from large multimodal models

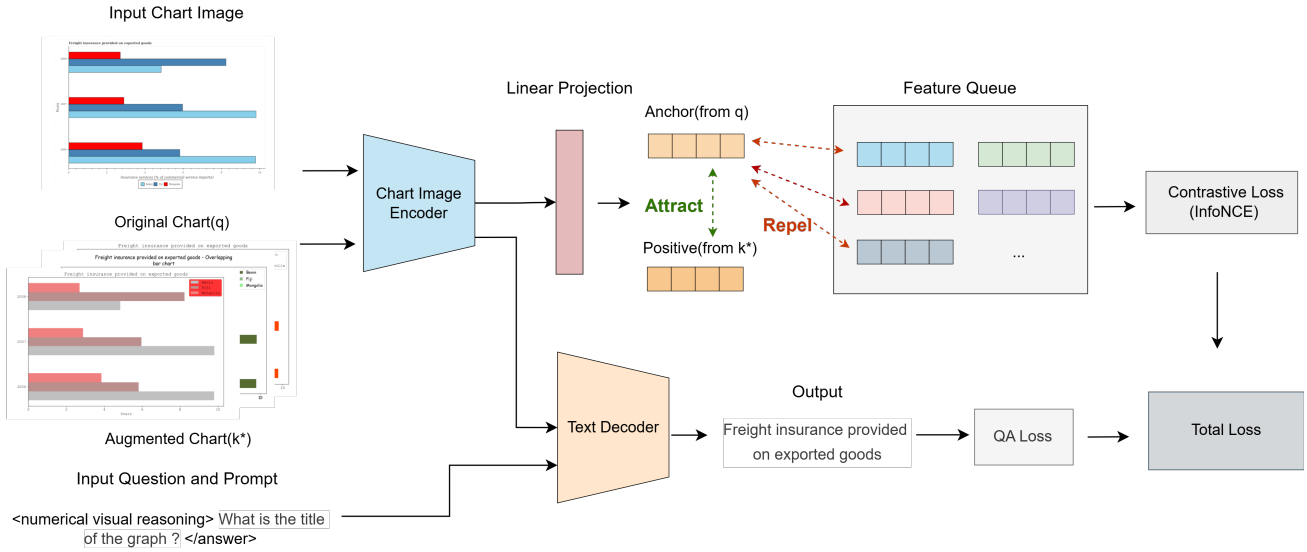


Fig. 2: **Overview of the ChartMVC framework.** The training pipeline consists of two parallel objectives: (1) **Contrastive Consistency Learning (Top):** An Original Chart (q) and its visually perturbed Augmented Chart (k^*) are processed by a shared Chart Image Encoder and mapped into an embedding space via a Linear Projection head. The framework explicitly enforces semantic robustness by attracting the Anchor (q) to its Positive counterpart (k^*) while repelling it from diverse negative samples stored in a dynamic Feature Queue (via InfoNCE loss). (2) **Chart Question Answering (Bottom):** Simultaneously, the visual features of the original chart and the input question are fed into the Text Decoder to generate the target answer. The model is trained end-to-end by optimizing a joint Total Loss, ensuring the learned representations are both style-invariant and semantically rich.

to lightweight encoders. To demonstrate its effectiveness, we instantiate ChartMVC using UniChart, a representative lightweight model specialized for chart understanding, as the backbone.

Below, we briefly describe the backbone architecture, followed by a detailed exposition of our proposed multi-view contrastive consistency mechanism.

A. Backbone Architecture: UniChart Instantiation

To validate our framework, we employ the UniChart architecture, which follows a unified *image encoder + text decoder* structure capable of end-to-end chart question answering.

a) Image Encoder.: We adopt the **Donut** [18] encoder, an OCR-free architecture based on the Swin Transformer [19]. It divides the input chart image into fixed-size patches and applies window-based multi-head self-attention to extract high-level semantic features. This encoder is particularly suitable for our framework as it captures the three key aspects of charts: textual elements (e.g., axis labels), graphical elements (e.g., bars, lines), and the overall spatial layout.

b) Text Decoder.: Following the UniChart design, we employ a **BART** [20] decoder for text generation. Conditioned on the visual features extracted by the encoder and task-specific prompts (e.g., the question), the decoder generates the answer sequence autoregressively. This unified encoder-decoder backbone serves as the foundation upon which we apply our ChartMVC pretraining strategy.

