Boosting Text-to-Chart Retrieval through Synthesized Hierarchical Insights

Anonymous ACL submission

Abstract

Text-to-chart retrieval enables users to locate relevant charts through natural language queries and has emerged as a critical capability in modern business intelligence systems. However, current benchmarks fail to simulate realistic user queries or test deep semantic understanding, leaving a significant evaluation gap. To address this, we introduce **CRBench**, the first real-world BI-sourced benchmark comprising 21,862 charts and 326 queries, which adopts a Target-and-Distractor paradigm to evaluate discriminative retrieval among highly similar candidates. Benchmarking reveals that existing methods, which rely primarily on visual features, fail to capture the rich analytical semantics of charts. We therefore propose a semantic insights synthesis pipeline that automatically generates three hierarchical levels of insights for alignment training: visual patterns, statistical properties, and practical applications. Using this pipeline, we produce 207,498 insights for 69,166 charts as training data. Leveraging this data, we develop **ChartFinder**, which significantly outperforms state-of-the-art methods on CRBench, achieving 66.9% NDCG@10 for precise queries (an 11.58% improvement) and an average gain of 5% across fuzzy query metrics.

1 Introduction

As business intelligence platforms accumulate thousands of charts, the ability to retrieve the right chart through natural language has become a fundamental yet underexplored challenge (Wu et al., 2024; Luo et al., 2018a; Shen et al., 2022; Qin et al., 2020). Consequently, text-to-chart retrieval systems that allow users to find relevant charts based on natural language queries (*i.e.,* text queries) have become increasingly valuable, attracting significant attention from both the database community (Ji et al., 2024b,a; Luo et al., 2023) and industry practitioners (Setlur et al., 2023). In real-world BI scenarios, queries range from *precise queries* targeting specific analytical information (*e.g.,* "Find charts showing hotel booking distribution in Europe for 2022") to *fuzzy queries* for exploratory analysis (*e.g.,* "Use a pie chart to compare sale methods") (Ji et al., 2024a,b). Both require a deep understanding of the rich analytical semantics embedded in charts, such as data relationships and statistical trends (Qin et al., 2020). Yet bridging the gap between natural language query intent and such visual semantics is inherently difficult. This challenge is exacerbated when charts exist only as static images without accessible metadata, a common scenario where visual-based methods struggle (Luo et al., 2023; Ji et al., 2024a; Li et al., 2022a).

A core obstacle impeding progress in this field is the lack of a dedicated benchmark for realistic evaluation. Existing chart-related datasets are primarily designed for tasks like captioning and fail to simulate real-world retrieval challenges. They lack concise, intent-driven queries and a large repository of visually and semantically similar charts to test a model's true discriminative power. To address this critical evaluation gap, we first curate and introduce CRBench, the first benchmark for text-to-chart retrieval sourced from real-world BI scenarios. CRBench contains 326 text queries and 21,862 chart images, with all query-chart labels verified by crowd workers to capture the diversity of real user needs.

Benchmarking existing models, including powerful Multimodal Large Language Models (MLLMs) (Liu et al., 2022; Xiao et al., 2024b; Zhou et al., 2024b; Zhang et al., 2024), on CRBench reveals their significant limitations. These models struggle to capture the deep analytical semantics required to differentiate between challenging charts. To address this performance bottleneck, we argue that the key lies in providing higher-quality training data that explicitly teaches models to bridge the gap between visual features and their analytical meaning. To this end, we propose our

second major contribution: an automatic semantic insights synthesis pipeline. Inspired by automatic data visualization (Luo et al., 2018a) and insight generation methods (Ma et al., 2021), this pipeline automatically synthesizes three hierarchical levels of insights: *(1) Visual-oriented*, *(2) Statistics-oriented*, and *(3) Task-oriented*.

To validate the challenge of our benchmark and the effectiveness of our data pipeline, we use our synthesized data to fine-tune several CLIP-based models. We name the model resulting from fine-tuning Long-CLIP (Zhang et al., 2025) as ChartFinder. The core idea is to leverage the synthesized insights exclusively during the training stage to learn robust visual-semantic representations. During retrieval, the model relies solely on chart images.

Contributions. We make the following contributions.

(1) **CRBench Benchmark.** We curate CRBench, *the first* benchmark for evaluating text-to-chart retrieval sourced from *real-world BI scenarios*. It includes 326 text queries and 21,862 charts, with query-chart labels verified by crowd workers to capture the diversity of real user needs (Section 3).

(2) **A Semantic Insights Synthesis Pipeline for Charts.** We develop an automatic pipeline for generating semantic insights from chart metadata, synthesizing three levels of semantic insights. This approach addresses the challenge of manual annotation by automating the process, enabling the creation of diverse insights that enrich the model's ability to comprehensively understand charts (Section 4).

(3) **A High-Performance Model Validating the Pipeline.** By fine-tuning the CLIP architecture with our synthesized data, we produce ChartFinder. It not only validates the effectiveness of our data synthesis pipeline by establishing a new state-of-the-art but also serves as a strong baseline for future research on CRBench.

(4) **Extensive Experiments.** (4) **Extensive Experiments.** Extensive experiments validate both the challenge of CRBench and the effectiveness of the proposed semantic insight synthesis pipeline (Section 5).

2 Related Work

Chart Retrieval. Chart retrieval aims to find relevant charts from a repository, typically using a chart, sketch, or underlying data as the query. Existing methods mainly fall into two paradigms. Perception-based approaches compare visual style, chart structure, deep visual embeddings, or hand-drawn trend sketches (Saleh et al., 2015; Hoque and Agrawala, 2019; Luo et al., 2023; Mannino and Abouzied, 2018). Data-based approaches instead assume access to the underlying data and retrieve charts by matching data patterns or series representations (Wang and Palpanas, 2021; Lekschas et al., 2020; Siddiqui et al., 2016). These studies are related to our setting because they require understanding chart structure and visual patterns. However, they do not address the key challenge of text-to-chart retrieval: aligning concise natural language intent with the analytical semantics of static chart images, especially when metadata or source data is unavailable.

Text-to-Chart Retrieval. Prior work related to text-to-chart retrieval includes visualization recommendation systems that guide user exploration (Vartak et al., 2017; Ye et al., 2024; Qin et al., 2018b,a; Bendeck et al., 2024; Srinivasan and Setlur, 2021) and natural language interfaces for visualization or dashboard authoring (Shen et al., 2022; Tang et al., 2022; Luo et al., 2018b, 2022b, 2021; Setlur et al., 2023; Srinivasan and Setlur, 2023). While these systems have made substantial progress in supporting visualization exploration and generation, they are not primarily designed for discriminative retrieval over large repositories of visually and semantically similar chart images. Recent Multimodal Large Language Models (MLLMs) (Li et al., 2022b, 2023; Yuan et al., 2021; Xiao et al., 2024a) and CLIP-style models (Radford et al., 2021) provide a general cross-modal retrieval framework. Several variants further improve visual encoders, text encoders, or retrieval-oriented pretraining, including EVA-CLIP (Sun et al., 2023; Fang et al., 2024), Jina-CLIP-v2 (Xiao et al., 2024b; Koukounas et al., 2024; Sturua et al., 2024), UniVL-DR (Liu et al., 2022), MagicLens (Zhang et al., 2024), VISTA (Zhou et al., 2024a), and MARVEL (Zhou et al., 2024b). However, these models are mostly trained on generic image-text data and are not explicitly optimized for chart-specific analytical semantics such as trends, distributions, cor-

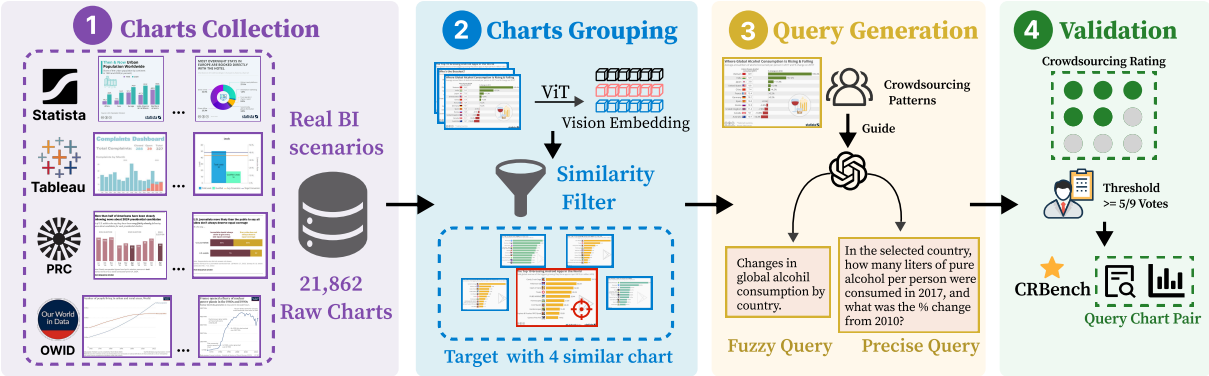


Figure 1: The construction workflow of the CRBench.

relations, and task-level usage. Our work addresses this gap by synthesizing hierarchical chart insights and using them as supervision for contrastive text-to-chart retrieval.

