Show Me the Infographic I Imagine: Intent-Aware Infographic Retrieval for Authoring Support

Jing Xu, Jiarui Hu, Zhihao Shuai, Yiyun Chen, and Weikai Yang

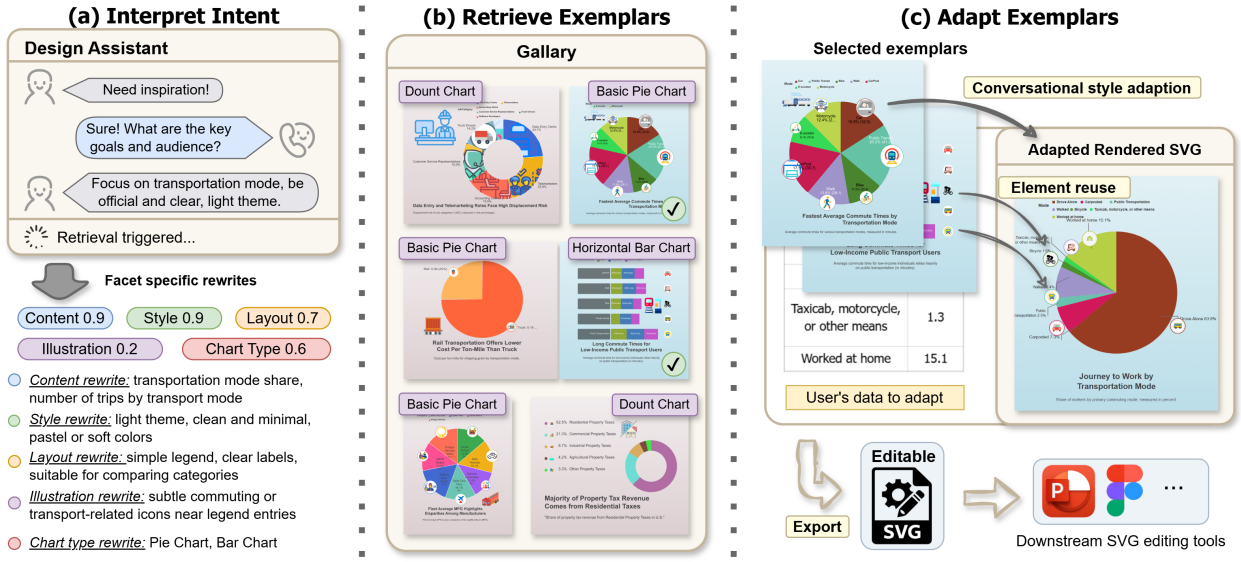

Fig. 1: Overview of our intent-aware retrieve-and-adapt workflow for infographic authoring. (a) A natural-language query is parsed into weighted intent facets; (b) These facet-specific cues guide exemplar retrieval over an infographic corpus; (c) The selected exemplar is adapted to the user's data, producing a rendered result together with an editable SVG that can be exported to downstream design tools.

Abstract—While infographics have become a powerful medium for communicating data-driven stories, authoring them from scratch remains challenging, especially for novice users. Retrieving relevant exemplars from a large corpus can provide design inspiration and promote reuse, substantially lowering the barrier to infographic authoring. However, effective retrieval is difficult because users often express design intent in ambiguous natural language, while infographics embody rich and multi-faceted visual designs. As a result, keyword-based search often fails to capture design intent, and general-purpose vision-language retrieval models trained on natural images are ill-suited to the text-heavy, multi-component nature of infographics. To address these challenges, we develop an intent-aware infographic retrieval framework that better aligns user queries with infographic designs. We first conduct a formative study of how people describe infographics and derive an intent taxonomy spanning content and visual design facets. This taxonomy is then leveraged to enrich and refine free-form user queries, guiding the retrieval process with intent-specific cues. Building on the retrieved exemplars, users can adapt the designs to their own data with high-level edit intents, supported by an interactive agent that performs low-level adaptation. Both quantitative evaluations and user studies are conducted to demonstrate that our method improves retrieval quality over baseline methods while better supporting intent satisfaction and efficient infographic authoring.

