ChartWalker: Benchmarking the Cross-Chart RAG Task with Hierarchical Knowledge Graphs

Ning Tang [*1,2] Chenghan Xie [*3] Hanyang Yuan [*4] Yi Li [+4] Renhong Huang [4]
Qian Kou [2] Xiaofeng Shi [2‡] Hua Zhou [2] Jiarong Xu [1‡]

[1]Fudan University [2]Beijing Academy of Artificial Intelligence [3]Stanford University [4]Zhejiang University

Abstract

Cross-Chart Retrieval-Augmented Generation (RAG) is critical for complex multimodal analysis in various domains. However, existing benchmarks either focus on well-structured tables or generate cross-chart queries via key-point extraction, leading to lexical overlap and logically weak reasoning. To address this, we propose **ChartWalker**, a novel framework for constructing challenging cross-chart RAG tasks. Specifically, ChartWalker constructs hierarchical knowledge graphs tailored to charts to preserve analytical structure. Furthermore, we employ a structure-aware sampling algorithm to synthesize semantically coherent multi-hop reasoning paths with controllable difficulty and granularity. Based on this framework, we introduce **ChartWalker-Bench**, a comprehensive benchmark spanning multiple cross-chart query types. Extensive evaluations across representative RAG paradigms reveal significant performance gaps. We further release **ChartWalker-Agent**, an agentic baseline to support analysis and future system development. Code is available at https://github.com/downing777/ChartWalker_Pub.git.

1. Inroduction

Charts are a primary medium for visualizing quantitative statistics in science, business, journalism, and policy (Das & Soylu, 2023; Norasaed & Siriborvornratanakul, 2024; Kastellec & Leoni, 2007). Unlike natural images or tables, they are information-dense and weakly structured. Answering questions by synthesizing evidence across multiple charts is a common requirement in real-world analysis. For instance, an analyst may need to relate a country's GDP growth trend in one chart to its inflation or unemployment trajectory in another, in order to characterize macroeconomic cycles.

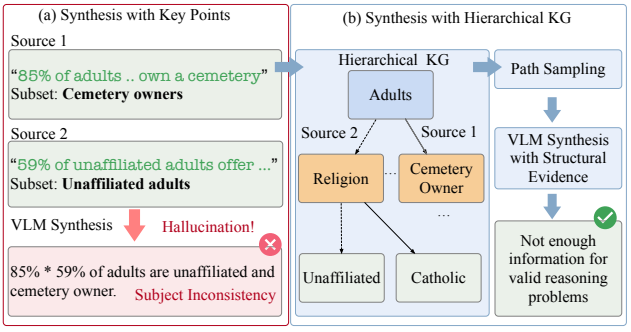


Figure 1. Compared to (a) concatenating isolated key statistics and prompting a VLM to synthesize cross-chart questions, our hierarchical KG (b) explicitly represents entities with their structural relations. Conditioning question generation on these structural paths makes entity dependencies clear and reduces the hallucination of incompatible subjects.

Recent advances in fundamental Vision Language Models (VLMs) and Multi-modal Language Models (MLMs) (Bai et al., 2025a; OpenAI, 2024) have substantially strengthened their capabilities in visual perception and reasoning. When fine-tuned on cross-chart QA questions (Masry et al., 2025), such models exhibit strong performance in complex numerical and multi-chart analytical tasks. However, fine-tuning alone cannot realistically endow a model with access to all chart-specific knowledge, nor does it generalize well to open-domain or long-tail chart collections encountered in real-world settings. Consequently, many practical applications adopt a Chart Retrieval-Augmented Generation (Chart RAG) paradigm (Yang et al., 2025), where charts are commonly retrieved as external knowledge sources to support analysis. Such scenarios are especially common in domains such as scientific surveys, market research, journalism forecasts, and political analysis (Kim et al., 2020; Kastellec & Leoni, 2007).

While the Cross-Chart RAG Task is of substantial practical importance, there is a lack of a good benchmark that fully captures the multimodal nature of charts and character-

[*]Equal contribution. [+]Core contributors. [‡]Corresponding authors.
Emails: ningtang24@m.fudan.edu.cn, jiarongxu@fudan.edu.cn, sxf1052566766@163.com, {yuanhanyang,3200105508,renh2}@zju.edu.cn

izes the underlying reasoning structure required by realistic queries. We identify two major limitations in existing benchmarks. First, most prior work focuses on tables rather than charts (Zou et al., 2025; Yu et al., 2025). In these settings, the underlying data is already explicitly structured, with clear entity boundaries and relations, enabling questions to be constructed and answered through symbolic or textual reasoning alone. Moreover, as illustrated in Figure 1, recent chart RAG benchmarks (Yang et al., 2025) simply linking semantically similar key points can yield brittle reasoning chains, where implicit referents drift across hops and may lead to subject-mismatch and logically invalid computations (*e.g.*, a follow-up clause refers to a subset while evidence is drawn from the full population).

To address these deficiencies, we aim to introduce a logically grounded and complex cross-chart RAG benchmark for evaluating multimodal RAG pipelines. Knowledge graph (KG) provides an intuitive approach to generate such multi-hop QAs: it extracts the implicit entity–relation structure embedded in charts and preserves explicit reasoning paths, which can be directly used to annotate questions with grounded reasoning chains.(Lu et al., 2025; Yang et al., 2018). However, existing KG-based QA generation pipelines often rely on random walks or naive PageRank to synthesize long-hop reasoning paths. The path sampling is largely blind to query design, offering limited control over the entity-level constraints that determine granularity and complexity. Moreover, such paths are frequently semantically incoherent: successive hops may be globally unrelated, causing meaningless analysis and sample waste.

To bridge these gaps, we present ChartWalker, a novel chart-centric framework designed to construct challenging cross-chart RAG tasks. Our framework introduces two tightly coupled innovations. First, we propose a hierarchical KG construction method tailored to chart data, which extracts chart entities and relations into explicit layers according to their information granularity. This design enables comprehensive semantic coverage while preserving the inherent structure of dense chart information. Second, building on this hierarchy, we introduce a structure-aware sampling algorithm for cross-chart reasoning path synthesis. Our sampler enforces semantic continuity along paths, ensuring that successive hops remain logically coherent and analytically meaningful. The resulting reasoning paths serve as supervision signals for generating multi-hop QA pairs.

Beyond this methodology, we further release a high-quality cross-chart RAG benchmark, ChartWalker-Bench, comprising 564 multi-hop QA instances across 4 query types. Comprehensive experiments are conducted across major RAG paradigms with different VLM generators. Experiments show that the best-performing model achieves only a 64% correctness rate in answering the cross-chart problems.

Table 1. Comparison between existing cross-chart benchmarks.

Benchmarks	Objective	Open Domain	Structrual Info.	Complex Reason
HeteQA(Yu et al., 2025)	Table	✓	✓	✓
Open-WikiTable(Kweon et al., 2023)	Table	✓	✗	✓
MTabVQA(Singh et al., 2025a)	Table	✗	✓	✗
MultiTableQA(Zou et al., 2025)	Table	✓	✗	✓
ChartQA-Pro(Masry et al., 2025)	Chart	✗	✗	✗
Chart-MRAG(Yang et al., 2025)	Chart	✓	✗	✓
ChartWalker(Ours)	Chart	✓	✓	✓

More critically, on complex reasoning queries, accuracy drops sharply, falling below 30% for the majority of cases, highlighting its potential for advancing the multimodal RAG system's capability in multi-step quantitative retrieval and reasoning. In addition, we provide ChartWalker-Agent, a VLM-based search agent baseline that facilitates the analysis of experience reuse and informs future ChartRAG system design. Our main contributions are summarized as follows:

- We introduce ChartWalker, a chart-centric framework that explicitly exposes the multi-granular structure of charts by organizing extracted entities and relations into a hierarchical knowledge graph, and synthesizes semantically coherent reasoning paths via structure-aware sampling.
- We release ChartWalker-Bench, a curated cross-chart RAG benchmark, with annotations grounded on explicit reasoning chains. Extensive experiments on mainstream RAG baselines show ChartWalker-Bench's difficulty in both retrieval and generation stages.
- We further present CharWalker-Agent, a VLM-based search agent for solving the multi-hop reasoning problem, offering insights and experimental analysis for future agent design.