3 CRBench Benchmark

This section introduces CRBench, a benchmark designed to evaluate text-to-chart retrieval under realistic and discriminative retrieval settings. We first describe the motivation and design principles, then present the benchmark construction process and compare CRBench with existing chart captioning datasets repurposed for retrieval.

3.1 Motivation and Design

Existing chart-related datasets are not well suited for evaluating text-to-chart retrieval. Datasets such as *VisText* (Tang et al., 2023) and *Chart-to-Text* (Kantharaj et al., 2022) were originally designed for chart captioning rather than retrieval. Their textual inputs are usually long descriptive captions, while real BI users often issue concise and intent-driven queries. Moreover, these datasets have low chart-to-query ratios, which fail to simulate a realistic retrieval scenario where a query must be matched against a large repository containing many visually and semantically similar charts.

CRBench is designed to fill this gap. It focuses on three requirements: (1) supporting both precise queries that specify concrete analytical details and fuzzy queries that express broader user intent, (2) maintaining a strict distributional separation between training data and benchmark data, and (3) ensuring that each query has a human-validated dominant target chart among hard distractors. This design emphasizes fine-grained discriminative retrieval, where models must distinguish charts that are visually similar but analytically different.

3.2 Benchmark Construction

As shown in Figure 1, CRBench is constructed through four stages: chart collection, similarity-based grouping, query generation, and human validation.

Chart Collection. We collect charts from four real-world BI and data analysis sources: *Tableau* (Tableau, 2003), *Statista* (statista, 2007), *Pew Research Center* (Research, 2004), and *Our World in Data* (OWID, 2011). In total, CRBench contains **21,862 charts** covering diverse domains such as finance, business, e-commerce, technology, health, and society.

Similarity-based Grouping. To construct hard retrieval cases, we adopt a target-and-distractor paradigm. We embed all charts into 768-dimensional vectors using a pre-trained *ViT* and compute pairwise cosine similarities. For each candidate chart, we retrieve its four most visually similar charts and form a group only when all four similarities are at least 0.90. The anchor chart is treated as the target, while the other four charts serve as hard distractors. This process produces 247 groups, each containing one target chart and four visually similar distractors.

Query Generation. For each target-distractor group, we generate queries that are both realistic and discriminative. We first conduct a pilot crowdsourcing study on Appen[1] and observe that human-written queries often become either too generic or too close to the chart title. Based on this observation, we use GPT-4o to generate one precise query and one fuzzy query per target chart, while requiring the query to include discriminative anchors such as axis labels, concrete values, or time ranges and avoid directly copying chart text.

[1] https://appen.com

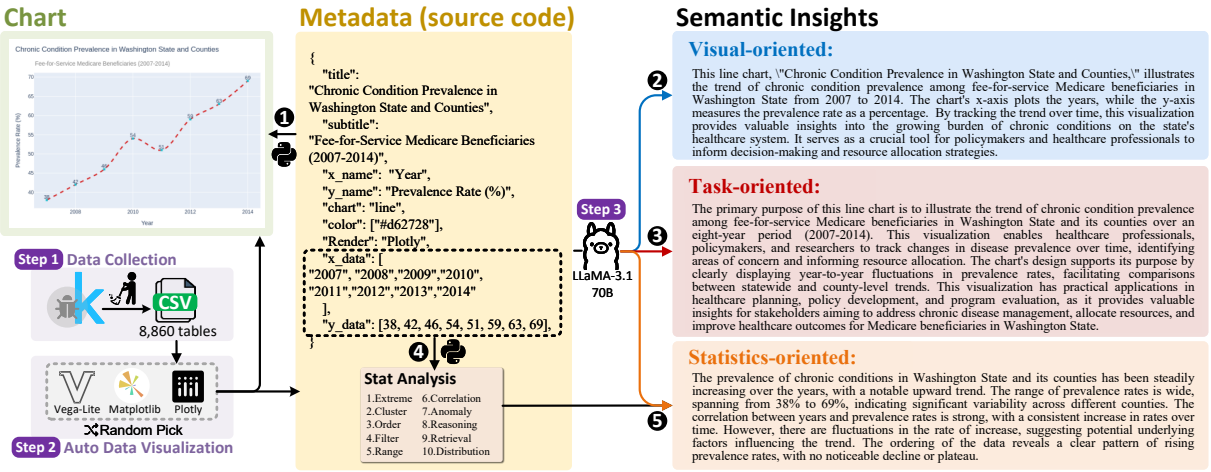


Figure 2: An overview of the semantic insights synthesis pipeline.

Human Validation. Each generated query is paired with its five candidate charts, including one target and four distractors, and evaluated by nine independent crowdworkers. A query is retained only if at least five workers select the same target chart. This consensus criterion serves as both a quality filter and evidence that the retained query has a dominant ground truth. After validation, CRBench contains **195 precise queries** and **131 fuzzy queries**.

3.3 Benchmark Overview

Table 1 compares CRBench with *Chart-to-Text* (Kantharaj et al., 2022) and *VisText* (Tang et al., 2023). Compared with these captioning datasets, CRBench has three distinctive properties. First, its queries are concise and intent-driven, with both precise and fuzzy query types, rather than long chart captions. Second, it provides a much larger retrieval repository, with a chart-to-query ratio of 67:1. Third, its target-and-distractor construction creates visually similar candidate charts, making the benchmark more suitable for evaluating fine-grained text-to-chart retrieval. Although the queries are generated with GPT-4o, they are validated through multi-worker consensus, and only query-chart pairs with a dominant human-agreed target are retained.

4 Semantic Insights Synthesis Pipeline

To train models for text-to-chart retrieval, we need supervision that connects chart images with their analytical semantics. Manual annotation is expensive and difficult to scale. We therefore develop an automatic synthesis pipeline that generates hierarchical semantic insights from chart metadata. As shown in Figure 2, the pipeline first constructs a synthetic chart corpus from tabular data, then derives three types of semantic insights for each chart: visual-oriented, statistics-oriented, and task-oriented insights.

4.1 Synthetic Chart Corpus

We collect high-quality tabular datasets from Kaggle to construct the training corpus. Kaggle datasets are selected because they are mostly derived from real applications or competitions and provide quality scores for filtering. We crawl 9,003 datasets with a score of 10 and remove tables with missing values, resulting in **8,860 CSV tables**.

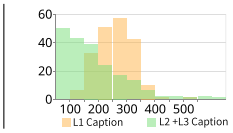
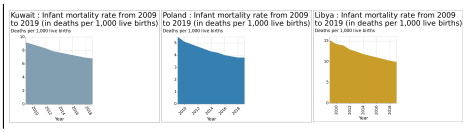
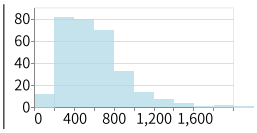
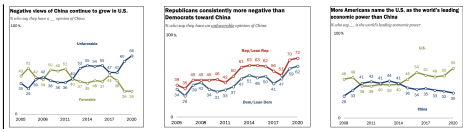
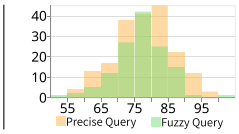
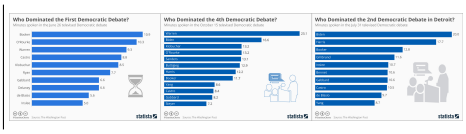
Given these tables, we use DeepEye (Luo et al., 2018a), an automatic data visualization system, to generate charts covering common visualization types, including bar charts, pie charts, line charts, scatter plots, grouped line charts, stacked bar charts, and grouped bar charts (Luo et al., 2022a). To improve visual diversity, we re-render the generated charts using Matplotlib, Plotly, and Vega-Lite, and randomize visual parameters such as line type, marker shape, and chart sub-type. For each chart, we also store metadata produced by DeepEye, including chart type, title, axis labels, and data points. This process produces **69,166 charts** with corresponding metadata.