B. The ChartMVC Framework

To enhance the semantic robustness of chart understanding models, we propose **ChartMVC**, a queue-based multi-view contrastive learning framework. As shown in Figure 2, the framework consists of three key components: multi-view representation generation, a shared encoder with a projection head, and a feature queue for contrastive optimization.

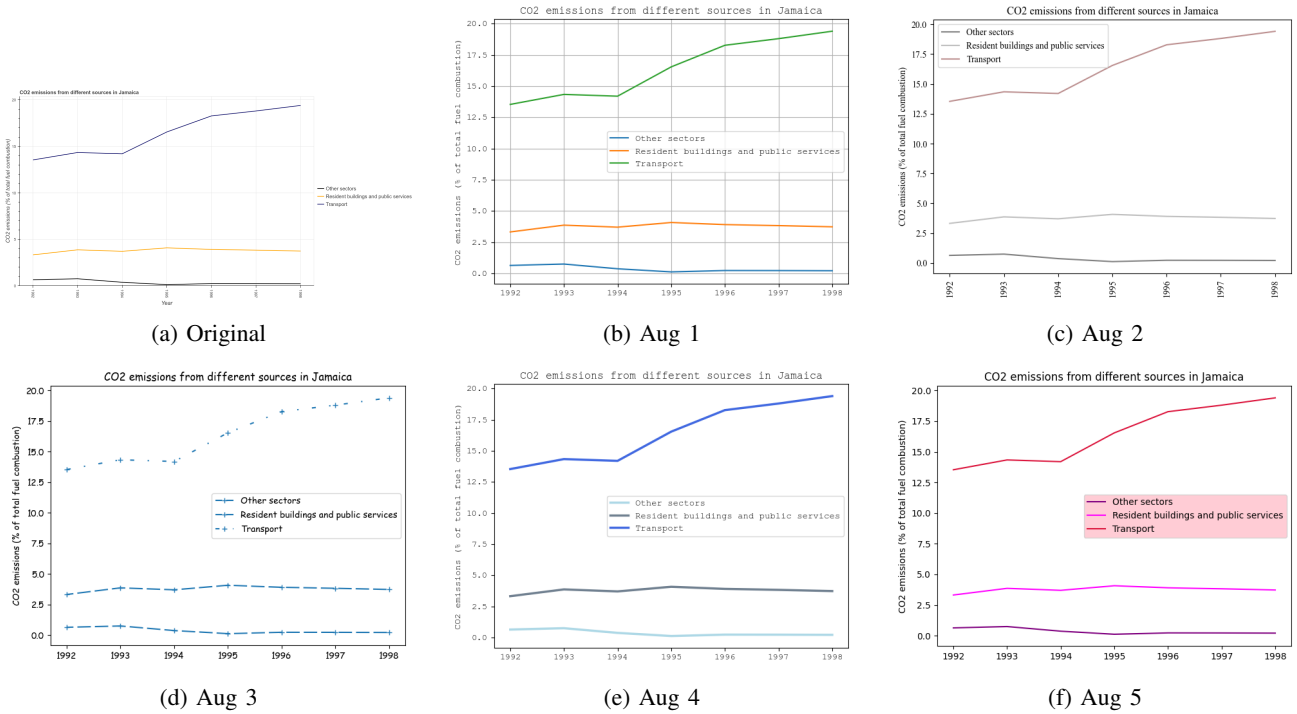
*1) **Multi-View Representation and Encoding**:* For each input chart x (**anchor**), we apply visual augmentations $\mathcal{T}$ (e.g., color, font, and layout shifts [24]) to generate a semantically equivalent **positive view** x^+, forming a positive pair (x, x^+) that shares identical data.

Both views are processed by a shared visual encoder $f(\cdot)$ and a lightweight MLP **projection head** $g(\cdot)$ to map the features into a metric space. The contrastive embeddings z_a and z_p are obtained as:

$$z_a = g(f(x)), \quad z_p = g(f(x^+)). \tag{1}$$

This design decouples representation learning from the downstream QA task, ensuring that the encoder learns robust, style-invariant features without interfering with the decoder's language generation process [23], [26].

*2) **Queue-based Contrastive Optimization**:* To stabilize training and provide a rich set of negative samples without requiring prohibitively large batch sizes, ChartMVC maintains a dynamic **feature queue**. This queue stores the encoded representations of charts from recent mini-batches.