Index Terms—Infographic retrieval, infographic authoring, design intent, interactive design adaptation

1 Introduction

Infographics are a widely used medium for communicating data-driven stories by combining data, text, and visual elements into a compact narrative [4, 7, 17]. Yet authoring a high-quality infographic remains challenging, as authors must coordinate multiple design decisions simultaneously, including narrative flow, layout composition, visual style, and illustration usage. In practice, both novice users and expert designers often start from existing designs by reusing templates or borrowing design patterns to reduce effort and improve quality [13]. This motivates exemplar retrieval as a practical authoring aid: given a user's design intent, relevant exemplars that match not only the topic but also design intent should be retrieved.

However, existing retrieval methods often fail to satisfy users' design intent for infographics. Traditional keyword-based search engines [1, 28] primarily optimize for topic matching and provide limited support for expressing and enforcing structural or stylistic constraints. Meanwhile, embedding-based text–image retrieval models [11, 26] typically collapse relevance into a single similarity, which offers limited control over which intent facet (e.g., layout composition or visual style) should dominate the retrieval. As a result, models frequently return exemplars that are semantically related but misaligned with the desired layout or style. These limitations motivate an intent-aware retrieval method that can better capture and enforce multiple design intent facets.

To characterize this mismatch, we first conducted a formative study to examine how people describe desired infographic exemplars in free-form natural language. Our analysis showed that queries commonly blend multiple intent facets, including content, chart type, layout, illustration, and style. Guided by these findings, we propose an intent-

- *Jing Xu, Zhihao Shuai, Yiyun Chen, and Weikai Yang are with The Hong Kong University of Science and Technology (Guangzhou).*
- *Jiarui Hu is with Nanjing University.*
- *Corresponding author: Weikai Yang.*

aware retrieval framework that represents a user's requirement as a weighted combination of these intent facets. Given a free-form query, the system generates facet-specific rewrites, estimates facet weights, and performs weighted multi-facet matching in a shared text–image embedding space. To better support infographic retrieval, we additionally apply a lightweight embedding alignment technique to better capture facet-specific query–infographic similarity.

Beyond retrieval, we present an exemplar-driven authoring workflow that helps users reuse and adapt retrieved designs to their own data. The workflow is supported by an interactive conversational agent that translates high-level edit intents into concrete modifications of the exemplar representation (e.g., SVG-based edits), thereby connecting inspiration seeking with iterative adaptation [41].

We evaluate our method through single-round and multi-round retrieval benchmarks, as well as an end-to-end authoring user study. The results indicate that our method improves retrieval quality over baseline techniques and better supports intent satisfaction during authoring. The contributions of our work are as follows:

- We conducted a formative study of how people describe desired infographic exemplars in free-form natural language and derived a structured intent taxonomy spanning five facets.
- We introduce an intent-aware infographic retrieval framework that leverages taxonomy-guided query rewriting, facet weighting, and lightweight embedding alignment for multi-facet matching.
- We build an interactive system that supports end-to-end exemplar-driven authoring in a chat session, enabling users to retrieve relevant exemplars and adapt them to their own data.

2 Related Work

2.1 Infographic Authoring

Research on infographic authoring can be divided into two lines of work: authoring from scratch [18, 27, 35–37] and authoring through exemplar reuse [4, 25]. In the first line, expressive authoring tools such as Data Illustrator [18] and Charticulator [27] give authors direct control over marks, bindings, and layout constraints, while systems like InfoNice [37] package similar capabilities into more novice-friendly workflows. Recent mixed-initiative systems further lower the barrier by interpreting higher-level user goals expressed in natural language [35, 36] or structured around the message to be conveyed [45]. However, these tools still require numerous design decisions, which can be overwhelming for authors without a clear design goal. Since designers frequently start from existing designs to reduce effort [13], a second line of work supports authoring through exemplar reuse. Retrieve-Then-Adapt [25] retrieves a proportion-related infographic exemplar and adapts it to new data, and Wang *et al.* [4] generalized this idea by treating existing infographic charts as reusable authoring assets. Complementary systems automate narrower reuse subproblems such as timeline structure extraction [46], design exploration [34], and palette recommendation [42]. In contrast to these methods, our method focuses on the missing retrieval layer between vague author intent and downstream adaptation.