4.2 Hierarchical Insight Synthesis

Based on chart metadata, we synthesize three levels of semantic insights. These insights are designed to cover different levels of chart understanding and support different forms of user queries.

Visual-oriented insights. These insights summa-

Table 1: Comparison of Existing Benchmarks. The statistics for Chart-to-Text and VisText are from the respective test sets. The column "*C/Q*" indicates the ratio for charts (C) and text queries (Q). The column "*Dist. of Query Len.*" shows the distribution of text query lengths in characters. The column "*CT*" denotes chart types.

Datasets	CT	#-Charts	#-Queries	C/Q	Dist. of Query Len.	Samples of Similar Charts
VisText	3	882	1764	0.5		
Chart-to-Text	6	1393	1393	1		
CRBench	11	21,862	326	67		

rize the directly observable visual patterns of a chart, such as chart type, axes, major trends, distributions, comparisons, and salient visual structures.

Statistics-oriented insights. These insights describe data-driven analytical properties. We first apply 10 statistical analysis tasks (Wu et al., 2024) to the chart metadata, including analyses of extremes, clusters, ordering, filtering, ranges, correlations, anomalies, reasoning, retrieval, and distributions. We then summarize the resulting statistical signals into natural language insights.

Task-oriented insights. These insights describe how a chart can support real-world analytical tasks, such as decision-making, monitoring, comparison, diagnosis, or resource allocation. They are generated from both chart metadata and the previously synthesized visual insights, allowing the model to associate chart content with practical user intent.

In total, the pipeline generates **207,498 semantic insights** for **69,166 charts**. These insights provide scalable supervision for learning chart-specific visual-semantic representations. We further conduct a human evaluation on a sampled subset of synthesized insights, where crowdworkers rate them on completeness, consistency, diversity, and readability. The results show that more than 74% of ratings are positive across all four dimensions. We report the full quality evaluation in Appendix **??**.

4.3 Training with Synthesized Insights

We use the synthesized insights as image-text supervision for contrastive learning. As shown in Figure 3(a), each chart image is paired with its visual-oriented, statistics-oriented, and task-oriented insights. During training, these insights are treated as textual descriptions aligned with the same chart image, encouraging the model to learn chart representations that capture not only visual but also statistical patterns and task-level semantics.

At inference time, as shown in Figure 3(b), no chart metadata or synthesized insight is required. Given a user query and a repository of chart images, ChartFinder encodes the query and chart images into a shared embedding space and retrieves charts according to embedding similarity. This design allows the model to benefit from rich metadata-derived supervision during training while remaining applicable to static chart images at retrieval.

5 Experiments

In this section, we conduct a set of experiments to evaluate whether the synthesized semantic insights can enhance text-to-chart retrieval performance across multiple benchmarks and settings. We also assess whether the ChartFinder model achieves state-of-the-art performance against all baselines. The following experiments are performed:

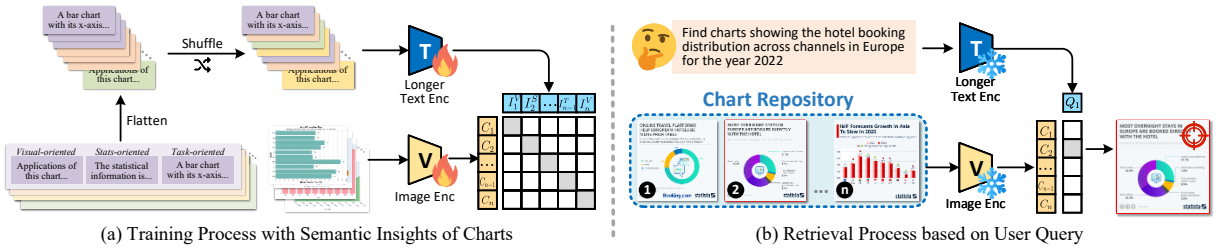
Exp-1: How does ChartFinder perform compared with cross-modal retrieval baselines on CRBench?

Exp-2: Can text-index or caption-index retrieval replace visual retrieval?

Exp-3: How does each type of synthesized semantic insight contribute to retrieval performance?

Exp-4: Can synthesized semantic insights improve different retrieval backbones?

Exp-5: Does ChartFinder generalize to external



(a) Training Process with Semantic Insights of Charts

(b) Retrieval Process based on User Query

Figure 3: (a) ChartFinder is trained using chart images and their synthesized semantic insights. (b) During retrieval, ChartFinder takes a user query and chart images as input, and ranks charts by embedding similarity.

chart-text benchmarks?

Exp-6: What is the impact of image preprocessing strategies on chart retrieval?

In the main body of the paper, we focus on Exp-1 to Exp-4. The results of Exp-5 and Exp-6 are reported in Appendix A, where we provide zero-shot results on external chart-text benchmarks and an ablation study on center crop versus direct resize. Full metric tables and additional analyses are also provided in the appendix.

Metrics. To evaluate the effectiveness of the model in text-to-chart retrieval, we use the following metrics: Recall@1, Recall@5, Recall@10, MRR@10, and NDCG@10. These metrics are widely adopted in information retrieval tasks (Qin et al., 2020)

Compared Methods. For Exp-1, we compare ChartFinder with representative cross-modal retrieval models, including VISTA (Zhou et al., 2024a), UniVL-DR (Liu et al., 2022), MARVEL (Zhou et al., 2024b), CLIP (Radford et al., 2021), CLIP-DPR (Liu et al., 2022), EVA-CLIP (Sun et al., 2023), CoCa (Cherti et al., 2023), Jina-CLIP-v2 (Koukounas et al., 2024), and Long-CLIP (Zhang et al., 2025). We report two versions of our model: ChartFinder-B, which uses ViT-B/16 as the visual encoder, and ChartFinder, which uses ViT-L/14.

For Exp-2, we additionally evaluate text-index retrieval baselines. Each chart is converted into either a generated title or a generated caption, and retrieval is performed using BM25, BGE, E5, and CLIP text encoders. These baselines test whether text-only retrieval can replace direct visual retrieval. For Exp-4, we further fine-tune CLIP-DPR and UniVL-DR with the same synthesized insight data to examine whether the proposed supervision is transferable across retrieval backbones.

Training Details. We trained ChartFinder using 4 A800 GPUs with a learning rate of $1e-6$ for 20 epochs. For the base version of the model, we set the batch size to 512, while for the larger version, we reduced the batch size to 256 due to the increased memory usage associated.

5.1 Exp-1: Overall Results on CRBench

Experiment Settings. The purpose of this experiment is to evaluate and compare the performance of different models in precise and fuzzy query tasks.

Overall Results. Experimental results are shown in Table 2. We can observe that ChartFinder and ChartFinder-B perform best in both query tasks, and multiple metrics are ahead of other methods by more than 5% or even 10%. For example, in the precise query task, ChartFinder's R@1, R@5, and R@10 are 47.18%, 80.00%, and 84.10%, respectively, significantly higher than the second-place Long-CLIP-L's 41.03%, 75.90%, and 79.49%. In the fuzzy query task, ChartFinder also surpasses Long-CLIP-L, with R@1, R@5, and R@10 being 3.05%, 8.39%, and 5.34% higher, respectively. In comparison, although Jina-CLIP-v2 (865M) has the largest number of parameters, it performs poorly in both types of tasks. The R@1 in the precise query task is only 4.10%, and the R@1 in the fuzzy query task is also only 3.05%.

5.2 Exp-2: Text-Index and Caption-Index Retrieval

Experiment Settings. We compare ChartFinder with a practical text-only alternative that first converts each chart into a textual index and then performs text-to-text retrieval. Specifically, we use Gemini-2.5-Flash to generate either a canonical title (*Title-index*) or a compact 70-token caption (*Caption-index*) for each chart in CRBench, without using any ground-truth query or relevance label. We then evaluate BM25, BGE, E5, and CLIP text encoders, and compare them with cross-modal re-

Table 2: Performance Comparison of Text-to-Chart Retrieval Models on CRBench Benchmark.