(a) Original (b) Aug 1 (c) Aug 2

(d) Aug 3 (e) Aug 4 (f) Aug 5

Fig. 3: **Example of a multi-view group in MV-ChartQA.** The first image is the original chart, and the remaining images show augmented versions with different colors, fonts, and layouts.

TABLE I: **Comparison of CQA Datasets.** "Total Images" counts all original and augmented images. "Multi-view Groups" indicates whether multiple views per original chart exist.

Dataset	Total Images	Multi-view Groups	Benchmark	Chart Types
ChartQA	28,299	✗	✓	Bar, Line, Pie, Scatter
ChartQA-PoT	140,584	✗	✓	Bar, Line, Pie, Scatter
PlotQA	157,070	✗	✓	Bar, Line, Scatter, Others
DVQA	200,000	✗	✗	Bar, Line, Pie, Scatter
OpenCQA	5,407	✗	✓	Bar, Line, Scatter
MV-ChartQA (Ours)	$6 \times 18{,}000 \approx 108{,}000$	✓	✓	Bar, Line, Scatter, Others

For a given anchor-positive pair (z_a, z_p), the framework treats all K features currently stored in the queue as **negatives** $\{z_k\}_{k=1}^{K}$. We employ the **InfoNCE** loss [24], [27] to maximize the similarity between the positive pair while minimizing the similarity with negative samples:

$$\mathcal{L}_{CL} = -\log \frac{\exp(\text{sim}(z_a, z_p)/\tau)}{\sum_{k=1}^{K} \exp(\text{sim}(z_a, z_k)/\tau)}, \tag{2}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, τ is a temperature hyperparameter that controls the sharpness of the distribution, and K is the queue size. The numerator encourages alignment between the original and augmented charts, while the denominator pushes the anchor away from other semantically distinct charts in the queue.

*3) **Joint Training Objective**:* During pretraining, the model is optimized in a multi-task manner. We combine the proposed contrastive consistency loss with the standard Chart QA supervision loss (typically cross-entropy loss for text generation). The total objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{QA} + \lambda \mathcal{L}_{CL}, \tag{3}$$

where λ is a hyperparameter balancing the two tasks. By integrating $\mathcal{L}_{CL}$, ChartMVC acts as a regularization term, forcing the backbone model to learn semantic features that are robust to visual perturbations, thereby improving generalization on real-world data.

IV. Dataset Construction

To facilitate contrastive learning for robustness, we construct **MV-ChartQA**, a multi-view dataset derived from PlotQA [21]. We leverage PlotQA's underlying tabular data to programmatically generate charts, ensuring semantic consistency between visual variants and their QA pairs. MV-ChartQA organizes data into "view groups," totaling approximately 108,000 images across diverse types (e.g., bar, line, scatter). As illustrated in Figure. 3, each view group consists of one original anchor chart and five visually augmented variants.

Our augmentation strategy is designed to balance visual diversity with semantic preservation. We focus on *mild structural perturbations*—such as changes in color palettes, font styles, grid visibility, and legend positioning—while avoiding drastic chart type conversions (e.g., bar-to-line) that might alter the semantic focus. Specifically, for each view, we randomly apply 3–4 perturbations from a predefined set, strictly within the same chart type. This controlled generation ensures that the model learns style invariance without confusing the distinct structural semantics of different chart categories. Table I highlights MV-ChartQA's unique contribution: providing aligned multi-view groups essential for training robustness, a feature absent in existing datasets.

V. Experiment

A. Implementation Details

We instantiate ChartMVC upon the UniChart architecture (input resolution 960×960) and pretrain it on the MV-ChartQA dataset. The contrastive module employs a two-layer MLP projection head with GELU activation to map encoder features into the embedding space, utilizing a dynamic feature queue of size 2048 for negative sampling. During training, we adopt an intra-group sampling strategy where one positive view is randomly selected from the augmented variants for each anchor. The model is trained for 3 epochs with a batch size of 20 and a learning rate of 5×10^{-5}. The contrastive loss temperature τ and the auxiliary loss weight λ are both set to 0.1.