2. Related Work

Chart RAG Benchmark. A chart is a general form of visual representation that combines data with graphical marks to convey information. Tables, graphs, and diagrams can all be viewed as instances of charts. Early work focused on QA over tables, typically assuming a given table context and relatively simple operations(Pasupat & Liang, 2015; Zhong et al., 2017). Research on complex reasoning over charts (Li et al., 2024) emerge with the advance in VLM's capability, where models are required to reason directly over graphical features(Masry et al., 2025) rather than relying on explicit textual or symbolic table schema (Xie et al., 2025; Singh et al., 2025a). Moving beyond the close domain understanding tasks, (Herzig et al., 2021) propose the first research of table RAG problem, where relevant tables must be retrieved from large corpora before reasoning. Building upon this line of work, subsequent studies extended table RAG to more complex cross-table settings (Kweon et al., 2023; Zou

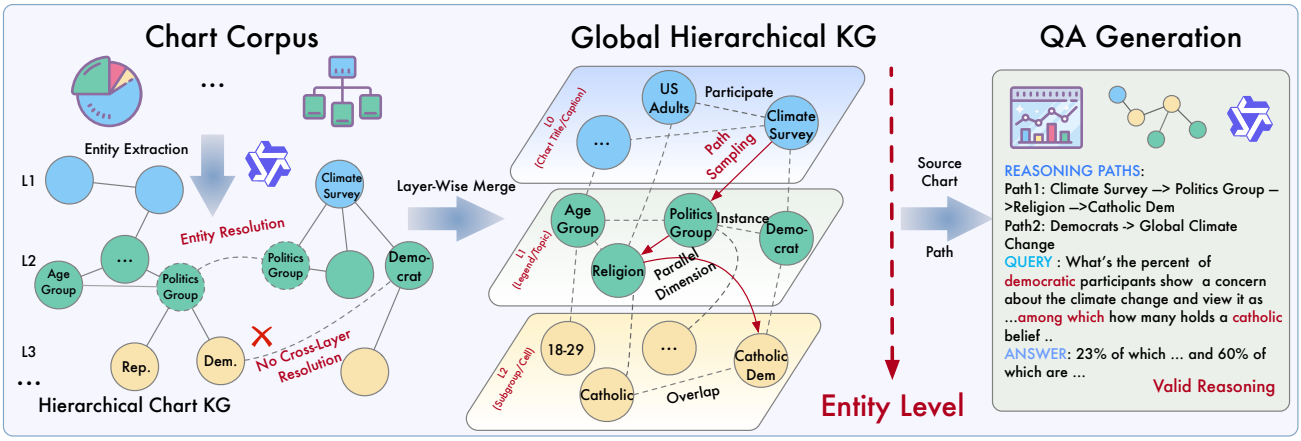


Figure 2. Illustration of the ChartWalker framework. Given a large chart corpus, a VLM first extracts entities and relations from each chart to build per-chart hierarchical knowledge graphs, where entities are organized by granularity levels. Then, identical entities are merged layer-wise to form a global hierarchical KG over the entire corpus. On top of this hierarchy, we perform structure-aware path sampling to construct multi-hop reasoning paths that traverse entities across levels and sources. Finally, we generate cross-chart QA pairs by conditioning on the original chart images and the sampled reasoning paths.

et al., 2025). The most related work is ChartMRAG (Yang et al., 2025), which is among the first to benchmark the cross-chart RAG tasks. However, its reliance on semantic similarity yields incoherent reasoning paths, limiting the reliability as an evaluation of cross-chart reasoning. Table 1 illustrates the main differences between existing Chart RAG Benchmarks and ChartWalker-Bench.

Multi-hop Question Generation (MHQG). MHQG aims to synthesize questions that require multi-step reasoning across multiple contexts (Mavi et al., 2024). Early methods largely relied on explicit structured representations to scaffold reasoning. (Kumar et al., 2019) formulate difficulty-controllable generation by sampling paths on knowledge graphs, where hop count is used as a proxy for reasoning difficulty. (Liu et al., 2023) applies graph convolution to capture global dependencies, enabling evidence aggregation without requiring explicit sentence-level labels. Recent studies target both diversity and logical tightness in multi-hop synthesis (Cheng et al., 2024). KCS (Wang et al., 2025) introduces Knowledge Composition Sampling, which stochastically samples different knowledge combinations from the same context to diversify generated questions while maintaining relevance. In parallel, RT-RAG (Shi et al., 2026) explicitly decomposes complex questions into hierarchical reasoning trees to mitigate error propagation across steps. Despite advances, existing MHQG methods still face a severe scalability challenge, especially in multimodal settings: as hop length increases, these methods suffer from information loss and semantic drift. Consequently, the generated questions often lack meaningful cross-modal grounding and fail to reflect valid multi-hop reasoning.

3. ChartWalker Benchmark

In this section, we introduce ChartWalker-Bench, a benchmark for chart RAG that provides query–answer pairs with diverse query granularities, along with logically coherent and verifiable rationales that support systematic evaluation. We first formalize the chart retrieval-augmented generation task in §3.1. We then describe our benchmark construction methodology, including: (1) constructing a knowledge graph that explicitly links entities within and across charts (§3.2); and (2) performing path sampling on this graph to obtain coherent multi-hop paths, which are then used to synthesize QA instances with grounded reasoning structures (§3.3). We provide an illustration in Figure 2. Finally, we detail how we instantiate this pipeline to construct ChartWalker from real-world data in §3.4, including the data sources, processing steps, and resulting dataset statistics.

3.1. Problem Formulation

In this work, we focus on the RAG task in the context of charts in their general form, *i.e.*, as images or text-based formats, rather than traditional tables in text form only (Kweon et al., 2023; Yu et al., 2025). Charts are more ubiquitous in real-world documents and convey information through richer and more flexible visual encodings (*e.g.*, marks, colors, shapes, spatial layouts), making chart RAG both more practical and more challenging to reason over.

Cross-Chart RAG Task Let $\mathcal{C} = \{c_j\}_{j=1}^{N}$ be a corpus of charts, and let q be a natural-language query. A ChartRAG system is composed of a retriever f_r and a generator f_g (*e.g.* a pre-trained VLM). The retriever ranks all charts in $\mathcal{C}$ by relevance to q and returns the top-k charts $\mathcal{C}^k = f_r(q; \mathcal{C})$.

Conditioned on the query and the retrieved set, the generator produces an answer $\hat{y} = f_g(q, \mathcal{C}^k)$. The objective of the cross-chart ChartRAG is to maximize answer correctness.

Existing evaluations (Yang et al., 2025) on cross-chart RAG can suffer from limitations, as the generated QAs may exhibit subject-mismatch and logically invalid reasoning. To address this problem, we leverage a knowledge graph to make information links explicit across charts and ultimately generate queries of varying granularity with logically coherent rationales. Specifically, to enable controllable granularity in query generation, we first construct a hierarchical knowledge graph over the entire chart corpus, where entities are organized by their information granularity. To ensure that each query is associated with a logically coherent rationale for the answer, we perform constrained path sampling for the final QA generation. Details are presented below.

3.2. Hierarchical Chart Knowledge Graph

Inspired by (Pasupat & Liang, 2015; Zhang et al., 2020), we seek to represent the chart corpus as a knowledge graph, which provides a uniform substrate for connecting entities across heterogeneous charts and makes multi-hop reasoning explicit. The KG nodes correspond to chart entities (*e.g.*, titles, legends, and individual units), and the edges encode semantic relations. In practice, as charts convey information at different granularities, user queries can naturally span multiple information scales, such as global context from titles or captions, series- or category-level patterns, and individual unit values; thus, an effective benchmark must account for how RAG methods retrieve evidence across different granularities. However, in a naive KG, entities extracted at different information scales are not explicitly distinguished or organized by granularity (Han et al., 2025). As a result, multi-hop sampling can drift across levels, making the semantic granularity of the resulting QA data difficult to control. To address this issue, we organize entities into a hierarchical KG that preserves the chart's inherent information scale. Specifically, we first construct a hierarchical local KG for each chart, and then obtain a unified KG through global integration. This design enables the subsequent generation of multi-hop reasoning paths with controllable information granularity.

Chart Graph Construction. Given a chart c, we first prompt a VLM-based extractor to identify structured entities in the chart and annotate each entity with a granularity level. Formally, the extractor returns the entity set as

$$\mathcal{V}_c = \mathsf{VLM}(c, \mathsf{Prompt}_{\mathsf{ent}}), \tag{1}$$

where $\mathsf{Prompt}_{\mathsf{ent}}$ denotes the prompt for entity extraction. For $v \in \mathcal{V}_c$, a corresponding granularity level is given as

$$l_v = \mathsf{VLM}(v, \mathsf{Prompt}_{\mathsf{lvl}}), \tag{2}$$

where $l_v \in \{0, 1, \ldots, L\}$. Smaller l corresponds to coarser, more globally informative entities (*e.g.*, title entities), and larger l corresponds to finer-grained units (*e.g.*, datapoints).

Subsequently, we instantiate edges between extracted entities to capture the relationships between chart components. Intuitively, entities at the same level can exhibit associative relations, while entities at different levels can reflect semantic progression. For example, titles and captions define the topic, axes and legends specify comparison dimensions, and marks and datapoints realize these dimensions with concrete values. We therefore connect two entities with an intra-level edge if they exhibit an associative relation, or with an inter-level edge if they reflect semantic progression. Because the same pair of entities can be linked multiple times, the resulting graph is essentially a multigraph and each edge is represented as a relation triple $(v, r, u) \in \mathcal{E}_c$. Together with the extracted entities, we obtain the resulting hierarchical chart KG $\mathcal{G}_c = (\mathcal{V}_c, \mathcal{E}_c)$.

Global Integration. The global KG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is the union of all chart subgraphs:

$$\mathcal{V} = \bigcup_{c \in \mathcal{C}} \mathcal{V}_c, \qquad \mathcal{E} = \bigcup_{c \in \mathcal{C}} \mathcal{E}_c. \tag{3}$$

The same entity may be mentioned across different charts, and to avoid duplicate entities, we merge identical entities at each level and rewrite edges accordingly. Note that the level does not correspond to a semantic or linguistic hierarchy (*e.g.*, abstract concepts vs. concrete instantiations). Instead, it reflects how informative the entity is to its source chart. Therefore, identical entities can legitimately occur in different layers, and we do not reconcile them across levels.