Models	#Param	Precise Query					Fuzzy Query				
		R@1	R@5	R@10	MRR@10	NDCG@10	R@1	R@5	R@10	MRR@10	NDCG@10
VISTA (Zhou et al., 2024a)	196M	8.21	20.00	27.18	13.15	16.42	9.16	19.08	24.43	13.16	15.8
Univl-DR (Liu et al., 2022)	151M	4.10	12.82	15.38	7.84	9.65	3.82	10.69	14.50	6.90	8.70
MARVEL (Zhou et al., 2024b)	310M	1.54	5.13	7.69	3.08	4.63	0.00	0.76	1.53	0.46	2.23
CLIP (Radford et al., 2021)	151M	12.31	26.15	33.33	19.01	22.43	10.69	21.37	25.95	15.69	18.15
CLIP-DPR (Liu et al., 2022)	151M	9.74	18.46	23.59	13.60	15.95	5.34	13.74	16.79	8.51	10.48
EVA-CLIP (Sun et al., 2023)	149M	5.13	19.49	24.10	11.08	14.21	3.05	9.92	13.74	6.40	8.14
CoCa (Cherti et al., 2023)	151M	8.72	21.54	25.1	14.11	16.76	3.05	8.40	9.92	5.19	6.33
Jina-CLIP-v2 (Koukounas et al., 2024)	865M	4.10	16.92	25.13	10.01	13.58	3.05	10.69	14.50	5.84	7.87
Long-CLIP-B (Zhang et al., 2025)	149M	26.67	53.85	62.56	37.99	43.90	22.90	40.46	51.91	30.29	35.33
ChartFinder-B	149M	38.97	62.05	70.26	49.30	61.30	29.77	50.38	61.83	38.96	44.41
Long-CLIP-L (Zhang et al., 2025)	427M	41.03	75.90	79.49	55.32	55.32	38.93	67.18	74.81	50.99	56.77
ChartFinder	427M	**47.18**	**80.00**	**84.10**	**61.23**	**66.90**	**41.98**	**75.57**	**80.15**	**55.31**	**61.40**

trieval models on the same benchmark.

Overall Results. Table 3 shows that text-index retrieval is a strong baseline, with BGE achieving the best text-only performance. However, ChartFinder still achieves the strongest overall ranking quality. For NDCG@10, ChartFinder reaches 0.669/0.614 on precise/fuzzy queries, outperforming the best Title-index baseline with 0.646/0.572 and the best Caption-index baseline with 0.620/0.589. These results indicate that generated titles and captions provide useful retrieval signals, but they remain indirect chart representations and can miss visual-statistical patterns. In contrast, ChartFinder directly retrieves over chart images while using synthesized insights only during training.

5.3 Exp-3: Ablation Study on Synthesized Semantic Insights

Experiment Settings. We conduct ablation experiments to analyze the contribution of three different insights. Using ChartFinder-B as the base model, we train seven variants with different insight combinations and compare them with the original Long-CLIP-B baseline. All variants are evaluated on CRBench under both precise and fuzzy queries.

Overall Results. Table 4 shows that each type of insight improves over the Long-CLIP-B baseline, confirming the effectiveness of synthesized semantic supervision. Among single-insight settings, statistics-oriented insights achieve the largest overall gain, increasing the overall score from 43.66 to 48.35, which suggests the importance of explicit statistical semantics for chart retrieval. The best performance is achieved when all three insights are used together, improving the overall score from 43.66 to 54.34. Specifically, the full setting improves NDCG@10 from 43.90 to 61.30 on precise queries and from 35.33 to 44.41 on fuzzy queries. These results indicate that the three insight types are complementary: visual-oriented insights capture chart structure, statistics-oriented insights encode data patterns, and task-oriented insights connect chart content with analytical intent.

5.4 Exp-4: Transferability of Synthesized Semantic Insights

Experiment Settings. We evaluate whether synthesized semantic insights can improve retrieval backbones beyond ChartFinder. Specifically, we fine-tune CLIP-DPR and UniVL-DR with the same insight data and compare them with their original versions. We also include Long-CLIP-B and ChartFinder-B as a stronger backbone pair. All models are evaluated on CRBench under precise and fuzzy query settings.

Overall Results. Table 5 shows that synthesized insights consistently improve different retrieval backbones. For example, NDCG@10 improves from 11.55 to 15.80 for CLIP-DPR and from 9.13 to 16.29 for UniVL-DR on precise queries. Similar gains are observed on fuzzy queries, where CLIP-DPR improves from 7.54 to 12.88 and UniVL-DR improves from 7.11 to 16.42. These results indicate that synthesized insights provide transferable chart-specific supervision rather than a model-specific optimization. The consistent gains across backbones suggest that the proposed insight synthesis pipeline can generally enhance cross-modal chart retrieval models.

6 Limitations

Dependence and Scope of Semantic Insight Generation. The insight generation process utilizes a LLM, meaning the quality of the output is subject

Table 3: Performance comparison between text-index retrieval and cross-modal retrieval on CRBench. For text-index retrieval, each chart is represented by either a generated title or a generated caption. **Bold** indicates the best overall result in each column, and underline indicates the best text-index retrieval result.

Index	Model	Precise Query				Fuzzy Query			
		R@1	R@5	R@10	NDCG@10	R@1	R@5	R@10	NDCG@10
Title	BM25	0.313	0.554	0.656	0.478	0.183	0.405	0.473	0.322
	BGE	**0.477**	0.733	0.815	0.646	0.374	0.656	0.786	0.572
	E5	0.379	0.656	0.723	0.551	0.321	0.603	0.672	0.492
	CLIP	0.328	0.605	0.703	0.501	0.282	0.496	0.603	0.429
Caption	BM25	0.374	0.610	0.718	0.538	0.176	0.397	0.466	0.322
	BGE	0.436	0.728	0.805	0.620	0.420	0.679	0.771	0.589
	E5	0.385	0.651	0.754	0.564	0.275	0.550	0.641	0.452
	CLIP	0.159	0.344	0.456	0.294	0.031	0.168	0.244	0.128
Cross-Modal Retrieval	CLIP	0.123	0.262	0.333	0.224	0.107	0.214	0.260	0.182
	Long-CLIP-L	0.410	0.759	0.795	0.553	0.389	0.672	0.748	0.568
	ChartFinder	0.472	**0.800**	**0.841**	**0.669**	**0.420**	**0.756**	**0.802**	**0.614**

Table 4: Ablation study on synthesized semantic insights on CRBench.

Visual	Statistics	Task	Precise Query			Fuzzy Query			Overall
			R@10	MRR@10	NDCG@10	R@10	MRR@10	NDCG@10	
			62.56	37.99	43.90	51.91	30.29	35.33	43.66
✓			69.64	44.39	48.69	56.28	27.27	34.29	46.67
	✓		66.15	45.18	50.18	57.25	32.75	38.60	48.35
		✓	65.64	44.44	49.52	58.78	32.24	38.59	48.20
✓		✓	**71.79**	44.96	51.40	60.31	35.24	41.17	50.81
✓	✓		66.15	43.26	48.74	54.96	32.20	37.66	47.16
	✓	✓	65.64	43.53	48.87	54.96	30.44	36.23	46.61
✓	✓	✓	70.26	**49.30**	**61.30**	**61.83**	**38.96**	**44.41**	**54.34**

Table 5: Transferability of synthesized semantic insights across retrieval backbones on CRBench.

Model	Precise Query		Fuzzy Query	
	R@10	NDCG@10	R@10	NDCG@10
CLIP-DPR	20.00	11.55	10.69	7.54
FT-CLIP-DPR	24.62	15.80	19.08	12.88
UniVL-DR	17.44	9.13	9.92	7.11
FT-UniVL-DR	26.15	16.29	20.61	16.42
Long-CLIP-B	62.56	43.90	51.91	35.33
ChartFinder-B	**70.26**	**61.30**	**61.83**	**44.41**

to the model's inherent knowledge and potential biases. Although the study validated a small sample of insights via crowdsourcing, it cannot guarantee that all 207,498 synthesized insights are entirely free of errors or misinterpretations.