B. Main Results

We evaluate our framework on two complementary benchmarks to assess both general reasoning capabilities and visual robustness. First, ChartQA [2] serves as the standard testbed, focusing on complex numerical reasoning and logic. Following prior works [4], [6], we report Relaxed Accuracy (RA), which tolerates numerical errors within 5%. Second, to rigorously test the model's resilience, we employ RobustCQA [10]. Built upon ChartQA, this benchmark systematically applies 75 unique perturbation types (grouped into ≈20 styles like color schemes, layout changes, and axis formatting) while preserving underlying data. Crucially, it categorizes samples into four difficulty levels (e.g., simple-simple to complex-complex) based on chart and question complexity, enabling a fine-grained evaluation of how visual perturbations impact performance across different cognitive loads.

On the standard ChartQA benchmark, our approach demonstrates strong competitiveness while maintaining a compact footprint (202M parameters). Without relying on auxiliary modules or multi-stage training pipelines, ChartMVC achieves a +2.08% accuracy improvement on human-annotated questions compared to the UniChart baseline. It also consistently outperforms or matches larger state-of-the-art models, including Pix2Struct [4] and MatCha [8]. This indicates that our contrastive pretraining does not compromise general reasoning ability; rather, by exposing the model to diverse visual variants, it enhances generalization across heterogeneous chart representations found in real-world data. The advantages of our framework are most evident on the RobustCQA benchmark. Following MV-ChartQA pretraining, the model achieves a substantial average accuracy increase of approximately 10 percentage points across diverse perturbation types. This confirms that our method effectively decouples visual style from data semantics, significantly improving stability against visual noise and stylistic inconsistencies prevalent in practical applications. While ChartMVC demonstrates strong gains across most scenarios, its impact on extremely challenging cases remains limited, suggesting that fundamental reasoning complexities persist beyond purely visual robustness.

TABLE II: **Performance Comparison on ChartQA Dataset**

Model	Size	Aug. Acc.	Human Acc.	Avg. Acc.
T5	223M	41.0	25.1	33.05
Chart-T5	400M	74.4	31.8	53.10
Donut	260M	78.1	29.8	53.95
UniChart	201M	88.56	43.92	66.24
Pix2Struct	282M	81.6	30.5	56.05
MatCha	282M	90.2	38.2	64.20
Ours	202M	88.56	**46.0(+2.08)**	**67.28(+1.04)**

TABLE III: **Average Accuracy (%) of RobustCQA under Different Perturbations**

Perturbation	Baseline	Ours
annotations	24.00	**29.50 (+5.50)**
area_plot	27.50	**42.00 (+14.50)**
benchmark	29.50	**37.25 (+7.75)**
color_random	25.75	**33.75 (+8.00)**
color_scheme	25.50	**32.25 (+6.75)**
data_pivot	18.75	18.00 (-0.75)
font	21.25	**31.25 (+10.00)**
grid	28.75	**32.50 (+3.75)**
hatching	19.00	**22.75 (+3.75)**
horizontal	25.50	**40.50 (+15.00)**
horizontal_grouped	21.50	**33.50 (+12.00)**
horizontal_stacked	11.50	9.50 (-2.00)
legend_position	26.75	**34.00 (+7.25)**
line_representation	27.75	**34.75 (+7.00)**
log_scale	10.25	9.50 (-0.75)
normal	28.00	**38.00 (+10.00)**
only_data_color_scheme	21.50	**28.50 (+7.00)**
replacing_legend_with_labels	27.00	**34.67 (+7.67)**
scaling	22.25	**30.75 (+8.50)**
scatter_representation	22.00	**32.25 (+10.25)**
stacked	17.50	16.00 (-1.50)
stacked_area	12.00	**13.50 (+1.50)**
stair_plot_normal	24.25	**25.25 (+1.00)**
stair_plot_with_marker	22.25	**27.50 (+5.25)**
stem_plot	22.00	**32.00 (+10.00)**
tick_orientation	22.75	**33.75 (+11.00)**
tick_position	20.50	16.25 (-4.25)

C. Ablation Study

To evaluate the individual effects of multi-view training and contrastive learning, we conducted an ablation study on the ChartQA dataset. Specifically, we designed four configurations: (1) the baseline UniChart model without contrastive learning, (2) a multi-view variant without contrastive learning,