3.3. Path Sampling and QA Synthesis

As we construct the knowledge graph, the next challenge lies in how to sample paths and use them as supervision signals for QA generation. A central consideration is to ensure that the generated QAs cover diverse information-query granularities and are accompanied by logically coherent rationales. However, existing sampling techniques, such as unconstrained random walks (Lu et al., 2025) or naive multi-hop expansion (Guo et al., 2025), may quickly drift to weakly related entities, producing paths that are difficult to convert into coherent, answerable questions, and they also lack a mechanism for granularity-controlled sampling. To address this, we impose constraints during sampling to control both the granularity of evidence through level transitions and the semantic coherence around an anchor topic. Specifically, before sampling begins, we first select an *anchor* entity as the starting point of the path, considering its importance within the knowledge graph. During the sampling process, we apply constraints to the next-hop sampling policy, which are derived from both semantic topic

coherence and information-level considerations. Finally, the sampled paths are converted into grounded QA records with explicit citations.

Anchor Selection. We begin by selecting a globally salient and well-connected anchor entity from the KG. A natural approach is to score entities using PageRank (Brin & Page, 1998), which favors central entities with high connectivity. However, in the chart KG, an entity may have a high degree because it is repeatedly connected within a single, information-dense chart. For example, the entity "Nation" may link to many country-name entities in a comparison chart, but these connections lack *source diversity*, making such entities less useful for cross-chart reasoning.

In this sense, we compute the modified PageRank score of entity nodes using a weighted transition probability that jointly considers their connectivity and source diversity. Formally, we define the transition matrix $\boldsymbol{M} = [M_{u,v}]$ as

$$M_{u,v} = \frac{N_{\text{edge}}(v,u) \cdot N_{\text{src}}(u)}{\sum_{u' \in \mathcal{N}(v)} N_{\text{edge}}(v,u') \cdot N_{\text{src}}(u')}, \quad (4)$$

where $\mathcal{N}(v)$ denotes neighbors of v, $N_{\text{edge}}(v,u)$ is the number of edges between v and u, measuring connectivity, and $N_{\text{src}}(u)$ is the number of distinct source charts among all edges related to u, representing source diversity. Qualitatively, higher source diversity encourages a stronger transition probability.

The final PageRank score is defined as the stationary distribution $\boldsymbol{\pi}$ of this transition matrix, which satisfies $\boldsymbol{\pi} = \boldsymbol{\pi}\boldsymbol{M}$. During path sampling, the starting entity v_1 is then chosen following the distribution $v_1 \sim \boldsymbol{\pi}$.

Path Sampling. Since the starting anchor only provides a structural foundation for generating the reasoning path, the quality of the resulting QA depends significantly on how we expand from the anchor to collect supporting evidence. To limit semantic drift and proactively control the query's granularity, we define the next-hop sampling policy *w.r.t* semantic topic coherence and the level of the entities being reached. Denote a sampled path of T hops as:

$$\mathcal{P}^T = \{(v_t, r_t, u_t) \mid v_{t+1} = u_t, (v_t, r_t, u_t) \in \mathcal{E}\}_{t=1}^{T}. \quad (5)$$

In this process, the next hop policy follows:

$$p(v_{t+1}, r_{t+1}, u_{t+1} \mid \mathcal{P}^t) \propto \pi(v_{t+1}) \cdot \phi_{\text{sem}} \cdot \phi_{\text{gran}}. \quad (6)$$

Here ϕ_{sem} is the cosine similarity between $(v_{t+1}, r_{+1}, u_{t+1})$ and current path $\mathcal{P}^t$, capturing semantic topic coherence. ϕ_{gran} is a dynamically adjusted scalar function that varies across different sampling processes. It controls the granularity of transitions, assigning different values for same-level and cross-level moves. For example, it can favor upward-level transitions (*i.e.* $l_{u_{t+1}} > l_{u_t}$) by assigning larger values, biasing the policy towards higher-level entities, and increasing the query's granularity. Alternatively, the opposite strategy can favor staying at a shallower level to generate coarser-grained queries. The sampling process terminates when either the maximum hop budget T is reached or the current entity has no outgoing edges.

QA Generation. In practice, to generate a high-quality QA instance, we sample multiple paths rooted at the same starting entity, as this provides more avenues to extract information from different perspectives. All the sampled paths are packed in a unified prompt skeleton (prompt variants across query types are listed in Appendix A.3). A VLM is instructed to formulate questions based on this specific information *w.r.t* different query types (see §3.4 for detailed query types), and output a complete QA record containing the question, answer, explanation, and explicit evidence usage. We also set constraints where questions must be decontextualized, meaning they are fully self-contained and do not rely on implicit references (*e.g.*, pronouns such as "this" or "that"). Additionally, queries exhibiting excessive lexical overlap with the original chart text are paraphrased to reduce direct lexical copying. Formally, we have:

$$\text{QA} = \text{VLM}\left(\{\mathcal{P}_i\}_{i=1}^{k}, \text{Prompt}_{\text{gen}}\right), \quad (7)$$

where $\{\mathcal{P}_i\}_{i=1}^{k}$ denotes the sampled k paths and $\text{Prompt}_{\text{gen}}$ denotes the prompt skeleton for QA generation. This is followed by post-verification to ensure answer correctness against the evidence, with resampling if verification fails.

3.4. Benchmark Construction

Utilizing the proposed construction pipeline, we next introduce how ChartWalker is built upon real-world data. The original chart corpus is collected from ChartMRAG (Yang et al., 2025) and ChartQA-pro (Masry et al., 2025). Based on this, we further curate the corpus by using a VLM to filter these source charts according to visual clarity, semantic richness, and by merging them based on rule-based duplicate detection. This process yielded a corpus of 806 charts, encompassing a wide variety of chart styles and in-chart information. Subsequently, we construct a subgraph per chart and merge into a global hierarchical chart KG of 4 layers with 8802 entities and 21436 relations. Based on the constructed KG, we perform path sampling with 4 paths per question and up to 4 hops per path, enforcing at least 2 unique chart sources and a maximum of 5 sources per QA pair, yielding 737 generated raw QAs.

Query types. Following the prior work (Li et al., 2024; Singh et al., 2025a; Zou et al., 2025), ChartWalker groups the generated queries into 4 types based on common scenarios: (i) Fact Check, (ii) Manipulation (sum/average/rank/-compare), (iii) Analysis and (iv) Complex Reasoning.

Table 2. Question statistics by category: **Src** = average sources; **Path** = average paths used for query construction; **Hop** = average reasoning hops; **Diff** = average subjective difficulty score; **Pass** = quality control pass rate.

Category	Num	Src	Path	Hop	Diff	Pass
FactCheck	72	2.03	2.44	4.04	2.00	0.92
Manipulation	141	2.09	2.79	4.56	2.07	0.84
Analysis	242	2.13	2.58	4.36	2.30	0.83
Reasoning	109	3.12	3.60	5.14	3.00	0.55
Overall	564	2.30	2.81	4.52	2.34	0.77

Post-verification. We apply an automatic quality control step to filter generated QA pairs before constructing the benchmark. The verifier is given the question, the proposed ground-truth answer, and the associated evidence. It outputs strict labels for "supported" and "meaningful" (each in {yes, no, uncertain}). We keep an instance only if both labels are "yes". The resulting overall pass rate is 0.77. The final filtered result contains 564 QA pairs, which constitute the final ChartWalker. We report key statistics in Table 2.

Reasoning Complexity Our benchmark exhibits substantial reasoning complexity. On average, each question draws on 2.30 evidence sources and spans 2.81 reasoning branches among the sampled paths. Notably, the *Complex Reasoning* subset is the most information-intensive, with the highest source charts usage (3.12) and reasoning hops (5.14), creating a challenging setting that stresses a RAG pipeline's ability to retrieve the right charts under distraction and then integrate multiple pieces of evidence into a coherent answer. To further characterize difficulty, we ask the VLM to provide a subjective difficulty score in $\{1, 2, 3\}$ for each question. This score is consistent with the statistics above, with Complex Reasoning showing the greatest difficulty.

4. ChartWalker Agent

A key challenge in cross-chart reasoning lies beyond chart perception, stemming from the heterogeneous reasoning requirements across query types. Different queries demand evidence to be retrieved at varying levels of granularity and composed through distinct reasoning processes, which limits the effectiveness of static retrieval pipelines. As demonstrated in our experiments (Table 3), existing static retrieval-based approaches exhibit notable performance degradation on complex multi-hop reasoning queries. Even with up to 10 retrieved source charts, the strongest model attains only 51% answer accuracy. In most settings, performance remains below 30%. These results suggest that a single static RAG pipeline is insufficient to support long-horizon reasoning and multi-granularity evidence aggregation across diverse cross-chart queries.

Recent advances in agentic retrieval-augmented generation have shown that treating retrieval as a sequential decision-making process can substantially improve long-horizon reasoning (Singh et al., 2025b). By allowing models to iteratively observe the source corpus, maintain searching memory, and adapt a dynamic retrieval strategy, agent-based approaches offer a promising mechanism for handling the complex tasks. Motivated by these developments (Geng et al., 2025; Lu et al., 2025), we design the ChartWalker Agent, which serves as a baseline for better benchmarking cross-chart long-horizon reasoning. The ChartWalker agent is a VLM-based search agent that navigates a chart knowledge graph to iteratively acquire evidence and perform step-wise reasoning, functioning as a diagnostic and analytical tool for studying cross-chart RAG behaviors and informing future research.

Environment and Action Space. The ChartWalker agent operates on a KG constructed from chart entities and their relations. To prevent potential information leakage from reusing the hierarchical KG built during benchmark construction, we follow (Guo et al., 2025) to rebuild a standard KG. In this process, we generate a summary for each chart and treat the summaries as textual passages for KG construction. Accordingly, the agent aims to iteratively acquire evidence by exploring the KG, retrieving relevant chart entities and source chart information, and terminating once sufficient evidence has been collected. To avoid context overflow and unstable credit assignment caused by concatenating long interaction trajectories, we formulate the problem as a partially observable Markov decision process (POMDP) and compress the agent's memory into the environment observation. Specifically, the observation provided to the agent summarizes the current search state, including the current entity, its neighboring entities, and relations in the KG, as well as a simplified record of past search history with retrieved source charts. At each turn, the agent decides to act from $\{\texttt{move}, \texttt{edge\_search}, \texttt{backward}, \texttt{stop}\}$. See more environment and action design at Appendix A.4.