Limitations of the CRBench Benchmark Construction. The queries in CRBench were generated by GPT-4o, not from organic user query logs. While this approach was designed to simulate realistic precise and fuzzy queries and was validated through a rigorous crowdsourcing process, these synthetic queries may not fully capture the diversity, ambiguity, and unforeseen intents of genuine user behavior.

Model Architecture and Application Scenario Constraints. The core architecture of ChartFinder is based on CLIP. Although the training strategy is innovative, its fundamental performance is still bound by the inherent capabilities and limitations of the underlying vision-language models.

Generalizability of Data Sources . The datasets used for generating the training data were sourced exclusively from Kaggle. While these are high-quality and derived from real-world applications, they may not fully represent the entire spectrum of data types and structures found within private, enterprise Business Intelligence (BI) systems, which can be more domain-specific or messy.

References

Alexander Bendeck, Dennis Bromley, and Vidya Setlur. 2024. Slopeseeker: A search tool for exploring a dataset of quantifiable trends. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 817–836.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 2818–2829. IEEE.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2024. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171.

Enamul Hoque and Maneesh Agrawala. 2019. Searching the visual style and structure of d3 visualizations. *IEEE transactions on visualization and computer graphics*, 26(1):1236–1245.

Daomin Ji, Hui Luo, Zhifeng Bao, and J. Shane Culpepper. 2024a. Navigating data repositories: Utilizing line charts to discover relevant datasets. *Proc. VLDB Endow.*, 17(12):4289–4292.

Daomin Ji, Hui Luo, Zhifeng Bao, and J. Shane Culpepper. 2024b. The story behind the lines: Line charts as a gateway to dataset discovery. *CoRR*, abs/2408.09506.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656.

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*.

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. *arXiv preprint arXiv:2203.06486*.

Andreas Koukounas, Georgios Mastrapas, Bo Wang, Mohammad Kalim Akram, Sedigheh Eslami, Michael Günther, Isabelle Mohr, Saba Sturua, Scott Martens, Nan Wang, and 1 others. 2024. jina-clip-v2: Multilingual multimodal embeddings for text and images. *arXiv preprint arXiv:2412.08802*.

Fritz Lekschas, Brant Peterson, Daniel Haehn, Eric Ma, Nils Gehlenborg, and Hanspeter Pfister. 2020. Peax: Interactive visual pattern search in sequential data using unsupervised deep representation learning. In *Computer Graphics Forum*, volume 39, pages 167–179. Wiley Online Library.

Haotian Li, Yong Wang, Aoyu Wu, Huan Wei, and Huamin Qu. 2022a. Structure-aware visualization retrieval. In *CHI*, pages 409:1–409:14. ACM.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

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

Zhenghao Liu, Chenyan Xiong, Yuanhuiyi Lv, Zhiyuan Liu, and Ge Yu. 2022. Universal vision-language dense retrieval: Learning a unified representation space for multi-modal retrieval. *arXiv preprint arXiv:2209.00179*.

Yuyu Luo, Xuedi Qin, Chengliang Chai, Nan Tang, Guoliang Li, and Wenbo Li. 2022a. Steerable self-driving data visualization. *IEEE Trans. Knowl. Data Eng.*, 34(1):475–490.

Yuyu Luo, Xuedi Qin, Nan Tang, and Guoliang Li. 2018a. Deepeye: Towards automatic data visualization. In *2018 IEEE 34th international conference on data engineering (ICDE)*, pages 101–112. IEEE.

Yuyu Luo, Xuedi Qin, Nan Tang, Guoliang Li, and Xinran Wang. 2018b. Deepeye: Creating good data visualizations by keyword search. In *SIGMOD Conference*, pages 1733–1736. ACM.

Yuyu Luo, Nan Tang, Guoliang Li, Chengliang Chai, Wenbo Li, and Xuedi Qin. 2021. Synthesizing natural language to visualization (NL2VIS) benchmarks from NL2SQL benchmarks. In *SIGMOD Conference*, pages 1235–1247. ACM.

Yuyu Luo, Nan Tang, Guoliang Li, Jiawei Tang, Chengliang Chai, and Xuedi Qin. 2022b. Natural language to visualization by neural machine translation. *IEEE Trans. Vis. Comput. Graph.*, 28(1):217–226.

Yuyu Luo, Yihui Zhou, Nan Tang, Guoliang Li, Chengliang Chai, and Leixian Shen. 2023. Learned data-aware image representations of line charts for similarity search. *Proc. ACM Manag. Data*, 1(1):88:1–88:29.

Pingchuan Ma, Rui Ding, Shi Han, and Dongmei Zhang. 2021. Metainsight: Automatic discovery of structured knowledge for exploratory data analysis. In *Proceedings of the 2021 International Conference on Management of Data*, SIGMOD '21, page 1262–1274, New York, NY, USA. Association for Computing Machinery.

Miro Mannino and Azza Abouzied. 2018. Qetch: Time series querying with expressive sketches. In *Proceedings of the 2018 International Conference on Management of Data*, pages 1741–1744.

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.

OWID. 2011. Our world in data.

Xuedi Qin, Yuyu Luo, Nan Tang, and Guoliang Li. 2018a. Deepeye: An automatic big data visualization framework. *Big Data Min. Anal.*, 1(1):75–82.

Xuedi Qin, Yuyu Luo, Nan Tang, and Guoliang Li. 2018b. Deepeye: Visualizing your data by keyword search. In *EDBT*, pages 441–444. OpenProceedings.org.

Xuedi Qin, Yuyu Luo, Nan Tang, and Guoliang Li. 2020. Making data visualization more efficient and effective: a survey. *VLDB J.*, 29(1):93–117.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Pew Research. 2004. Pew research public.

Babak Saleh, Mira Dontcheva, Aaron Hertzmann, and Zhicheng Liu. 2015. Learning style similarity for searching infographics. *arXiv preprint arXiv:1505.01214*.

Vidya Setlur, Andriy Kanyuka, and Arjun Srinivasan. 2023. Olio: A semantic search interface for data repositories. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–16.

Leixian Shen, Enya Shen, Yuyu Luo, Xiaocong Yang, Xuming Hu, Xiongshuai Zhang, Zhiwei Tai, and Jianmin Wang. 2022. Towards natural language interfaces for data visualization: A survey. *IEEE transactions on visualization and computer graphics*, 29(6):3121–3144.

Tarique Siddiqui, Albert Kim, John Lee, Karrie Karahalios, and Aditya Parameswaran. 2016. Effortless data exploration with zenvisage: an expressive and interactive visual analytics system. *arXiv preprint arXiv:1604.03583*.

Arjun Srinivasan and Vidya Setlur. 2021. Snowy: Recommending utterances for conversational visual analysis. In *The 34th annual ACM symposium on user interface software and technology*, pages 864–880.

Arjun Srinivasan and Vidya Setlur. 2023. Bolt: A natural language interface for dashboard authoring. In *EuroVis (Short Papers)*, pages 7–11.

statista. 2007. Chart of the day.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and 1 others. 2024. jina-embeddings-v3: Multilingual embeddings with task lora. *arXiv preprint arXiv:2409.10173*.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint: 2303.15389*.

Tableau. 2003. Tebleau public. Website. https://public.tableau.com/app/discover.

Benny J Tang, Angie Boggust, and Arvind Satyanarayan. 2023. Vistext: A benchmark for semantically rich chart captioning. *arXiv preprint: 2307.05356*.

Jiawei Tang, Yuyu Luo, Mourad Ouzzani, Guoliang Li, and Hongyang Chen. 2022. Sevi: Speech-to-visualization through neural machine translation. In *SIGMOD Conference*, pages 2353–2356. ACM.

Manasi Vartak, Silu Huang, Tarique Siddiqui, Samuel Madden, and Aditya Parameswaran. 2017. Towards visualization recommendation systems. *Acm Sigmod Record*, 45(4):34–39.

Qitong Wang and Themis Palpanas. 2021. Deep learning embeddings for data series similarity search. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 1708–1716.

Yifan Wu, Lutao Yan, Leixian Shen, Yunhai Wang, Nan Tang, and Yuyu Luo. 2024. Chartinsights: Evaluating multimodal large language models for low-level chart question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12174–12200.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. 2024a. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829.