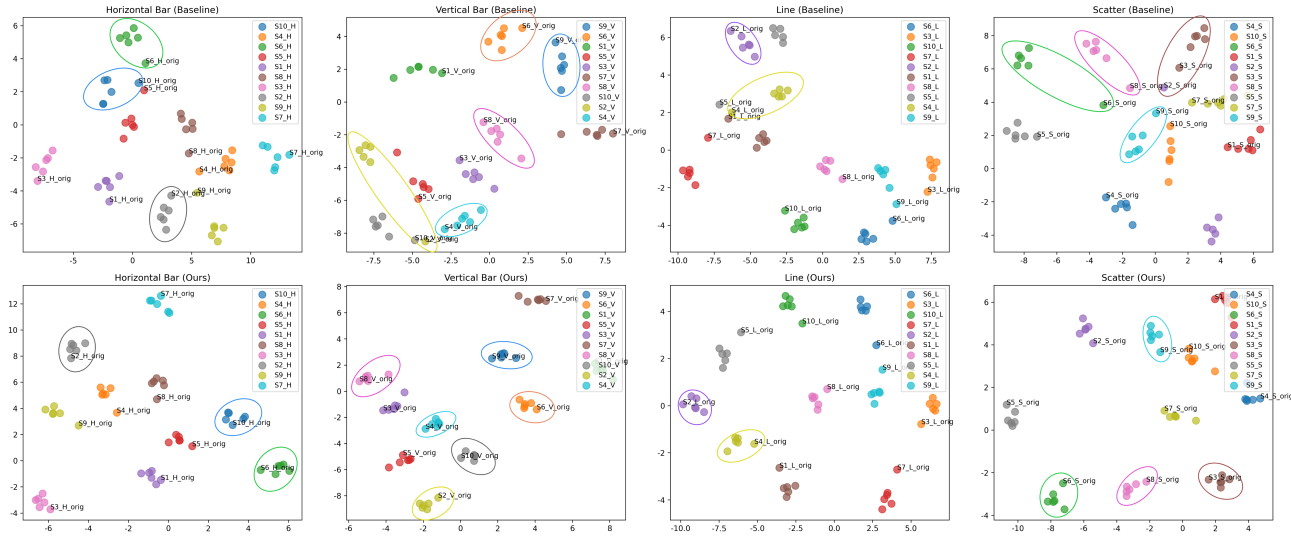


Fig. 4: **t-SNE Visualization Comparison.** The top row corresponds to the Base model, and the bottom row to Ours ChartMVC. From left to right: Horizontal Bar, Vertical Bar, Line, and Scatter chart types.

(3) a single-view contrastive variant, and (4) our proposed multi-view contrastive learning model. As shown in Table IV, both multi-view training and contrastive learning contribute positively to model performance. The introduction of multiple visual views improves the model's generalization across diverse chart appearances, while contrastive learning further enhances semantic consistency. Combining the two yields the best results, demonstrating that multi-view representation and contrastive objectives are complementary in improving robustness on ChartQA.

TABLE IV: **Ablation Study on ChartQA Dataset**

Model	Aug. Acc.	Human Acc.	Avg. Acc.
Baseline (UniChart)	88.56	43.92	66.24
Single-view	88.56	45.20	66.88
Multi-view without Contrastive	88.48	45.00	66.74
Multi-view with Contrastive	88.56	**46.00**	**67.28**

D. Similarity Analysis

To evaluate representational robustness, we visualized chart embeddings via t-SNE (Fig. 4). Compared to the baseline, ChartMVC forms significantly tighter clusters across all chart types, effectively pulling visually diverse variants into compact groups. Notably, for samples like S_2 in the horizontal bar chart and S_6 in the scatter plot, the baseline disperses their augmented versions across the embedding space due to style sensitivity. In contrast, ChartMVC maps these semantically identical variants to consistent latent representations while maintaining clear separation between different groups, demonstrating superior style-invariance.

Quantitative analysis further supports this. We computed pairwise cosine similarity between embeddings within view groups. As illustrated in Fig. 5, the resulting heatmap reveals that ChartMVC achieves an average intra-group similarity exceeding 93%, significantly surpassing the baseline. These results confirm that our contrastive framework robustly encodes charts into a structured embedding space, enhancing intra-class compactness (invariance to style) and inter-class discriminability.