2.2 Visualization Recommendation

Visualization recommendation automatically generates and ranks chart specifications from structured data. Research in this area has progressed from rule- and specification-based chart suggestion [19, 29, 38–40], to knowledge-based recommenders [10, 15, 20], and more recent natural-language, user-adaptive systems [5, 6, 8, 21, 32, 43]. Early systems such as *Show Me* [19] and *Voyager* [39] frame design rules as search over a structured space of encodings and transformations, while grammars such as Vega-Lite [29] make it easier to steer recommendation from partial specifications. While this line of work makes design choices easier, it still depends heavily on hand-authored grammars and explicitly specified constraints. Later work responds by learning or formalizing richer ranking criteria. For example, *VizML* [10] learns likely design choices from large corpora, while *KG4Vis* [15] makes these criteria more explicit and interpretable using knowledge graphs. Recent systems make user intent easier to express, either by mapping natural-language inputs to chart specifications [6, 21, 30] or by incorporating preference signals [5, 8, 32, 43]. These advances are informative for our setting, but recommendation systems typically output chart specifications over structured data, whereas we retrieve whole infographic references whose usefulness depends on holistic design factors that such systems usually abstract away.

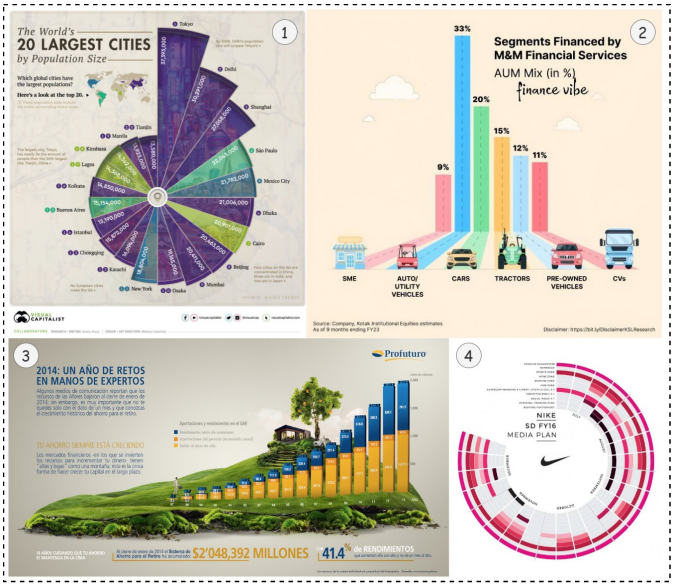

Fig. 2: Four infographic exemplars used in the formative study to elicit participants' queries of desired infographic exemplars.

2.3 Visualization Retrieval and Cross-Modal Search

Unlike recommendation that constructs specifications within a predefined grammar space, retrieval searches a corpus of existing visualizations to find relevant examples. This is particularly suited for infographic authoring, where users often seek design references that are difficult to specify in advance. Cross-modal visual retrieval in this context includes three related lines of work: generic vision-language retrieval models [11, 26, 33, 44], example-centric and context-aware design search [1, 12, 13, 31], and visualization-specific retrieval methods [3, 14, 22–24, 28]. Generic models such as *CLIP* [26] align text and images in a shared embedding space to enable open-ended retrieval, and later large-scale training efforts such as *SigLIP* [33] and *MegaPairs* [44] further strengthen this paradigm. However, generic embedding similarity typically collapses multiple relevance cues into a single score, offering limited transparency and limited control over which aspect of a design dominates matching. A complementary line of work in design studies and design-support tools highlights the central role of examples in creative practice [1, 13]. Building on this insight, Kovacs *et al.* [12] ranked candidate assets based on their compatibility with the surrounding composition, while Son *et al.* [31] supported more expressive search by helping users concretize vague intents and search with generated or edited visual queries. In the visualization domain, Saleh *et al.* [28] modeled style similarity for infographic search, capturing aesthetic resemblance but not broader semantic or structural intent. Later work incorporates richer retrieval signals, such as structural and visual cues [14], comparisons of chart retrieval pipelines [23], and explicit consideration of which visualization properties should determine similarity [22]. Our work builds on these works but focuses on infographic exemplar search for authoring, where the query is often under-specified and must be decomposed into controllable intent facets rather than collapsed into a single similarity notion.

3 Formative Study

3.1 Study Goals and Research Questions

We conducted a formative study to inform the design of our intent-aware infographic retrieval system by understanding how users describe their search intent. Specifically, we elicited participants' queries of

Table 1: **Five-facet intent taxonomy derived from the formative study.** The taxonomy summarizes recurring facets in users' natural-language descriptions of desired infographic exemplars. Facets can be partially specified: a query may omit one or more facets, or express them only implicitly.