Training Objective and Optimization. During exploration, the agent observes the current KG context along with retrieved evidence, takes an action to further explore the KG, and receives a scalar reward from the environment at each turn. The training objective is to maximize the expected discounted return (Schulman et al., 2017). In our setting, the agent follows a policy parameterized by a VLM, which autoregressively outputs a token sequence as the action. Following (Wang* et al., 2025), we adopt non-concatenated PPO rollouts and a turn-level advantage assignment scheme. Specifically, we estimate the advantage function using temporal-difference learning and assign the resulting turn-level advantage uniformly to all action tokens within each turn. Implementation details are deferred to Appendix A.4. We evaluate the performance of the

Table 3. Retrieval Recall of different RAG baselines on ChartWalker Bench. The optimal performance is in bold, and the second-best performance is underlined.

Category	Methods	Overall		Fact Check		Manipulation		Analysis		Complex Reasoning	
		R@5	R@10	R@5	R@10	R@5	R@10	R@5	R@10	R@5	R@10
Plain RAG	BM25	37.55	49.80	37.50	57.64	53.13	66.31	38.22	51.24	15.93	20.06
	Text-EmBed	44.59	59.41	36.81	51.85	45.45	58.75	44.56	58.82	48.69	66.61
	VL-Embed	**53.90**	**71.87**	<u>49.07</u>	**73.15**	**61.82**	**79.91**	**52.00**	<u>70.52</u>	**51.09**	<u>63.64</u>
Graph-based RAG	RA-Local	40.19	50.51	37.27	48.38	47.46	58.75	44.28	52.69	23.61	36.44
	RA-Mix	42.38	54.19	33.10	45.60	42.97	54.26	45.11	54.13	41.68	59.89
	HippoRAG	<u>53.25</u>	<u>71.78</u>	**53.24**	<u>71.76</u>	<u>60.75</u>	<u>76.00</u>	<u>50.28</u>	**70.80**	<u>50.15</u>	**68.52**

ChartWalker agent and present analyses and insights on the cross-chart RAG task in §5.3.

5. Experiments

We evaluate (i) the retrieval quality of different RAG pipelines and (ii) the effectiveness of VLM generators in leveraging retrieved sources for correct answering.

5.1. Experimental Setup

Baselines. We compare against five representative retrievers, grouped into two families: plain RAG (by similarity) and graph-based RAG. Plain RAG involves BM25 (Robertson & Walker, 1994), Dense textual embedding (Zhang et al., 2025) and Vision-Language embedding (Li et al., 2026); while graph-based RAG involves RagAnything (local/mix) (Guo et al., 2025) and HippoRAG (Gutiérrez et al., 2025).

For VLM generators, we test three VLMs with different scales and openness: Qwen3-VL-8B, Qwen3-VL-32B (Bai et al., 2025a), and GPT-4o(OpenAI, 2024). Unless otherwise stated, we keep the prompting format fixed across the models to isolate the effect of model capability.

Evaluation Metrics To assess retrieval quality, we report Recall@k (R@k), which measures the fraction of annotated gold evidence sources that appear in the top-k retrieved list. To evaluate end-to-end success, we report Correctness@k (Cor@k), defined as the proportion of queries that the VLM generators answer correctly when conditioned on the top-k retrieved sources. Correctness is evaluated using an LLM-as-a-judge paradigm (Zheng et al., 2023). Unless otherwise specified, we report results at $k = 5$ and 10, and all metrics are reported as percentages (%).

Implementation Details. For the text-based RAG framework, we use Qwen3-VL-8B-Instruct (Bai et al., 2025a) to extract structured summaries and entities from each chart to build the multimodal database and knowledge graph. Dense embeddings are generated by Qwen3-Embedding-8B (Zhang et al., 2025) and Qwen3-VL-Embedding-8B (Li et al., 2026). All generator models use a fixed zero-shot prompt template and greedy decoding (temperature=0.0) for deterministic outputs. For agent training and correctness comparison 7, we use Qwen2.5-VL-3B-Instruct (Bai et al., 2025b) as the base policy model and VLM generator. A more detailed description and parameter settings of baselines can be found in Appendix A.2.

All experiments are conducted on a machine with Ubuntu 22.04 system, equipped with AMD EPYC 7742 64-Core Processor and 8× NVIDIA A100 GPUs (40GB memory). VL-Embedding is implemented in PyTorch version 2.8.0 with CUDA version 12.8 and Python 3.13.11, and others are all implemented in PyTorch version 2.9.1 with CUDA version 12.8 and Python 3.10.19.

5.2. Retrieval Performance

Table 3 reveals that retrieval remains far from saturated (best R@10 $\approx$ 72), confirming that ChartWalker-Bench is non-trivial under a limited top-k budget.

Enhanced multimodal representations outweigh complex retrieval heuristics. The table shows that the VL-Embedding retriever achieves the strongest average recall among all methods (53.90/71.87), and ranks first across most categories. This suggests that directly aligning query text with chart visuals in a shared embedding space substantially strengthens relevance matching.

Graph-aware retrieval provides clear benefits. HippoRAG consistently outperforms the text-only retriever by a large margin (Overall: +8.66 R@5 and +12.37 R@10), with especially pronounced gains on Fact Check and Manipulation, indicating that propagation over the knowledge graph effectively aggregates multi-hop evidence and recovers missing supporting sources that a single-pass embedding search tends to miss. Moreover, HippoRAG is particularly competitive on the more adversarial regimes: it attains the best R@10 on Analysis and Complex Reasoning, aligning with the intuition that graph-based relevance diffusion is helpful when queries are de-lexicalized and evidence is scattered across multiple charts.

Table 4. Entity extraction quality judged against chart images (VLM JUDGE = `gpt-5.4`). Metrics are averaged over extracted entities. N_{ent}: total extracted entities evaluated per dataset.

Dataset	P	R	Halluc.	Lvl.	N_{ent}	Parse err.
ChartQA	0.587	0.606	0.185	0.655	2,709	0
ChartMRAG	0.501	0.427	0.094	0.613	1,730	1

Notes. **Halluc.** = micro hallucination rate; **Lvl.** = micro level (granularity) accuracy. Macro chart-mean: ChartQA (P=0.633, R=0.593, Halluc.=0.109, Lvl.=0.649); ChartMRAG (P=0.495, R=0.425, Halluc.=0.089, Lvl.=0.602). ChartQA entity vs. relation runs use *different* 100-chart samples (see text).

Table 5. Relation triple quality (same judge). **Micro** metrics over evaluated triples; at most 40 relations sampled per chart. N_{rel}: total relation judgements.

Dataset	P	R	Halluc.	Type OK	N_{rel}	Parse err.
ChartQA	0.647	0.609	0.082	0.748	2,595	1
ChartMRAG	0.620	0.542	0.092	0.751	2,288	1

Notes. **Type OK** = micro type appropriateness. Macro chart-mean: ChartQA (P=0.631, R=0.576, Halluc.=0.082, Type=0.734); ChartMRAG (P=0.619, R=0.516, Halluc.=0.082, Type=0.741).

5.3. Generation Performance

Table 6 reports answer correctness under different retrievers and VLM generators. Overall, ChartWalker-Bench remains challenging at the generation stage: even with the strongest configuration, overall Cor@10 peaks at ∼65, while Complex Reasoning is consistently the hardest subset (best Cor@10 ≈ 51), indicating that multi-source retrieval and evidence composition are still major bottlenecks.

Stronger retrieval quality translates into substantial gains in correctness, but recall alone does not fully predict success. For Qwen3-VL-8B, moving from BM25 to dense/multimodal retrieval yields large jumps, and HippoRAG further improves to 45.74/59.40. A similar pattern holds for Qwen3-VL-32B. These results suggest that HippoRAG's graph-based relevance propagation provides more than "extra hits": even when its retrieval recall is not the highest, it tends to surface structurally connected evidence that completes multi-hop chains, which is easier for the generator to integrate and reason over the information.

Scaling generator improves both accuracy and robustness. Qwen3-VL-32B consistently outperforms Qwen3-VL-8B under the same retriever, and GPT-4o paired with VL-Embedding achieves the best overall scores, showing strong synergy between high-capacity VLMs and unified vision-language retrieval. Interestingly, even with GPT-4o, graph-based retrieval remains valuable for the most challenging reasoning regimes: HippoRAG becomes best at Cor@10 on Analysis (70.66) and Complex Reasoning (51.38), suggesting that structured, multi-hop evidence retrieval complements stronger generative reasoning rather than being replaced by it.

5.4. Agent Performance and Analysis

We evaluate ChartWalker-Agent on ChartWalker-Bench with an 8:2 split for training/testing, resulting in a held-out test set of 105 questions (Table 7). For lightweight VLMs (3B), increasing the retrieval budget does not monotonically improve correctness: feeding more charts quickly runs into multi-image context/visual token limitations and introduces distractors, so k=10 can be worse than k=5 (e.g., HippoRAG drops from 31.74 at k=5 to 29.79 at k=10). After PPO training, the agentic policy improves evidence acquisition via deeper KG exploration and achieves higher overall accuracy than static pipelines (33.14 vs. 31.74 best static), with the largest gains on the most search-intensive subset, Complex Reasoning.