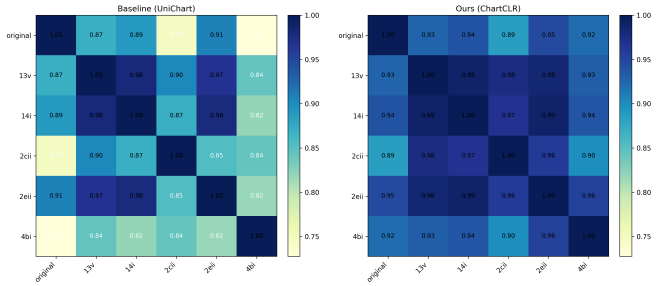


Fig. 5: **Feature Similarity Comparison.** Average pairwise cosine similarity heatmap for ChartMVC versus Baseline. Each cell represents the average cosine similarity between different views (orig, aug1-aug5) across multiple samples. High diagonal values and within-group similarity demonstrate superior style invariance.

VI. Conclusion

In this paper, we introduce **ChartMVC**, a generalizable multi-view contrastive learning framework designed to enhance the robustness of chart understanding. By leveraging a queue-based consistency mechanism and our large-scale MV-ChartQA dataset, the framework effectively decouples visual styles from intrinsic data semantics to learn perturbation-invariant representations. Extensive experiments on ChartQA and RobustCQA benchmarks demonstrate that ChartMVC significantly boosts the accuracy and stability of backbone models against realistic visual disturbances.

Crucially, ChartMVC serves as a model-agnostic, plug-and-play solution that can be seamlessly integrated into various vision-language architectures. Our findings highlight that decoupling semantics from style is a critical step towards robust chart understanding, offering a scalable path to enhance the reliability of multimodal models in complex real-world scenarios.

REFERENCES

[1] Enamul Hoque, Parsa Kavehzadeh, and Ahmed Masry. 2022. Chart question answering: State of the art and future directions. *Computer Graphics Forum (Proc. EuroVis)*, pages 555–572.

[2] A. Masry, D. Long, J. Q. Tan, S. Joty, and E. Hoque, "ChartQA: A benchmark for question answering about charts with visual and logical reasoning," in *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, Dublin, Ireland, 2022.

[3] S. Kantharaj, X. L. Do, R. T. K. Leong, J. Q. Tan, E. Hoque, and S. Joty, "OpenCQA: Open-ended question answering with charts," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

[4] K. Lee, M. Joshi, I. Turc, H. Hu, F. Liu, J. Eisenschlos, U. Khandelwal, P. Shaw, M.-W. Chang, and K. Toutanova, "Pix2Struct: Screenshot parsing as pretraining for visual language understanding," *arXiv preprint arXiv:2210.03347*, 2022.

[5] J. Obeid and E. Hoque, "Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model," in *Proceedings of the 13th International Conference on Natural Language Generation (INLG)*, pp. 138–147, Dublin, Ireland, 2020.

[6] S. Kantharaj, R. T. K. Leong, X. Lin, A. Masry, M. Thakkar, E. Hoque, and S. Joty, "Chart-to-text: A large-scale benchmark for chart summarization," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

[7] D. Shi, X. Xu, F. Sun, Y. Shi, and N. Cao, "Calliope: Automatic visual data story generation from a spreadsheet," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 453–463, 2020.

[8] F. Liu, F. Piccinno, S. Krichene, C. Pang, K. Lee, M. Joshi, Y. Altun, N. Collier, and J. M. Eisenschlos, "Matcha: Enhancing visual language pretraining with math reasoning and chart derendering," *arXiv preprint arXiv:2212.09662*, 2022.

[9] A. Masry, P. Kavehzadeh, X. L. Do, E. Hoque and S. Joty, "UniChart: A Universal Vision-language Pretrained Model for Chart Comprehension and Reasoning," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Singapore, 2023, pp. 14662–14684.

[10] E. Hoque, P. Kavehzadeh, and A. Masry, "RobustCQA: A benchmark for style-robust chart question answering," in *Proceedings of . . .* , 2024.

[11] Y. Wu, L. Yan, L. Shen, Y. Wang, N. Tang, and Y. Luo, "ChartInsights: Evaluating multimodal large-language models for low-level chart question answering," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.

[12] A. Masry, M. Thakkar, A. Bajaj, A. Kartha, E. Hoque, and S. Joty, "ChartGemma: Visual Instruction-tuning for Chart Reasoning in the Wild," *arXiv preprint arXiv:2407.04172*, 2024.

[13] Y. Han *et al.*, "A Multimodal LLM for Chart Understanding and Generation," *2025 International Joint Conference on Neural Networks (IJCNN)*, Rome, Italy, 2025, pp. 1–8.