Facet	What it captures	Typical cues in user queries
Content	Communicative goal and information organization: what should be shown and what takeaway should be emphasized.	category breakdown/share; trend or growth over time; compare A vs. B; highlight the largest component; include key numbers or summaries
Style	Holistic visual aesthetics: overall tone and look-and-feel, including palette and typography cues.	clean or minimalist; editorial or magazine-like; warm or playful; brand-like; pastel or muted colors; 3D or colorful; typography-forward
Layout	Spatial composition and narrative organization: how the infographic is arranged and read, beyond chart type alone.	vertical poster; clear hierarchy or sections; center chart with side summaries; radial layout with fixed-angle segments; dense labels or annotations; start at the positive y-axis and go clockwise
Illustration	Whether and how icons or illustrations are used, including their density and explanatory versus decorative role.	icons for each category; illustration-heavy; replace labels with pictorial symbols; scene-based background; minimal decoration
Chart Type	Primary chart or visual form used to encode data, often as shorthand for the intended visual encoding.	bar chart; stacked bars; pie or donut; rose/Nightingale chart; radial chart; ring heatmap; timeline-like chart

desired exemplars under two interfaces: a conventional keyword-based image search interface, and an imagined AI-assisted retriever that can interpret long natural language prompts. This study design allows us to characterize the facets of intent that users naturally include and identify gaps in expression when users are constrained to keyword-based search.

Accordingly, we formulated two research questions:

- **RQ1:** What facets of intent do users usually express in natural language descriptions of desired infographics?
- **RQ2:** Which facets are under-expressed in keyword queries compared to natural language queries, and how does this affect the retrieval performance and users' confidence?

Natural language queries are hypothesized to capture richer intent context than keyword queries. By answering RQ1 and RQ2, we seek to validate this hypothesis and identify specific intent facets that a keyword-based interface might fail to capture. These insights will directly inform the design of the query interpretation module in our system and highlight where intent-aware features are most needed.

3.2 Participants

We recruited **14** participants (n=14; 10 male / 4 female), who completed the online questionnaires remotely. Participants were aged 21–45 years (M=27.1, SD=6.8) and reported majors/primary fields in computer science and design. Most participants were students (1 undergraduate / 7 master's / 3 PhD), and the remainder were practitioners working in data analysis or software engineering.

To contextualize participants' familiarity with data visualization, we asked them to self report on 1) how often they create charts/visualizations and 2) how often they search online for reference charts/infographics for design inspiration. For chart creation, some $(5/14)$ reported doing so frequently, and most $(9/14)$ reported doing so occasionally. For searching reference charts, most $(8/14)$ reported doing so frequently, some $(5/14)$ reported doing so occasionally, and only one $(1/14)$ reported doing so almost never. This indicates that our participants have a wide range of prior exposure to chart-making and design-inspiration seeking behaviors, which is relevant for interpreting differences in how participants articulate infographic intent.

3.3 Procedure

To cover a diverse range of infographic designs for eliciting query descriptions, we selected four infographic exemplars that differ in chart type, layout, illustration density, and overall aesthetic. As shown in Figure 2, they include: 1) a minimalist radial schedule / calendar-like visualization, 2) an illustrated, perspective bar-chart poster, 3) a scene-based, 3D-styled growth infographic, and 4) an editorial-style radial ranking infographic with dense annotations.

For each stimulus image, participants completed the following tasks:

1. **Keyword-style query.** Participants imagined using a typical image search website (e.g., Pinterest or Google Images) and wrote the query they would realistically type based on their usual search habits. They then rated, on a 5-point Likert scale (1 = very unlikely, 5 = very likely), how likely this keyword query would be to retrieve their desired results based on their prior experience with such websites.
2. **Natural language query.** Participants imagined an AI-powered infographic retrieval system that could understand long, prompt-like natural language descriptions. Ignoring current technical limitations, they wrote what they would ideally like to input to retrieve the target infographic or highly similar ones. We did not collect confidence ratings for the natural language query because participants generally lacked prior experience with such systems.

We randomized the order of the four stimuli to mitigate potential order effects. The study yielded 56 (4×14) keyword-style queries with confidence ratings and 56 natural language queries.

3.4 Results and Analysis

We conducted iterative qualitative coding to identify recurring intent facets in both keyword queries and natural language queries. Two researchers independently performed open coding on a subset of queries, marking phrases that describe *what* the infographic should communicate and *how* it should be visually realized. The two researchers then reconciled differences through discussion, consolidated an initial codebook, and iteratively refined it while applying the codes to the full set of queries. Here, multiple codes per query are allowed because participants frequently combined multiple requirements in a single query. Finally, the two researchers grouped codes into higher-level categories that capture major intent facets and examined how these facets co-occur within individual queries. This coding process yields three findings that directly motivate an intent-aware retrieval formulation.