6. Conclusion

In this paper, we introduced ChartWalker, a novel framework and benchmark for Cross-Chart RAG. By leveraging hierarchical knowledge graphs and structure-aware sampling, ChartWalker generates complex, multi-hop reasoning paths that challenge existing systems. Our evaluations on ChartWalker-Bench reveal that current Vision-Language Models struggle with multi-chart analysis, exposing the limitations of static retrieval. To bridge this gap, we proposed ChartWalker-Agent, demonstrating the power of iterative, graph-based evidence acquisition. This work provides a rigorous foundation for advancing multimodal RAG, with future efforts directed toward enhancing multimodal embeddings and agentic reasoning for complex chart analysis.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal

Table 6. Answer correctness under different VLM generators on ChartWalker Bench.

VLM	Methods	Overall		FactCheck		Manipulation		Analysis		Complex Reasoning	
		Cor@5	Cor@10	Cor@5	Cor@10	Cor@5	Cor@10	Cor@5	Cor@10	Cor@5	Cor@10
Qwen3-VL-8B	BM25	18.44	22.69	25.00	33.33	15.60	18.44	25.62	29.75	1.83	5.50
	Text-EmBediing	26.59	35.46	36.11	41.67	24.11	32.62	33.05	44.63	9.17	14.68
	VL-Embedding	42.02	49.65	59.72	72.22	44.68	56.03	48.76	55.37	11.93	13.76
	Local	27.84	35.11	37.50	50.00	24.11	29.79	36.77	41.74	6.42	17.43
	Mix	37.94	41.14	47.22	48.61	38.30	42.55	46.28	48.76	12.84	17.43
	HippoRAG	**45.74**	**59.40**	**56.94**	**76.39**	**46.10**	**60.28**	**51.65**	**63.64**	**24.77**	**37.61**
Qwen3-VL-32B	BM25	20.74	25.36	22.22	38.88	17.02	22.70	27.27	29.34	10.09	11.01
	Text-EmBediing	26.59	37.06	34.72	45.83	19.15	34.04	32.23	44.62	18.35	18.35
	VL-Embedding	48.58	59.4	63.89	79.17	50.35	**68.79**	**54.13**	62.81	23.85	26.61
	Local	28.37	37.05	38.88	51.39	23.40	33.33	36.36	44.21	10.09	16.51
	Mix	40.42	45.39	55.55	55.55	36.17	41.13	48.76	53.72	17.43	25.69
	HippoRAG	**50.18**	**64.72**	**68.06**	**77.78**	**51.06**	66.67	52.07	**68.60**	**33.03**	**44.95**
GPT-4o	BM25	29.79	33.87	38.88	47.22	22.69	22.70	36.36	38.43	18.35	29.36
	Text-EmBediing	26.77	36.52	31.94	40.28	22.69	30.50	32.23	44.21	16.51	24.77
	VL-Embedding	**56.03**	**64.89**	**66.67**	**76.39**	**49.65**	**66.67**	**61.57**	67.36	**44.95**	49.54
	Local	27.68	35.44	31.94	43.86	24.29	34.48	36.40	45.77	10.09	10.28
	Mix	42.93	46.31	54.29	66.67	35.00	39.47	54.58	57.45	20.18	27.52
	HippoRAG	49.82	63.30	59.72	70.83	46.81	56.03	55.37	**70.66**	34.86	**51.38**

Table 7. Answer correctness comparison between HippoRAG and Agent on ChartWalker Bench.

Method	Overall	FactCheck	Manipulation	Analysis	Complex
VL-Emb(k=5)	23.76	34.72	20.56	28.51	10.09
VL-Emb(k=10)	25.89	37.50	20.57	32.23	11.01
HippoRAG(k=5)	31.74	44.44	29.79	36.36	15.60
HippoRAG(k=10)	29.79	37.50	29.79	34.30	14.68
Agent(3b)	33.14	37.50	34.37	34.08	18.75

consequences of our work, none which we feel must be specifically highlighted here.

References

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., and Zhu, K. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., and Lin, J. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Brin, S. and Page, L. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1):107–117, 1998. ISSN 0169-7552. doi: https://doi.org/10.1016/S0169-7552(98)00110-X. URL https://www.sciencedirect.com/science/article/pii/S016975529800110X. Proceedings of the Seventh International World Wide Web Conference.

Cheng, K., Lin, G., Fei, H., zhai, Y., Yu, L., Ali, M. A., Hu, L., and Wang, D. Multi-hop question answering under temporal knowledge editing, 2024. URL https://arxiv.org/abs/2404.00492.

Das, R. and Soylu, M. A key review on graph data science: The power of graphs in scientific studies. *Chemometrics and Intelligent Laboratory Systems*, 240:104896, 2023. ISSN 0169-7439. doi: https://doi.org/10.1016/j.chemolab.2023.104896. URL https://www.sciencedirect.com/science/article/pii/S0169743923001466.

Geng, X., Xia, P., Zhang, Z., Wang, X., Wang, Q., Ding, R., Wang, C., Wu, J., Zhao, Y., Li, K., Jiang, Y., Xie, P., Huang, F., and Zhou, J. Webwatcher: Breaking new frontier of vision-language deep research agent, 2025. URL https://arxiv.org/abs/2508.05748.

Guo, Z., Ren, X., Xu, L., Zhang, J., and Huang, C. Rag-anything: All-in-one rag framework, 2025. URL https://arxiv.org/abs/2510.12323.

Gutiérrez, B. J., Shu, Y., Qi, W., Zhou, S., and Su, Y. From rag to memory: Non-parametric continual learning for large language models, 2025. URL https://arxiv.org/abs/2502.14802.

Han, H., Wang, Y., Shomer, H., Guo, K., Ding, J., Lei, Y., Halappanavar, M., Rossi, R. A., Mukherjee, S., Tang, X., He, Q., Hua, Z., Long, B., Zhao, T., Shah, N., Javari, A., Xia, Y., and Tang, J. Retrieval-augmented generation with graphs (graphrag), 2025. URL https://arxiv.org/abs/2501.00309.

Herzig, J., Müller, T., Krichene, S., and Eisenschlos, J. M. Open domain question answering over tables via dense retrieval, 2021. URL https://arxiv.org/abs/2103.12011.

Kastellec, J. P. and Leoni, E. L. Using graphs instead of tables in political science. *Perspectives on Politics*, 5(4): 755–771, 2007. doi: 10.1017/S1537592707072209.

Kim, D. H., Hoque, E., and Agrawala, M. Answering questions about charts and generating visual explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pp. 1–13, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376467. URL https://doi.org/10.1145/3313831.3376467.

Kumar, V., Hua, Y., Ramakrishnan, G., Qi, G., Gao, L., and Li, Y.-F. Difficulty-controllable multi-hop question generation from knowledge graphs. In *The Semantic Web – ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part I*, pp. 382–398, Berlin, Heidelberg, 2019. Springer-Verlag. ISBN 978-3-030-30792-9. doi: 10.1007/978-3-030-30793-6_22. URL https://doi.org/10.1007/978-3-030-30793-6_22.

Kweon, S., Kwon, Y., Cho, S., Jo, Y., and Choi, E. Open-wikitable: Dataset for open domain question answering with complex reasoning over table, 2023. URL https://arxiv.org/abs/2305.07288.

Li, M., Zhang, Y., Long, D., Keqin, C., Song, S., Bai, S., Yang, Z., Xie, P., Yang, A., Liu, D., Zhou, J., and Lin, J. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026.

Li, Z., Du, Y., Zheng, M., and Song, M. Mimotable: A multi-scale spreadsheet benchmark with meta operations for table reasoning, 2024. URL https://arxiv.org/abs/2412.11711.

Liu, S., Xie, X., Siow, J., Ma, L., Meng, G., and Liu, Y. Graphsearchnet: Enhancing gnns via capturing global dependencies for semantic code search, 2023. URL https://arxiv.org/abs/2111.02671.

Lu, R., Hou, Z., Wang, Z., Zhang, H., Liu, X., Li, Y., Feng, S., Tang, J., and Dong, Y. Deepdive: Advancing deep search agents with knowledge graphs and multi-turn rl, 2025. URL https://arxiv.org/abs/2509.10446.

Masry, A., Islam, M. S., Ahmed, M., Bajaj, A., Kabir, F., Kartha, A., Laskar, M. T. R., Rahman, M., Rahman, S., Shahmohammadi, M., et al. Chartqapro: A more diverse and challenging benchmark for chart question answering. *arXiv preprint arXiv:2504.05506*, 2025.

Mavi, V., Jangra, A., and Jatowt, A. Multi-hop question answering, 2024. URL https://arxiv.org/abs/2204.09140.

Norasaed, W. and Siriborvornratanakul, T. Market movement prediction using chart patterns and attention mechanism. *Discover Analytics*, 2(1), 2024. doi: 10.1007/s44257-023-00007-6. URL https://doi.org/10.1007/s44257-023-00007-6.

OpenAI. Hello GPT-4. https://openai.com/index/hello-gpt-4o/, 2024.

Pasupat, P. and Liang, P. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1470–1480, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1142. URL https://aclanthology.org/P15-1142.

Robertson, S. E. and Walker, S. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In *Proceedings of the 17th International ACM SIGIR Conference on Research and Development in Information Retrieval (Special Issue of the SIGIR Forum)*, pp. 232–241. Springer-Verlag, 1994. ISBN 3-540-19889-X.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Shi, Y., Sun, M., Liu, Z., Yang, M., Fang, Y., Sun, T., and Gu, X. Reasoning in trees: Improving retrieval-augmented generation for multi-hop question answering, 2026. URL https://arxiv.org/abs/2601.11255.