Han Xiao, Georgios Mastrapas, and Bo Wang. 2024b. Jina clip: Your clip model is also your text retriever. In *Multi-modal Foundation Model meets Embodied AI Workshop@ ICML2024*.

Yilin Ye, Jianing Hao, Yihan Hou, Zhan Wang, Shishi Xiao, Yuyu Luo, and Wei Zeng. 2024. Generative AI for visualization: State of the art and future directions. *Vis. Informatics*, 8(1):43–66.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, and 1 others. 2021. Florence: A new foundation model for computer vision. *arXiv preprint: 2111.11432*.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. 2025. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*, pages 310–325. Springer.

Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. 2024. Magiclens: Self-supervised image retrieval with open-ended instructions. *arXiv preprint arXiv:2403.19651*.

Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. 2024a. Vista: Visualized text embedding for universal multi-modal retrieval. *arXiv preprint arXiv:2406.04292*.

Tianshuo Zhou, Sen Mei, Xinze Li, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, Yu Gu, and Ge Yu. 2024b. Marvel: Unlocking the multi-modal capability of dense retrieval via visual module plugin. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14608–14624.

A Complete Experiments

A.1 Exp-5: ChartFinder in the Zero-Shot Setting

To evaluate the generalization ability of ChartFinder, we conducted zero-shot experiments on two external public benchmarks, VisText (Tang et al., 2023) and Chart-to-Text (Kantharaj et al., 2022). In this setting, we directly applied the model trained on our synthetic data to these new tasks without any fine-tuning to test the transferability of its learned semantic understanding ability.

A.1.1 Experiments on VisText Benchmark.

We use the test set of VisText (Tang et al., 2023), comprising 882 charts with each chart corresponding to two captions mentioned in Section 3.3, to conduct this group of experiments. We take the two-level caption as the query, and the annotated chart as the retrieval target to perform text-to-chart retrieval tasks.

Overall Results. As shown in Table 6, ChartFinder performs best in all metricss in L1 caption, significantly ahead of other models. For example, Recall@1 reaches 96.15%, far exceeding other models. Recall@10 reaches 99.66%, close to a full match. MRR@10 and NDCG@10 are 96.90% and 97.44%, respectively, further verifying their advantage in ranking quality. ChartFinder-B still achieves excellent performance with fewer parameters. Recall@1 reaches 93.20%, which is comparable to Long-CLIP-L and about 3 percentage points higher than Long-CLIP-B. On Recall@10, ChartFinder-B reaches 98.85%, surpassing all models with similar parameter scales. L2+L3 captions are more inclined to detailed analytical descriptions. Thus, the retrieval difficulty is significantly higher than the L1 task, and the overall metrics have declined. ChartFinder still leads in this task with R@1 reaching 67.80%, R@10 reaching 89.12%, MRR@10 and NDCG@10 reaching 75.20% and 78.78%, respectively.

A.1.2 Experiments on Chart-to-Text Benchmark

We use the Chart-to-Text dataset (Kantharaj et al., 2022) to conduct this experiment. Specifically, this dataset has 1,393 text and chart pairs in the test set. We use the same metrics as other experiments.

Overall Results. As shown in Table 7, ChartFinder outperforms other models in all metrics. For example, Recall@1 reaches 97.06%, 2.09% higher than the second-best model (Long-CLIP-L, 94.97%). Recall@10 reaches 99.86%, almost close to the full score, showing the excellent performance of the model in the top-k retrieval task. MRR@10 and NDCG@10 are 98.41% and 98.80%, respectively, further verifying the effectiveness of Liter in ranking quality. ChartFinder-B also demonstrates obvious performance, surpassing Long-CLIP-B in all metrics, especially in Recall@1 (94.90% vs. 90.81%) and NDCG@10 (97.53% vs. 95.35%).

Although the Chart-to-Text dataset can effectively test the performance of the model in clear semantic scenarios, its task design also has certain limitations. For example, the same pictures exist in the dataset, which narrows the search scope. Most of the queries in Chart-to-Text are descriptive sentences generated based on charts. There is a lack of user intent-driven queries, which is not enough to fully evaluate the generalization and upper-limit capabilities of the model. In contrast, the queries in CRBench are more complex and closer to actual applications, which can better test the performance of the model in real retrieval scenarios.

A.2 Exp-6: The Impact of Center Crop Strategy

Experiment Settings. We mentioned in Section 5 Training Details that for the text-to-chart retrieval task, we replaced the traditional center crop preprocessing method in the CLIP-based models with a direct resize strategy. To verify the necessity of

Table 6: Performance Comparison of Text-to-Chart Retrieval Models on VisText Benchmark.

Models	L1 Caption				L2+L3 Caption			
	R@5	R@10	MRR@10	NDCG@10	R@5	R@10	MRR@10	NDCG@10
VISTA	87.76	89.80	79.85	82.27	63.95	70.86	37.61	41.80
Univl-DR	81.75	86.62	71.57	75.22	52.83	60.77	42.61	46.93
MARVEL	66.55	72.79	55.61	59.73	39.80	48.64	31.25	35.37
CLIP	89.46	92.74	83.05	85.41	68.25	74.83	56.81	61.14
CLIP-DPR	88.10	90.93	80.59	83.14	64.74	72.11	53.51	57.97
EVA-CLIP	87.07	89.68	78.50	81.26	49.43	57.26	39.39	43.64
CoCa	84.35	87.87	79.69	81.65	58.96	66.44	50.54	54.30
Jina-CLIP-v2	90.93	94.67	84.85	87.23	74.49	80.50	62.72	67.01
Long-CLIP-B	94.78	95.35	81.88	83.50	78.12	83.45	67.99	71.73
ChartFinder-B	97.85	98.75	95.17	96.05	79.48	84.01	69.64	73.11
Long-CLIP-L	98.07	99.09	95.21	96.15	82.99	88.66	73.43	77.09
ChartFinder	**99.43**	**99.66**	**96.90**	**97.44**	**84.58**	**89.12**	**75.20**	**78.78**

Table 7: Experimental Results on Chart-to-Text Benchmark.

Models	R@5	R@10	MRR@10	NDCG@10
VISTA	68.46	73.20	58.94	62.39
Univl-DR	67.91	73.08	57.60	61.35
MARVEL	38.19	46.30	28.97	33.08
CLIP	88.16	92.03	80.25	83.11
CLIP-DPR	82.56	85.86	73.09	76.21
EVA-CLIP	83.42	87.80	74.73	77.90
CoCa	83.92	87.65	74.49	77.92
Jina-CLIP-v2	83.78	88.23	73.51	77.08
Long-CLIP-B	98.28	98.85	94.18	95.35
ChartFinder-B	99.07	99.43	96.89	97.53
Long-CLIP-L	99.35	99.78	97.06	97.75
ChartFinder	**99.86**	**99.86**	**98.41**	**98.80**

Table 8: Ablation Study on Center Crop. *W/* direct resize means we directly resize the image to the required size instead of cropping the chart to the target size.

Models	R@10	MRR@5	MRR@10	NDCG@10
Urban1k				
Long-CLIP-B	94.1	97.4	85.8	88.6
w/ direct resize	93.2	96.1	85.1	87.8
	↓0.9	↓1.3	↓0.7	↓0.8
Long-CLIP-L	96.7	98.2	90.7	92.6
w/ direct resize	95.1	97.5	88.9	90.9
	↓1.6	↓0.7	↓1.8	↓1.5
Chart-to-Text				
Long-CLIP-B	87.29	90.38	79.87	82.44
w/ direct resize	98.28	98.85	94.18	95.35
	↑10.99	↑8.47	↑14.31	↑12.91
Long-CLIP-L	88.51	91.03	81.20	83.61
w/ direct resize	99.35	99.78	97.06	97.75
	↑10.84	↑8.75	↑15.86	↑14.14

this modification, we selected Long-CLIP-B and Long-CLIP-L as the base models and conducted comparative tests on the urban1k and Chart-to-Text datasets, respectively. Urban1k (Zhang et al., 2025) is a retrieval dataset containing natural images of urban scenes, which contains a variety of urban buildings, streets, and landscapes, representing a typical natural image retrieval task.