F1: Users describe infographics through multiple interacting facets. Participants often went beyond topic keywords and described design-relevant aspects, such as information organization, compositional structure, illustration usage, and overall aesthetic. For example, one participant described the second infographic as "each bar extends in perspective," which captures both the chart type and the stylistic rendering rather than only the topical content. Participants also used chart-type terms as an explicit part of their intent (e.g., "rose chart," "ring heatmap," "bar chart") rather than as an afterthought. From these descriptions, we derived a five-facet taxonomy of infographic search intent: **content** (communicative goal and what should be conveyed), **chart type** (the primary chart/visual form used to encode data), **layout** (spatial composition, hierarchy, and narrative organization), **illustration** (icons/illustrations and their usage), and **style** (overall aesthetics such as typography, palette, and visual tone). Table 1 summarizes the facets and typical linguistic cues observed in the study.

We further observe that facets often co-occur within a single query, and that a query can mix explicit constraints (e.g., "radial layout") with implicit signals (e.g., "editorial-style" implying dense annotations and typography-heavy design). Figure 3 shows a facet co-occurrence matrix and an annotated example query.

F2: Natural language queries encode richer multi-facet intent, whereas keyword queries compress intent and under-specify

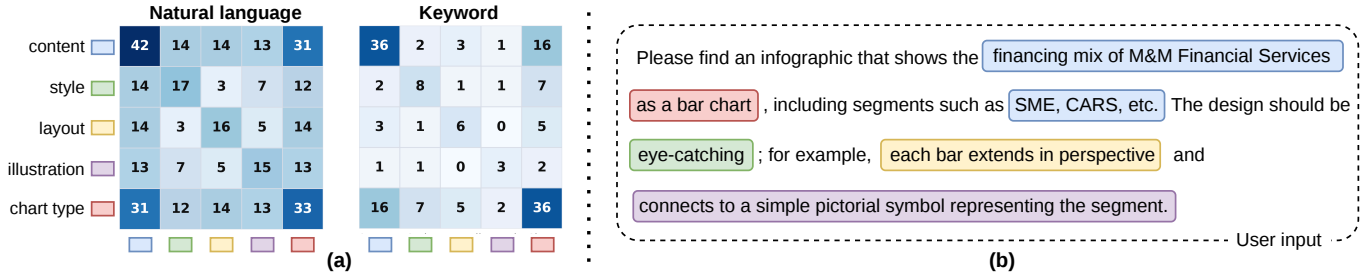

Fig. 3: **Facet patterns in participants' queries.** (a) A facet co-occurrence matrix computed from the natural language queries and keyword queries. (b) An annotated example query illustrates how a single query can contain multiple facet signals, motivating facet-specific rewriting in retrieval.

design constraints. When comparing natural language queries and keyword queries, natural language ones expressed more facet-level constraints and more often included design-critical requirements beyond topical semantics. Specifically, natural language queries expressed more facets per query (2.29 facets/query) than keyword queries (1.59 facets/query). Moreover, half of the natural language queries contained ≥ 3 facets, while only 8.9% of keyword queries reached this level, indicating that keyword inputs often compress intent into one or two high-level descriptors. A detailed analysis reveals that this compression disproportionately affected design-critical facets. Natural language queries frequently specified *layout* (39.3%), *illustration usage* (28.6%), and *style/aesthetic tone* (32.1%), whereas these facets were rare in keyword queries (layout 10.7%, illustration 5.4%, and style 14.3%). Aggregating these three facets, 69.6% of natural language queries mentioned at least one design facet (layout, illustration, or style), compared to 26.8% of keyword queries. In contrast, content and chart-type terms remained common in both conditions (71.4% and 64.3% for natural language and keyword queries, respectively).

F3: Users are not confident that keyword search will return their intended designs. Participants' confidence ratings provide quantitative evidence for this gap. Across 56 keyword-based queries, the median confidence was only 2 (IQR: 1–3; M=2.32, SD=1.15). Moreover, 82.1% of the ratings were 3 or below, suggesting that participants generally did not expect conventional keyword-based search to reliably return their intended results even when a clear target was shown. This aligns with a common failure mode in practice: results can be semantically related yet misaligned with design-critical facets such as chart type, layout, illustration usage, and overall style.