Singh, A., Biemann, C., and Strich, J. Mtabvqa: Evaluating multi-tabular reasoning of language models in visual space, 2025a. URL https://arxiv.org/abs/2506.11684.

Singh, A., Ehtesham, A., Kumar, S., and Khoei, T. T. Agentic retrieval-augmented generation: A survey on agentic rag, 2025b. URL https://arxiv.org/abs/2501.09136.

Wang*, K., Zhang*, P., Wang*, Z., Gao*, Y., Li*, L., Wang, Q., Chen, H., Wan, C., Lu, Y., Yang, Z., Wang, L., Krishna, R., Wu, J., Fei-Fei, L., Choi, Y., and Li, M. Vagen:reinforcing world model reasoning for multi-turn vlm agents, 2025. URL https://vagen-ai.github.io/.

Wang, Y., Liu, J., Tang, C., Yan, L., and Jiang, J. KCS: Diversify multi-hop question generation with knowledge composition sampling. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 23173–23185, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1181. URL https://aclanthology.org/2025.emnlp-main.1181/.

Xie, T., Lin, M., Liu, M., Ye, Y., Chen, C., and Liu, S. Infochartqa: A benchmark for multimodal question answering on infographic charts, 2025. URL https://arxiv.org/abs/2505.19028.

Yang, Y., Zhong, J., Jin, L., Huang, J., Gao, J., Liu, Q., Bai, Y., Zhang, J., Jiang, R., and Wei, K. Benchmarking multimodal rag through a chart-based document question-answering generation framework. *arXiv preprint arXiv:2502.14864*, 2025.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., and Manning, C. D. Hotpotqa: A dataset for diverse, explainable multi-hop question answering, 2018. URL https://arxiv.org/abs/1809.09600.

Yu, X., Jian, P., and Chen, C. Tablerag: A retrieval augmented generation framework for heterogeneous document reasoning, 2025. URL https://arxiv.org/abs/2506.10380.

Zhang, X., Shou, L., Pei, J., Gong, M., Wen, L., and Jiang, D. A graph representation of semi-structured data for web question answering, 2020. URL https://arxiv.org/abs/2010.06801.

Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., Huang, F., and Zhou, J. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

Zhong, V., Xiong, C., and Socher, R. Seq2sql: Generating structured queries from natural language using reinforcement learning, 2017. URL https://arxiv.org/abs/1709.00103.

Zou, J., Fu, D., Chen, S., He, X., Li, Z., Zhu, Y., Han, J., and He, J. Rag over tables: Hierarchical memory index, multi-stage retrieval, and benchmarking, 2025. URL https://arxiv.org/abs/2504.01346.

A. Appendix.

A.1. Notations

Table 8: Summary of Notations and Symbols used in ChartWalker.

Symbol	Description
Problem Formulation	
$\mathcal{C}, c$	The corpus of charts, chart.
q	Natural language query
$\mathcal{C}_k()$	The set of top-k retrieved charts.
$y, \hat{y}$	The ground-truth answer and The predicted answer generated by the reader model.
Hierarchical Knowledge Graph	
$\mathcal{G}_c, \mathcal{V}_c, \mathcal{E}_c$	The chart hierarchical knowledge graph, Entities and Edges for chart c
l	Granularity level of an entity.
$\mathcal{G}$	The global knowledge graph integrated from all chart subgraphs.
Path Sampling & QA Synthesis	
M	The transition matrix for PageRank calculation, $M = [M_{u,v}]$.
$\mathcal{N}(v)$	The set of neighbor entities of entity v.
$N_{src}(u)$	Number of distinct source charts associated with entity u.
π	The stationary distribution (PageRank score) used for anchor selection.
$\mathcal{P}_T$	A sampled reasoning path of length T.
ϕ_{sem}	Semantic topic coherence function (cosine similarity).
ϕ_{gran}	Granularity control function for regulating level transitions.

A.2. More experiment settings

Baseline Details (1) BM25 (Robertson & Walker, 1994): a sparse lexical retriever that ranks candidates by term-matching scores. (2) Dense textual embedding (Zhang et al., 2025): we convert each candidate visual source into text, embed both query and candidates, and rank by cosine similarity. (3)Vision-Language embedding (Li et al., 2026): unify textual and visual features into one embedding space and compute the direct similarity. (4) RagAnything (local/mix) (Guo et al., 2025): graphrag supporting multimodal retrieval. (5) HippoRAG(Gutiérrez et al., 2025): a neuro-inspired graph retriever traverses the graph with Personalized PageRank.

Retrieval hyperparameters are as follows: BM25 uses standard Okapi parameters (k1=1.5, b=0.75).RagAnything (Local) retrieves the top-10 seed entities with one hop expansion. RagAnything (Mix) starts with top-10 seed candidates, then expands to the top-8 neighbors per seed. HippoRAG uses its hierarchical mechanism with Personalized PageRank (damping factor=0.5) to propagate relevance and enable multi-hop traversal to relevant passages.

A.3. Prompt Templates

Hierarchical Entity Extraction Prompt

Task Overview. You are given a chart (table, figure, or a combination of both), its caption, and optional surrounding context. Your task is to construct a **hierarchical knowledge graph** that represents the semantic structure of the chart.
Overall Objective. Your goal is to build a **layered knowledge graph**:

- **Level 1 entities** represent the chart's core topic and primary comparison dimensions, derived from the caption and top-level headers.
- **Level 2 entities** represent secondary dimensions or detailed subcategories, such as rows, subheaders, or nested categories.
- **Level 3 and beyond** represent supporting attributes at a finer level of granularity.

The number of levels should be **adaptively determined** by the chart structure. If a level can be decomposed further (e.g., via multi-level headers), you must recursively create deeper levels.
Entity Extraction Guidelines.

1. Extract **semantic-rich and disambiguated entities**.
    - Level 1: Theme entities and key dimensions.
    - Level 2: Row or column entities specifying subcategories.
    - Level 3+: Fine-grained attributes.
    - Always include qualifiers for uniqueness (e.g., "Revenue (2023)").
    - All entities must belong to the required entity types `{entity_types}`.
2. Skip generic or referential phrases (e.g., "this table", "the data") and presentation-only text (e.g., "see below").
3. Provide concise and meaningful descriptions for each entity.

Relationship Guidelines.

1. Extract **intra-level** relationships, especially among Level 1 entities.
2. Extract **inter-level** relationships linking higher-level entities to their children.
3. All relationships must belong to the required types `{relation_types}`.
4. Each relationship must include:
    - source_entity and target_entity,
    - relation_type,
    - relationship_description,
    - relationship_keywords.

Output Format. Return a JSON object with two fields: `entities` and `relationships`.
Additional Constraints.

- Do not extract numerical data points.
- Ensure all relationships refer to extracted entities.
- If the input chart is too short or lacks sufficient detail, return an empty JSON object.

Return **only valid JSON**. Do not include explanations, markdown, or code blocks. The output must start with `{` and end with `}`.
Chart Input. The chart image or text is provided as `{input_text}`.

Unified Hierarchical Entity Extraction & QA Generation Prompt

SYSTEM CONTEXT: You are designing a high-quality multi-chart QA pair. **Inputs provided:** `reasoning_paths`, `available_sources`, `chart_index`, `text_evidence`.

UNIVERSAL CONSTRAINTS (Apply to ALL tasks):

- **Self-contained:** No deictic references (e.g., "this chart"). Include scope qualifiers.
- **Multi-source:** Must combine evidence from at least TWO sources (Chart+Chart or Chart+Text).
- **CRITICAL: Do NOT reveal specific numbers/percentages in the question text.**

Select ONE Strategy Module based on Task Type:

Module A: QA_FactCheck

Goal: Verify a claim using evidence.

- **Question:** Must present a **CLAIM** (e.g., "Is it true that...").
- **Answer:** Must be binary ("True"/"False", "Yes"/"No") + Evidence.

Module B: QA_Manipulation

Goal: Require arithmetic calculation.

- **Question:** Ask for difference, sum, ratio, or growth rate.
- **Answer:** Specific numerical value derived from calculation.

Module C: QA_Complex_Reasoning (Stealth)

Goal: Low lexical overlap (Hard Retrieval).

- **Requirement: MANDATORY PARAPHRASING**.
- **Forbidden:** No Entity Names (use descriptions), No Years (use relative time), No Keywords (e.g., "increase").

Module D: QA_Analysis / QA_Trend

Goal: High-level pattern recognition.

- **Question:** Ask about **Logical Relationships** (correlations) or **Temporal Trends** (peaks, fluctuations).
- **Logic:** Synthesize separate observations into a conclusion.

UNIFIED OUTPUT FORMAT (Strict JSON):

```
{
  "query_type": "FactCheck" | "Manipulation" | "Analysis" | "Trend",
  "question": "...",          // Follows Module constraints above
  "answer": "...",            // Includes specific numbers/calcs/verdict
  "explanation": "...",       // References reasoning edges
  "entities_used": ["..."],   // List of entity IDs
  "evidence": ["..."],        // List of chunk IDs
  "difficulty_subjective": 1-3
}
```

Figure 3. Unified Prompt Template for Multi-Chart QA Generation. The system shares a common context and output format, but branches into four distinct modules (A-D) with specific logic, constraints, and paraphrasing requirements depending on the desired query type.