[14] L. Zhang, A. Hu, H. Xu, M. Yan, Y. Xu, Q. Jin, J. Zhang, and F. Huang, "TinyChart: Efficient Chart Understanding with Visual Token Merging and Program-of-Thoughts Learning," *arXiv preprint arXiv:2404.16635*, 2024. [Online]. Available: https://arxiv.org/abs/2404.16635

[15] J. Ma, P. Wang, D. Kong, Z. Wang, J. Liu, H. Pei, and J. Zhao, "Robust visual question answering: Datasets, methods, and future challenges," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5575–5594, 2024.

[16] Y. Zhao, T. Pang, C. Du, X. Yang, C. Li, N.-M. Cheung, and M. Lin, "On evaluating adversarial robustness of large vision-language models," in *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[17] A. Gupta, V. Gupta, S. Zhang, Y. He, N. Zhang, and S. Shah, "Enhancing question answering on charts through effective pretraining tasks," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[18] G. Kim, T. Hong, M. Yim, J. Nam, J. Park, J. Yim, W. Hwang, S. Yun, D. Han, and S. Park, "OCR-free document understanding transformer," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 498–517, Springer, 2022.

[19] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021.

[20] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *CoRR*, vol. abs/1910.13461, 2019.

[21] N. Methani, P. Ganguly, M. M. Khapra, and P. Kumar, "PlotQA: Reasoning over scientific plots," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020.

[22] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 776–794, 2020.

[23] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1597–1607, 2020.

[24] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 9640–9649, 2021.

[25] C. Peng, X. Long, Y. Li, "MSVQ: Self-Supervised Learning with Multiple Sample Views and Queues," arXiv preprint arXiv:2305.05370, 2023.

[26] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[27] H. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," arXiv preprint arXiv:1807.03748, 2018.

[28] A. Masry, D. Long, J. Q. Tan, S. Joty, and E. Hoque, "ChartQA: A benchmark for question answering about charts with visual and logical reasoning," in *Findings of the Association for Computational Linguistics: ACL*, Dublin, Ireland, 2022, pp. 2263–2279.

[29] J. Cho, J. Lei, H. Tan, and M. Bansal, "Unifying vision-and-language tasks via text generation," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

[30] M. Zhou, Y. R. Fung, L. Chen, C. Thomas, H. Ji, and S.-F. Chang, "Enhanced chart understanding via visual language pre-training on plot table pairs," in *Findings of the Association for Computational Linguistics: ACL*, Toronto, Canada, 2023, pp. 1314–1326.

[31] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[32] W. Chen, X. Ma, X. Wang, and W. W. Cohen, "Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks," *arXiv preprint arXiv:2211.12588*, 2022.

[33] A. Masry, M. Shahmohammadi, M. R. Parvez, E. Hoque, and S. Joty, "ChartInstruct: Instruction tuning for chart comprehension and reasoning," *arXiv preprint arXiv:2403.09028*, 2024.

[34] F. Meng, W. Shao, Q. Lu, P. Gao, K. Zhang, Y. Qiao, and P. Luo, "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*, pp. 7775–7803, 2024.

[35] Z. Wang, M. Xia, L. He, H. Chen, Y. Liu, R. Zhu, K. Liang, X. Wu, H. Liu, S. Malladi, A. Chevalier, S. Arora, and D. Chen, "CharXiv: Charting gaps in realistic chart understanding in multimodal LLMs," *arXiv preprint arXiv:2406.18521*, 2024.

[36] F. Liu, X. Wang, W. Yao, J. Chen, K. Song, S. Cho, Y. Yacoob, and D. Yu, "MMC: Advancing multimodal chart understanding with large-scale instruction tuning," *arXiv preprint arXiv:2311.10774*, 2023.

[37] R. Xia, B. Zhang, H. Ye, X. Yan, Q. Liu, H. Zhou, Z. Chen, M. Dou, B. Shi, J. Yan, and Y. Qiao, "ChartX & ChartVLM: A versatile benchmark and foundation model for complicated chart reasoning," *arXiv preprint arXiv:2402.12185*, 2024.

[38] A. Bansal, D. Soselia, D. Nguyen, and T. Zhou, "ChartAB: A Benchmark for Chart Grounding & Dense Alignment," *arXiv preprint arXiv:2501.07684*, 2025.