Together, these findings motivate an intent-aware retrieval approach that 1) makes intent facets explicit, 2) supports partial specification when some facets are absent, and 3) performs facet-aware matching rather than relying on a single overall similarity score.

4 Intent-Aware Infographic Retrieval Framework

Motivated by the multi-faceted nature of infographic search intent, we propose an intent-aware retrieval framework that 1) parses a free-form query into five intent facets with per-facet weights, 2) computes facet-aware similarities, and 3) ranks exemplars based on a weighted combination of facet-aware similarities. Figure 4 summarizes the pipeline.

4.1 Intent Representation and Query Parsing

Given a free-form query q, we parse it into five intent facets $\mathcal{F} = \{$**content** (C), **chart type** (T), **layout** (L), **illustration** (I), **style** (S)$\}$. Among these, chart type takes values from a relatively fixed label space $\mathcal{T}$, so we treat it as a multi-choice constraint and infer a set of labels $q_{\mathrm{T}} \subseteq \mathcal{T}$ from q when specified. Following ChartGalaxy [16], we use a compact pool $\mathcal{T}$ of 13 coarse chart types: *Bar Chart, Line Chart, Area Chart, Radar Chart, Pie Chart, Scatterplot, Gauge Chart, Treemap, Diagram, Histogram, Range Chart, Funnel Chart, Pyramid Chart*. The remaining four facets are naturally expressed in open vocabulary, so we use LLMs to rewrite q into concise facet-focused descriptions $\{q_f\}_{f\in\{\mathrm{C,L,I,S}\}}$. For those unspecified facets, we set $q_f = \emptyset$.

In addition to the facet-focused descriptions, the parser also predicts a non-negative facet weight vector $\mathbf{w} = \{w_f\}_{f\in\mathcal{F}}$ to reflect the importance of each facet. We set $w_f = 0$ if the facet f is unspecified.

In our implementation, we use Qwen3-32B with a taxonomy-guided prompt and a fixed output schema to parse the query. We validate the schema and automatically retry on invalid outputs. The first-time schema invalid rate is about 4e-4. Prompt templates and the full schema are provided in Supplementary Section 2.

4.2 Facet-Aware Retrieval Scoring

After parsing q into facets, we score each exemplar infographic x_i by combining multi-facet similarities. The final relevance score is a weighted sum $S(q,x_i) = \sum_{f\in\mathcal{F}} w_f \cdot s_f(q,x_i)$, where $s_f(q,x_i)$ is the similarity score for facet f between the query q and the exemplar x_i. We next describe how we compute $s_f(q,x_i)$ for each facet.

4.2.1 Chart Type Similarity

In ChartGalaxy, each exemplar x_i is associated with a chart type set $T_{x_i} \subseteq \mathcal{T}$. When the user specifies chart type as a multi-choice set $q_{\mathrm{T}} \subseteq \mathcal{T}$, we compute a chart-type similarity score $s_{\mathrm{T}}(q,x_i)$ between q_{T} and T_{x_i}. Rather than requiring set-overlap agreement with strict matches, we allow *soft* matches between visually similar chart types, since they can be ambiguous at coarse granularity and users may not always name the intended type precisely (e.g., an area chart can be described as a line chart with the region filled).

Concretely, we define a type-to-type kernel $\kappa(t,t') \in [0,1]$ with $\kappa(t,t') = 1$ if $t = t'$, and non-zero values only for expert-specified similar pairs (listed in Supplementary Section 3). We then match each queried type to the best-matching type present in the exemplar and average across queried types:

$$s_{\mathrm{T}}(q,x_i) = \frac{1}{|\mathcal{T}_q|} \sum_{t\in\mathcal{T}_q} \max_{t'\in T_{x_i}} \kappa(t,t'), \tag{1}$$

4.2.2 Embedding-Based Facet Similarities.

For the open-vocabulary facets $\{\mathrm{C,L,I,S}\}$, we compute a separate text–image similarity per facet. A naive solution would be to compute similarity between the text embedding of facet-focused description q_f and the image embedding of infographic x_i. However, this method may not faithfully measure similarity along that specific facet, because different facets rely on different linguistic cues in the query and different visual evidence in the image. We therefore seek facet-aware embeddings that model similarity separately for different facets. To obtain them, we condition the text encoder with facet-specific tokens and project the image embedding through facet-specific heads, yielding facet-conditioned similarities.