Overall Results. The results in Table 8 show that the preprocessing strategies have very different effects on different retrieval tasks. On the Urban1k dataset, direct resize caused a slight decrease in model performance, with Long-CLIP-B and Long-CLIP-L decreasing by 0.9 and 1.5 percentage points on average, respectively. However, on the Chart-to-Text dataset, the same strategy brought significant improvements, with the two models gaining approximately 12.3 and 13.1 percentage points in performance, respectively. This shows that in chart retrieval tasks, retaining the complete structure of the chart (including edge axis labels, legends, etc.) is more important than maintaining precise proportions, proving that our adjustments to chart-specific preprocessing strategies are necessary and effective.

B Case Study

Figure 4 compares the top-1 retrieval performance of our model and CLIP on CRBench. In Query #1, ChartFinder successfully retrieves a chart that includes both *urban* and *rural population*, while CLIP retrieves a chart focusing only on *urban population*, failing to meet the query requirements. In Query #2, ChartFinder correctly aligns with the semantic meaning of *booking distribution*, retrieving a pie chart about hotel booking channels. In contrast, CLIP retrieves a chart about optimizing distribution channels, which does not match the query's focus. In Query #3, ChartFinder captures the key-

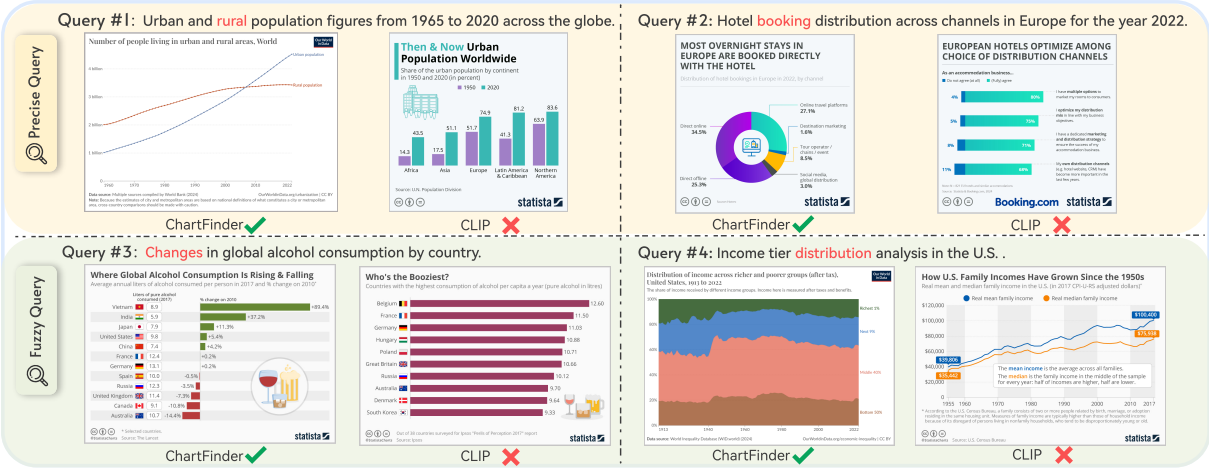


Figure 4: Case study on Top-1 Text-to-Chart Retrieval Results of the ChartFinder and CLIP.

word *change* and retrieves a chart showing trends in alcohol consumption. In contrast, CLIP retrieves a chart ranking alcohol consumption, ignoring the notion of *change*. In Query #4, ChartFinder identifies *distribution* and retrieves a stacked area chart illustrating income distribution. In contrast, CLIP retrieves a chart focusing on income growth trends, failing to address the *distribution* aspect.

(i) For precise query #1, ChartFinder successfully retrieves the charts that accurately match the query by capturing the keywords and filter conditions *urban and rural population* correctly. However, the chart CLIP retrieved only reflects the urban population, failing to meet the retrieval requirements of the rural population.

(ii) For precise query #2, our model aligns the semantic information about the hotel booking channels described by the pie chart with the *booking distribution* in the query. In contrast, although the retrieved chart of the CLIP is also about hotel information, the description is not about the different ways to book hotels, but the optimization of hotel distribution channels.

(iii) For fuzzy query #3, our model successfully captures the keyword *change* and retrieves a chart that clearly shows the trend in alcohol consumption. By contrast, although the chart retrieved by CLIP is also related to alcohol consumption, its content only focuses on the absolute ranking of alcohol consumption ("Who's the Booziest"), completely ignoring the core element of *change*.

(iv) For fuzzy query #4, *distribution* in red refers directly to the concept of data distribution. Our model successfully retrieves a stack area chart showing the distribution of income levels. However, the chart retrieved by CLIP is biased toward the trend of household income growth, although this graph is related to income and does not reflect the analytical information on income level distribution.

Through the four cases above, we can see the effectiveness of our model when matching both precise queries and fuzzy queries with their target charts.

C Complementary Analysis

C.1 Comparative Analysis with ChartQA

While prior benchmarks such as FigureQA (Kahou et al., 2017), DVQA (Kafle et al., 2018), and PlotQA (Methani et al., 2020) have established robust frameworks for chart understanding, they primarily focus on "single-chart question answering". In these settings, the model is required to extract graphical primitives, numerical values, or local relations from a specific provided image. ChartInsights (Wu et al., 2024) further extends this by evaluating multimodal large language models on low-level analytical perception. However, our proposed text-to-chart retrieval task addresses a distinct and complementary challenge: the discriminative selection of the most appropriate chart from a massive repository containing visually and semantically similar candidates. This paradigm shifts the focus from localized fact extraction within a single chart to the global alignment of natural language intent with latent visual representations across a vast candidate pool. By utilizing our semantic insights synthesis pipeline to bridge the gap between query intent and visual information through multi-level semantic supervision, we provide a new perspective for assessing how models understand the analytical semantics of charts in realistic Business

Intelligence (BI) scenarios. In summary, while ChartQA primarily evaluates the comprehension of a single chart in isolation, text-to-chart retrieval demands the simultaneous understanding and comparative analysis of multiple charts to identify the correct target, thereby offering a more rigorous and comprehensive benchmark for the development of real-world retrieval-and-ranking systems.

C.2 Error Analysis

To gain a deeper understanding of the model's limitations, we performed a fine-grained error analysis by conducting a word-level diagnosis on failed queries within CRBench. As summarized in Table 9, retrieval failures are not uniformly distributed but are significantly concentrated in queries involving implicit intent or complex logical structures. Specifically, "opinion-oriented" queries exhibit the highest failure bias of +20.5%, suggesting that the model struggles to align subjective business concerns with concrete visual features when the statistical target is not explicitly stated. Similarly, "multi-dimensional breakdown" queries show a failure bias of +9.8%, highlighting the difficulty in parsing intricate grouping relations where the analytical goal remains implicit .

In contrast, queries with explicit statistical cues such as *trend*, *growth*, or *ranking* yield negative failure biases, indicating that the model effectively captures well-defined analytical patterns with high reliability. These results underscore the necessity of our proposed multi-level semantic supervision to bridge the gap between high-level query intent and latent visual representations.

C.3 The Target-and-Distractor Paradigm

We emphasize that our design is intentionally structured to simulate realistic Business Intelligence (BI) scenarios and to evaluate the fundamental capabilities of retrieval-and-ranking systems.

- **Reason 1: Simulation of Intent-Driven BI Workflows.** In practical BI environments, users do not perform 1:1 static mappings; rather, they use natural language to search across a vast repository of existing chart assets. Our task setup mirrors this $1 \times N$ search process, shifting the focus from localized fact extraction to the global alignment of natural language intent with visual information across a massive candidate pool.

- **Reason 2: Assessment of Discriminative and Ranking Capabilities.** By matching 247 high-quality queries against 21,328 candidates—which include numerous visually and semantically similar distractors—we rigorously assess a model's ability to distinguish and rank appropriate targets. If a query were generated for every chart, the task would degenerate into a simple image-to-text description matching exercise, failing to evaluate the model's capacity for comparative analysis among similar candidates.

- **Reason 3: Task Complexity**. While benchmarks like Chart-QA evaluate the ability to "read" a single chart in isolation, CRBench addresses the distinct challenge of simultaneous understanding and comparative analysis. Maintaining a large candidate pool relative to the number of queries is essential for developing practical retrieval-and-ranking systems that must operate in high-interference, real-world scenarios.