A.4. Agent Environment

Following the multi-turn VLM-agent training paradigm, we model visual search on the global chart KG as a partially observable Markov decision process (POMDP) $\langle \mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{R} \rangle$. At turn t, the agent receives an observation $o_t \in \mathcal{O}$ (the current KG context and retrieved evidence), generates an action $a_t \in \mathcal{A}$ to further explore the KG, and the environment returns a scalar reward $r_t = \mathcal{R}(s_t, a_t)$. The objective is to maximize the expected discounted return $\max_\theta \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T-1} \gamma^t r_t\right]$. In

In the standard *concatenation* rollout, the context grows with turns, $c_t^{\text{concat}} = [\text{sys}, o_0, y_0, \ldots, o_t]$, which quickly exceeds the VLM context window and induces high-variance credit assignment over an ultra-long token axis. We instead adopt a *non-concatenated* rollout: at each turn t the policy conditions only on the current prompt context

$$c_t = [\text{sys}, o_t], \tag{8}$$

and autoregressively generates a response/action token sequence $y_t = (y_{t,1}, \ldots, y_{t,K_t})$ with

$$\pi_\theta(y_t \mid c_t) = \prod_{j=1}^{K_t} \pi_\theta(y_{t,j} \mid c_t, y_{t,<j}). \tag{9}$$

Accordingly, PPO is computed *per turn* without requiring a forward pass over the entire concatenated trajectory. Let

$$u_{t,j}(\theta) = \frac{\pi_\theta(y_{t,j} \mid c_t, y_{t,<j})}{\pi_{\text{old}}(y_{t,j} \mid c_t, y_{t,<j})}, \tag{10}$$

and let $m_{t,j}^{\text{act}} \in \{0, 1\}$ mask response/action tokens. The PPO objective is

$$J_{\text{PPO}}(\theta) = \frac{1}{\sum_{t,j} m_{t,j}^{\text{act}}} \sum_{t,j} m_{t,j}^{\text{act}} \min\Big(u_{t,j}(\theta)A_t,\ \text{clip}(u_{t,j}(\theta), 1-\epsilon, 1+\epsilon)A_t\Big), \tag{11}$$

where the advantage A_t is defined at the *turn* level (below) and broadcast to tokens in the same turn.

Turn-level GAE and broadcasting. We learn a critic $V_\phi(c_t)$ defined on the turn context c_t. Given turn rewards $\{r_t\}_{t=1}^T$, we compute TD residuals and GAE over turns:

$$\delta_t = r_t + \gamma V_\phi(c_{t+1}) - V_\phi(c_t), \tag{12}$$

$$A_t = \delta_t + \gamma\lambda A_{t+1}, \tag{13}$$

with $V_\phi(c_{T+1}) = 0$ for terminal. We then assign each response token in turn t the same advantage:

$$A_{t,j} = m_{t,j}^{\text{act}} A_t. \tag{14}$$

Value regression at turn boundaries. Let the turn return target be $G_t = A_t + V_\phi(c_t)$ (stop-gradient on the RHS). We regress the critic only once per turn using a value-mask $m_t^{\text{val}} = 1$:

$$L_V(\phi) = \frac{1}{\sum_t m_t^{\text{val}}} \sum_t m_t^{\text{val}} \big(V_\phi(c_t) - G_t\big)^2. \tag{15}$$

In implementation, $V_\phi(c_t)$ can be read from a designated anchor position (e.g., the first response token) while the objective remains a turn-level value function.

our setting, the policy π_θ is parameterized by a VLM that This design yields stable bootstrapping targets while keeping advantage estimation aligned with turn-level interaction dynamics, mitigating the instability induced by ultra-long concatenated contexts.

ChartWalker Agent: Navigation & Reasoning System Prompt

SYSTEM CONTEXT: You are an intelligent agent navigating a multi-modal, multi-level Knowledge Graph (MMKG).
Goal: Explore the graph to find the correct answer for a given query.

CORE OBJECTIVES:

1. **Identify** required information for the query.
2. **Navigate** to entities providing this info.
3. **Extract** concrete facts/numerical values from evidence.
4. **Stop** once sufficient info is collected.

Note: Exploration actions that do not fill missing information are discouraged.

Interaction Phase Templates:

Phase 1: Initialization (Start)

Input: `[Initial Graph Observation], Candidate Start Entities.`

- **Context:** Choose an entry point.
- **Valid Action:** `start`
- **Grammar:** Must choose exactly ONE entity name from the candidate list.

Example Output:
`<answer>start CountryX</answer>`

Phase 3: Termination (Stop)

Input: Collected evidence, Observation.

- **Condition:** Collected enough info OR reached max stops.
- **Valid Action:** `stop`
- **Requirement:** Output final answer text immediately after operator.

Example Output:
`<answer>stop The answer is Paris</answer>`

Phase 2: Navigation & Reasoning Loop (Intermediate Steps)

Input: `[Current Graph State]` (Entity desc, Relations, Memory, Visited sources).
Reasoning: Decide next action based on context/history. **WARNING: DO NOT REPEAT ACTIONS OR SEARCH FORBIDDEN RELATIONS.**

Action Grammar Options:

- **`edge_search <int>`**: Inspect a relation index from `[Searchable Relations]`.
  Ex: `<answer>edge_search 1</answer>`
- **`move <int>`**: Move to a relation's target entity (to get more info). Targets can be from Searchable OR Forbidden lists.
  Ex: `<answer>move 1</answer>`
- **`backward <Name>`**: Return to a previously visited entity.
  Ex: `<answer>backward CountryY</answer>`

UNIFIED OUTPUT FORMAT: All responses must be wrapped in XML-style tags.

`<answer> [action_keyword] [arguments] </answer>`

Figure 4. The ChartWalker Agent Prompt Structure. The prompt guides the agent through three distinct phases: (1) Selecting a start entity, (2) An iterative navigation loop involving edge searching and entity traversal, and (3) A termination phase to output the final answer.

A.5. Showcase

 # Manipulation

KG Path

Democrats/leaning Democrats (L1) –horizontal [src=chart_00189_07]–> Global Climate Change (L1)
Global Climate Change (L1) –vertical_up [src=chart_00048_01]–> Survey (L0)
Democrats/leaning Democrats (L1) –vertical_down [src=chart_00189_07]–> Catholic Democrats

Query

In 2022, what percentage of U.S. Catholic adults who identify as Democrats or leaning Democratic viewed global climate change as an extremely or very serious problem, and how does this compare to the percentage of all U.S. adults who identify as Democrats or leaning Democratic who considered dealing with global climate change a top priority for the president and Congress in 2024?

Answer

82% in 2022 among Catholic Democrats, which is 23 percentage points higher than the 59% in 2024 among all Democrats.

Evidence

chart_00189_07 (2022, Catholic Dem/lean Dem)

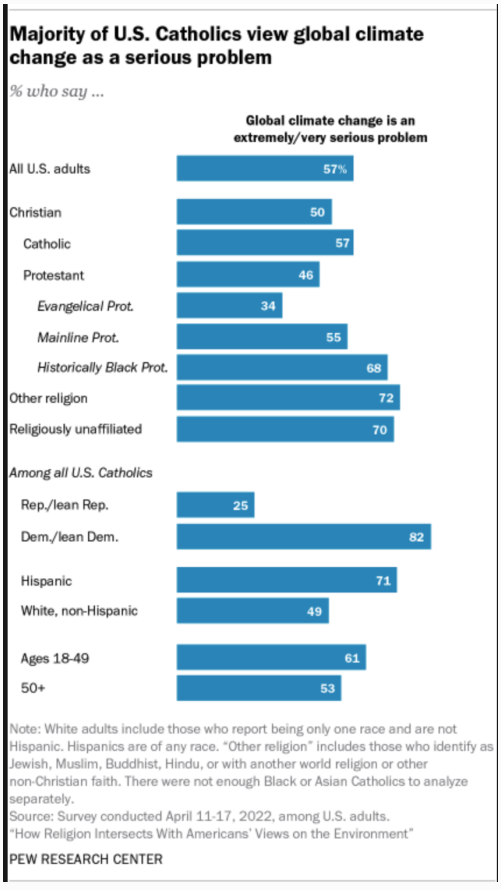


chart_00048_01 (2024, all Dem/lean Dem)

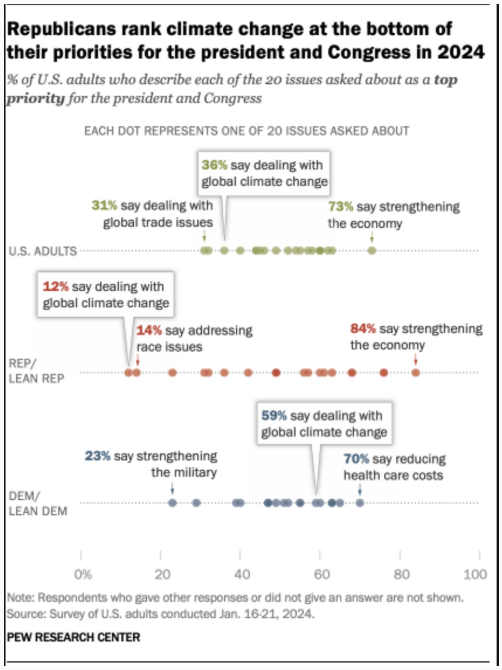


Figure 5. Manipulation showcase. Due to the relation of Democrats and Catholic Democrats being highlighted in the sampled path, our problem does not contain the logical inconsistency shown in (Lu et al., 2025) (Figure 6).

```
# Intra-Document_Text-only_00571

## Query
In Japan, what percentage of adults say their family owns a cemetery? Among
these adults, how many of those with no religious affiliation have made
offerings to their ancestors in the past 12 months?

## Answer
85% of Japanese adults claim that their family owns a cemetery. Among these
adults, 59% of those with no religious affiliation have prepared food, water, or
beverages for their ancestors in the past 12 months.

## Source

### Source 1
"id": "paragraph_00128_01"
"text": "People in Japan are preparing to celebrate Obon - a festival devoted to
celebrating ancestors that features lighting lanterns and maintaining family
gravesites. In Japan, 85% of adults say their family has such a gravesite, and
79% say they have looked after this gravesite by sweeping or cleaning it in the
past year, according to a recent Pew Research Center survey. Obon …"

### Source 2
"id": "paragraph_00128_02"
"text":" We also asked survey respondents … For instance, 59% of Japan's
religiously unaffiliated adults say they have offered food, water or drinks in
the past 12 months to care for their ancestors. Christians generally are less
likely to engage in these sorts of activities. However, many Vietnamese
Christians have burned incense, offered flowers or lit candles to care for
ancestors in the last year. "

## keypoints
"paragraph_00128_01_02": "85% of adults in Japan say their family has a
gravesite",

"paragraph_00128_02_07": "59% of Japan's religiously unaffiliated adults have
offered food, water, or drinks in the past 12 months to care for their
ancestors"
```