Facet-aware text embeddings. We use a shared text encoder $E_T(\cdot)$ conditioned by facet-specific tokens to construct this. Specifically, for each open-vocabulary facet f, we prepend a special token $\langle f\rangle$ to the facet rewrite q_f to obtain the text embedding:

$$\mathbf{e}_{q,f} = \begin{cases} \mathrm{norm}\big(E_T(\langle f\rangle \oplus q_f)\big), & q_f \neq \emptyset, \\ \mathbf{0}, & q_f = \emptyset, \end{cases} \tag{2}$$

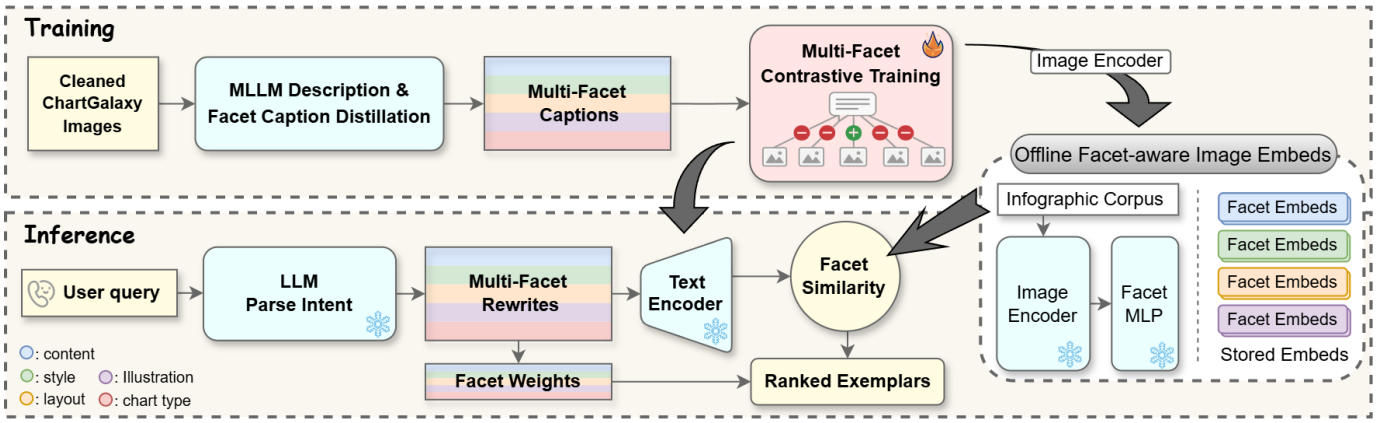

Fig. 4: **Overview of our intent-aware infographic retrieval framework.** *Top (training):* ChartGalaxy images are described by a multimodal LLM and distilled into short facet-specific captions (two per facet). We fine-tune the full model (text encoder, image encoder and facet MLP heads) with a multi-facet in-batch contrastive loss (Eq. 5). *Bottom (inference):* A user query q is parsed into facet rewrites $\{q_f\}$, facet weights $\mathbf{w}$, and an optional multi-choice chart-type set $\mathcal{T}_q$. We compute facet-conditioned text embeddings and compare them with precomputed facet-specific image embeddings to obtain $\{s_f\}_{f\in\mathcal{F}_e}$, and compute a multi-choice discrete chart-type similarity s_T. We fuse all facets by a weighted sum to rank exemplars.

where $\mathrm{norm}(\cdot)$ denotes ℓ_2 normalization, $\oplus$ denotes concatenation, and $\mathbf{0}$ is a zero vector of the corresponding dimensionality.

Facet-aware image embeddings. For each exemplar infographic x_i, we compute a base image embedding using an image encoder $E_I(\cdot)$: $\mathbf{h}_{x_i} = \mathrm{norm}(E_I(x_i))$. We then attach a lightweight MLP head per embedding facet to produce facet-specific image features:

$$\mathbf{e}_{x_i,f} = \mathrm{norm}\big(\mathrm{MLP}_f(\mathbf{h}_{x_i})\big). \tag{3}$$

This multi-head design provides facet-conditioned projections of the same image representation, enabling different relevance notions for different facets.