C.4 Clarification of Paper Contribution

In this section we wish to further clarify the core value of our research across three dimensions—task definition, empirical significance, and methodological depth:

- **Contribution 1: Bridging the Gap Between Academic Focus and Industrial Requirements.** While Chart Question Answering (CQA) is widely studied in academia, the challenge of Chart Retrieval—a critical and frequent task in the Business Intelligence (BI) industry—has not received sufficient attention. Our proposed CRBench is the first large-scale benchmark specifically designed for this scenario. It is not merely a dataset but a formal academic modeling of the industrial pain point: "precise target selection from massive chart repositories".

- **Contribution 2: Revealing Performance Bottlenecks in Realistic BI Scenarios.** Through systematic evaluation on CRBench, we provide the first objective evidence that even state-of-the-art multimodal models fail to satisfy industrial-level performance requirements when faced with complex queries and highly similar distractors. This finding highlights a critical shift for the field: moving from

Table 9: Diagnosis of error-prone word sets in failed retrieval queries.

Error-prone word sets	Success rate	Failure rate	Failure bias
opinion / perception / concern / belief	15.4%	35.9%	**+20.5%**
by / among / across	43.4%	53.2%	**+9.8%**
distribution / percentage / share / proportion / composition / rate	23.1%	19.8%	-3.3%
compare / compared / versus (vs) /difference	15.9%	5.9%	-10.0%
trend / change / growth / decline / over time / between / from...to	61.0%	47.7%	-13.3%

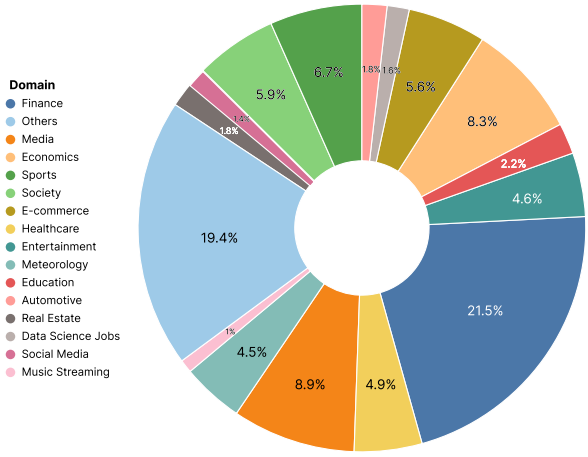


Figure 5: Domain distribution of data source.

"single-chart perception" to "multi-chart discriminative ranking".

- **Contribution 3: Proposed Data-Driven Pipeline and Scientific Insights.** To address these identified performance gaps, we propose a Multi-level Semantic Insights Synthesis Pipeline. Beyond achieving significant performance gains (e.g., nearly doubling results in fuzzy intent scenarios), our exploration has yielded several high-value scientific observations. These include the ··Scaling Law of modality superiority relative to encoder capacity"and the "absolute dependence of fuzzy analytical semantics on latent visual representations". These insights offer the community a reproducible path for utilizing large-scale synthetic data to solve complex cross-modal alignment problems. In summary, the contribution of this research lies in defining a novel and practical task, revealing an underestimated performance gap, and providing both a feasible solution and a set of foundational scientific insights.

C.5 Data Source Diversity

To quantitatively assess the semantic diversity of our collected corpus, we employed a scalable **hybrid annotation pipeline** on the 8,860 Kaggle CSV datasets. Given the scale of the data, we first utilized an unsupervised learning approach: metadata features (filenames and column headers) were extracted and mapped via TF-IDF, followed by K-Means clustering to group semantically similar datasets. Subsequently, we leveraged *GPT-4o* to automatically assign domain labels to each cluster by analyzing representative samples. To further resolve the long-tail distribution, we conducted a fine-grained reclassification, successfully identifying distinct sub-domains such as Society and Education.

As shown in Figure 5, the resulting taxonomy spans over 15 distinct domains. While Finance (21.5%) and Media (8.9%) constitute major proportions, the substantial presence of the "Others" category (19.4%) highlights the high-entropy and long-tail nature of our data sources, ensuring the benchmark's robustness across diverse semantic contexts.

D Prompt

D.1 Query Generation Prompts for GPT-4o

Query Generation Prompt for GPT-4o

```
You are a user searching for visualizations in a
      database. You will see 5 visualizations, but
      you should ONLY focus on the FIRST image when
      generating queries.
The other four images are very similar to the first
      one, make sure your queries CAN NOT match any
      of the other four images.
Generate two meaningful queries (10-15 words each):
1. Precise Query:
Generate a query about specific content in the FIRST
      visualization that includes:
- Meaningful combination of axis labels/categories
- Important data values or time ranges
- Key categories or measurements
- If the image contains time, include the time range
      in the query.
- When covering the time range, do not directly use
      the initial time range in the chart.
- The time range must be within the original chart
      time span.
- Do not use commas in your query.
- Do not directly copy text from the image. Use
      similar wording instead.

Example (good):
global temperature change 1990-2020 in the United
      States
```

```
Bad example:
global temperature change

2. Fuzzy Query:
Generate a query about the visualization purpose that
       includes:
- Line charts: trend analysis or comparison over time
- Bar charts: value comparison or ranking
- Pie/stacked charts: distribution or proportion
      comparison
- Scatter plots: correlation or pattern analysis
Example (good):
annual economic growth comparison between countries

Bad example:
growth comparison

Format your response as:
{
  "Precise query": "10-15 word query about content",
  "Fuzzy query": "10-15 word query about purpose"
}
Remember:
- ONLY consider the FIRST image
- Include relevant context
- Be specific and meaningful
- Articles and prepositions count as words
```

D.2 Semantic Insights Generation Prompts for LLaMA

Generation Prompts of Visual-oriented Insights for LLaMA

```
system_prompt = """You are an expert data
      visualization analyst who excels at crafting
      clear, engaging narratives.
Your task is to write a fluid, well-structured
      paragraph that describes a data visualization.
Write as if explaining the visualization to a
      professional audience."""

user_prompt = f"""Based on the following chart
      information, write a single cohesive paragraph
      explaining the visualization:

Title: {chart_data['describe']}
Subtitle: {chart_data['sub_title']}
Chart Type: {chart_data['chart']}
X-axis: {chart_data['x_name']}
Y-axis: {chart_data['y_name']}
Categories: {', '.join(chart_data.get('classify', [])
      ) if chart_data.get('classify') else 'Single
      category'}

Requirements:
- Begin with "This {chart_data['chart']} chart"
- Naturally describe relationships between variables
- Use professional yet accessible language
- Approximately 100 words
"""
```

Generation Prompts of Task-oriented Insights for LLaMA

```
system_prompt = """You are a professional data
      analyst. Describe the chart's purpose and
      practical applications.

Format strictly:

Main Purpose:
[Single paragraph of 50-100 words describing the
      visualization's core objective and data
      presentation approach. Avoid introductory
      phrases.]
"""

user_prompt = f"""Chart information:
Title: {chart_data['describe']}
Subtitle: {chart_data['sub_title']}
Chart Type: {chart_data['chart']}
X-axis Label: {chart_data['x_name']}
Y-axis Label: {chart_data['y_name']}
Data Categories: {', '.join(chart_data['classify'])
      if chart_data['classify'] else 'None'}
"""
```

Generation Prompts of Statistics-oriented Insights for LLaMA

```
system_prompt = """You are an expert data analyst.
      Based on key statistical metrics, provide
      analysis.

Format strictly:

Statistical Analysis:
[Single paragraph of 50-100 words analyzing ordering,
       relationships, ranges, and correlations. Avoid
       introductory phrases.]
"""

user_prompt = f"""Chart Information:
Title: {chart_data['describe']}
Chart Type: {chart_type}
X-axis: {chart_data['x_name']}
Y-axis: {chart_data['y_name']}

Key Statistics:
{stats_info}
"""
```

E Ethical and Open Source Statements

E.1 Ethical Statements

Regarding the concern on data compliance, we emphasize that all Kaggle datasets utilized in this work were acquired exclusively through the official Kaggle API. This acquisition process strictly adheres to Kaggle's Terms of Service and official data access protocols, ensuring that our data collection and utilization are fully compliant with legal requirements and standard research ethics guidelines.

E.2 Open Source Statements

Regarding the inquiry about data availability, we confirm that all research artifacts—including the CRBench benchmark, the fine-tuning datasets, model weights, and the complete training pipeline—will be fully open-sourced upon the acceptance of this paper. We are committed to ensuring transparency and reproducibility to facilitate further research within the community.