Figure 6. Wrong case

 # Complex_Reasoning

KG Path

USA (L1) –horizontal [src=test_000774]–> Coal (33%)
USA (L1) –horizontal [src=test_000774]–> Natural gas (33%)
SocialMediaPlatform (L1) –horizontal [src=test_001503]–> Clean & renewable energy (35%)

Query

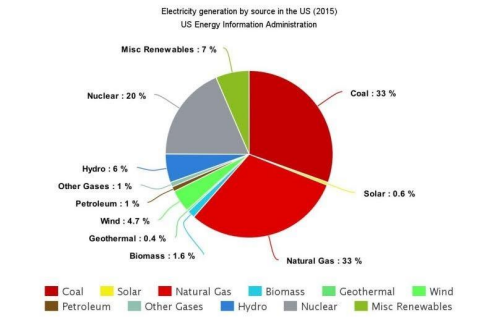
For the same year (2015), let G_{top2} be the set of the top-2 energy sources by share in the national electricity generation mix, and let P_{top1} be the top-1 energy source by share in the social media platform's operations energy consumption breakdown. Check whether $P_{\text{top1}} \in G_{\text{top2}}$.

Answer

No. In 2015, $G_{\text{top2}} = \{\text{Coal},\ \text{Natural gas}\}$, while $P_{\text{top1}} = \text{Clean \& renewable energy}$, so $P_{\text{top1}} \notin G_{\text{top2}}$.

Evidence

test_000774 (2015, USA electricity generation mix)



test_000699 (supporting chart / context)

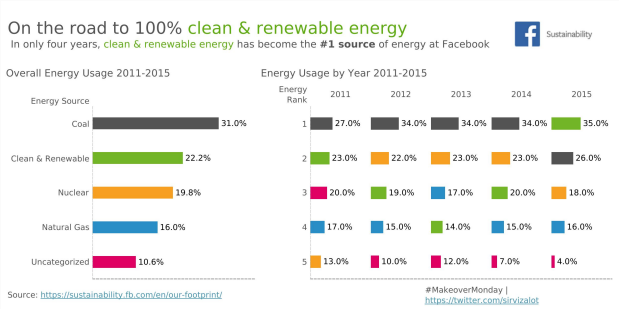


Figure 7. Complex Reasoning showcase.

 # FactCheck

KG Path

Peru (L1) –horizontal [src=chart_00260_02]–> Favorable opinion of Brazil (58%)
Peru (L1) –horizontal [src=chart_00260_04]–> Confidence in Brazil's president (30%)
Spring 2024 Pew surveys (L1) –horizontal [src=paragraph_00260_01]–> Covered countries include Peru

Query

In the spring 2024 surveys, does the share of people in Peru with a favorable opinion of Brazil exceed the share expressing confidence in Brazil's president?

Answer

Yes (58% vs 30%).

chart_00260_02 (Peru favorable opinion of Brazil, Spring 2024)

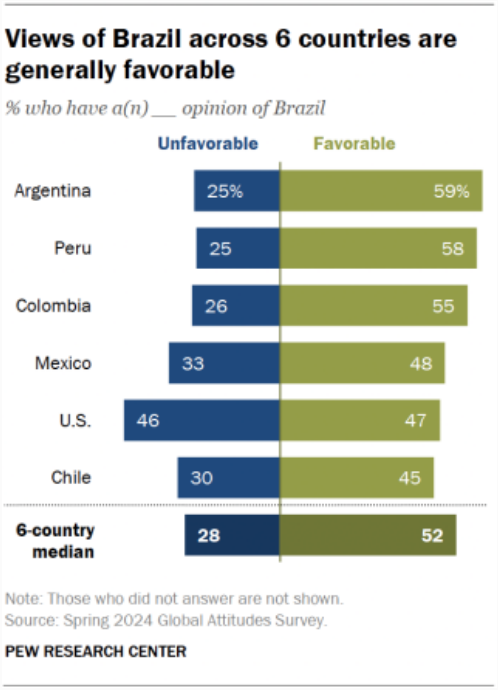
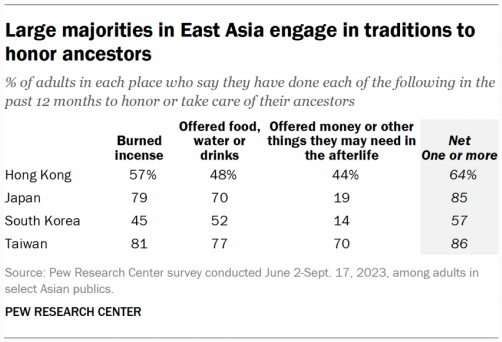


chart_00261_04 (Peru confidence in Lula, Spring 2024)



Large majorities in East Asia engage in traditions to honor ancestors

% of adults in each place who say they have done each of the following in the past 12 months to honor or take care of their ancestors

	Burned incense	Offered food, water or drinks	Offered money or other things they may need in the afterlife	*Net One or more*
Hong Kong	57%	48%	44%	64%
Japan	79	70	19	85
South Korea	45	52	14	57
Taiwan	81	77	70	86

Source: Pew Research Center survey conducted June 2-Sept. 17, 2023, among adults in select Asian publics.
PEW RESEARCH CENTER

Figure 8. Factcheck showcase.

Analysis

KG Path

Israel (L1) –horizontal [src=chart_00076_03]–> EU favorability, Ideological left (85%)
Israel (L1) –horizontal [src=chart_00076_03]–> EU favorability, Ideological right (49%)
Israel (L1) –horizontal [src=chart_00235_04]–> Biden approval/confidence, Ideological left (48%)
Israel (L1) –horizontal [src=chart_00235_04]–> Biden approval/confidence, Ideological right (61%)

Query

Among Israeli adults, does a more favorable view of the European Union correlate with lower approval of U.S. President Biden?

Answer

Yes — Israeli adults on the ideological left show higher EU favorability (85%) but lower Biden approval/confidence (48%), while the ideological right shows lower EU favorability (49%) but higher Biden approval/confidence (61%).

Evidence

chart_00076_03 (Israel, EU favorability by ideology)

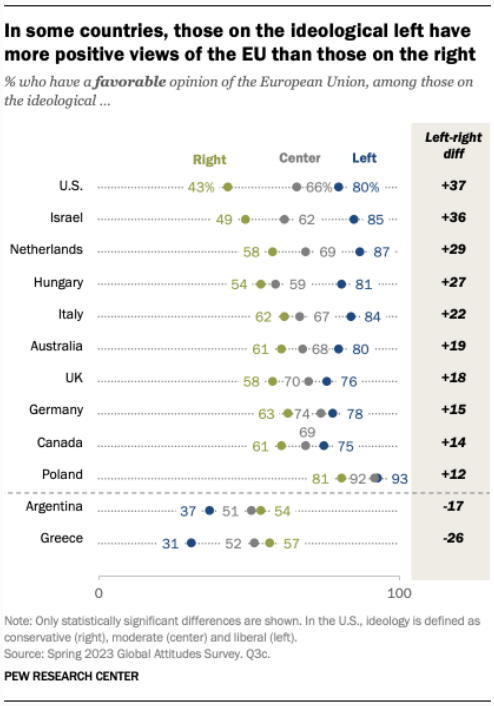


chart_00235_04 (Israel, Biden approval/confidence by ideology)

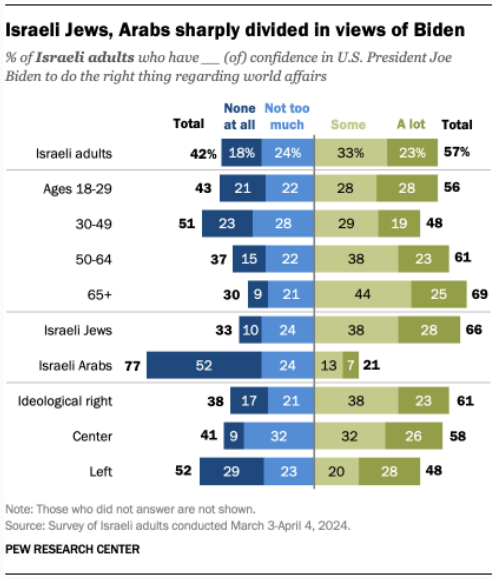


Figure 9. Analysis showcase