Facet-conditioned similarities. For each embedding facet f, we compute cosine similarity between the text embedding $\mathbf{e}_{q,f}$ and the image embedding $\mathbf{e}_{x_i,f}$:

$$s_f(q,x_i) = \mathbf{e}_{q,f}^\top \mathbf{e}_{x_i,f}. \tag{4}$$

Please note that the facet-aware scoring function in Eq. 4 is parameterized by 1) facet tokens $\{\langle f\rangle\}_{f\in\mathcal{F}_e}$ that condition the shared text encoder, and 2) facet-specific projection heads $\{\mathrm{MLP}_f\}_{f\in\mathcal{F}_e}$ that map a shared image embedding to facet spaces. These additional degrees of freedom are not explicitly supported by generic CLIP pretraining, especially for design-centric facets such as layout and style. Therefore, we construct in-domain facet-level supervision and optimize the same facet-aware similarities, as described next.

4.3 In-domain Alignment with Synthetic Facet Captions

To learn the facet-specific text token and MLP heads, we construct in-domain facet-level supervision from ChartGalaxy [16] and perform contrastive alignment for each facet.

Alignment data construction. We removed near-duplicates in image-embedding space from the original ChartGalaxy dataset and randomly sampled 52,000 infographics for training. For each training infographic x_i, we then generated facet-specific supervision for the four embedding facets by first producing a rich multimodal description with Qwen3-VL-8B and then distilling it into short facet-focused captions, following large-scale retrieval training practices such as MegaPairs [44]. For each open-vocabulary facet f, we created two paraphrased captions $\{c_f^{(1)}(x), c_f^{(2)}(x)\}$, yielding eight training texts per image, and trained on these distilled captions to provide cleaner facet-targeted supervision with less cross-facet leakage.

Multi-facet contrastive alignment. We optimize the same facet-aware similarity used at inference. For each image x in a minibatch $\mathcal{B}$, each facet f, and each caption variant $k \in \{1,2\}$, we treat $(c_f^{(k)}(x), x)$ as a positive pair under Eq. 4, and use other images in the batch as negatives:

$$\mathcal{L} = \sum_{x\in\mathcal{B}} \sum_{f\in\mathcal{F}_e} \sum_{k\in\{1,2\}} -\log \frac{\exp\big(\mathbf{e}_{c,f}^{(k)}(x)^\top \mathbf{e}_{x,f}/\tau\big)}{\sum_{x'\in\mathcal{B}} \exp\big(\mathbf{e}_{c,f}^{(k)}(x)^\top \mathbf{e}_{x',f}/\tau\big)}, \tag{5}$$

where $\mathbf{e}_{c,f}^{(k)}(x) = \mathrm{norm}\big(E_T(\langle f\rangle \oplus c_f^{(k)}(x))\big)$, τ is a temperature, and the image embedding $\mathbf{e}_{x,f}$ is computed via the same MLP_f heads in Eq. 3. This training encourages the model to learn facet-sensitive similarity functions that better reflect how users describe infographic exemplars.

5 Conversational Retrieval-and-Adaptation System

Retrieval alone is often insufficient to support practical infographic authoring. Even when a highly relevant exemplar is retrieved, users must still adapt its structure and visual style to their own data. To bridge the gap between inspiration seeking and downstream reuse, we develop an exemplar-driven infographic authoring system, enabling users to 1) explore inspirations through retrieval, 2) commit to a small set of infographic exemplars, and 3) adapt those exemplars to their own data.

5.1 Interface and Typical Workflow

Figure 5 illustrates a concrete chat-centered workflow for exemplar retrieval and adaptation. In this example, the user first asks in the chat panel (A) for "chart exemplars suitable for my data." The system parses this request into weighted retrieval facets shown in the retrieval panel (D1), supports chart-type filtering in (D2), and retrieves relevant exemplars in the candidate gallery (D3). After the user manually selects two exemplars, they remain pinned in the exemplar panel (B) as persistent references for downstream adaptation. The user then specifies a concrete combination strategy in the chat: keep the screen icon from exemplar 1, adopt the bar-chart style from exemplar 2, and adapt the design to the user's screen-time data. The output panel (C) first shows a draft SVG (Version #1) that already integrates cues from both references, combining the illustration from one exemplar with the chart style of the other. After the user provides follow-up feedback to correct problems in the initial SVG output, the system generates a refined SVG (Version #2) with proportional horizontal bars and a tighter canvas while preserving the selected visual style. This example highlights how retrieval controls (D), pinned exemplars (B), conversational instructions (A), and iterative SVG refinement (C) work together in a single end-to-end authoring workflow.

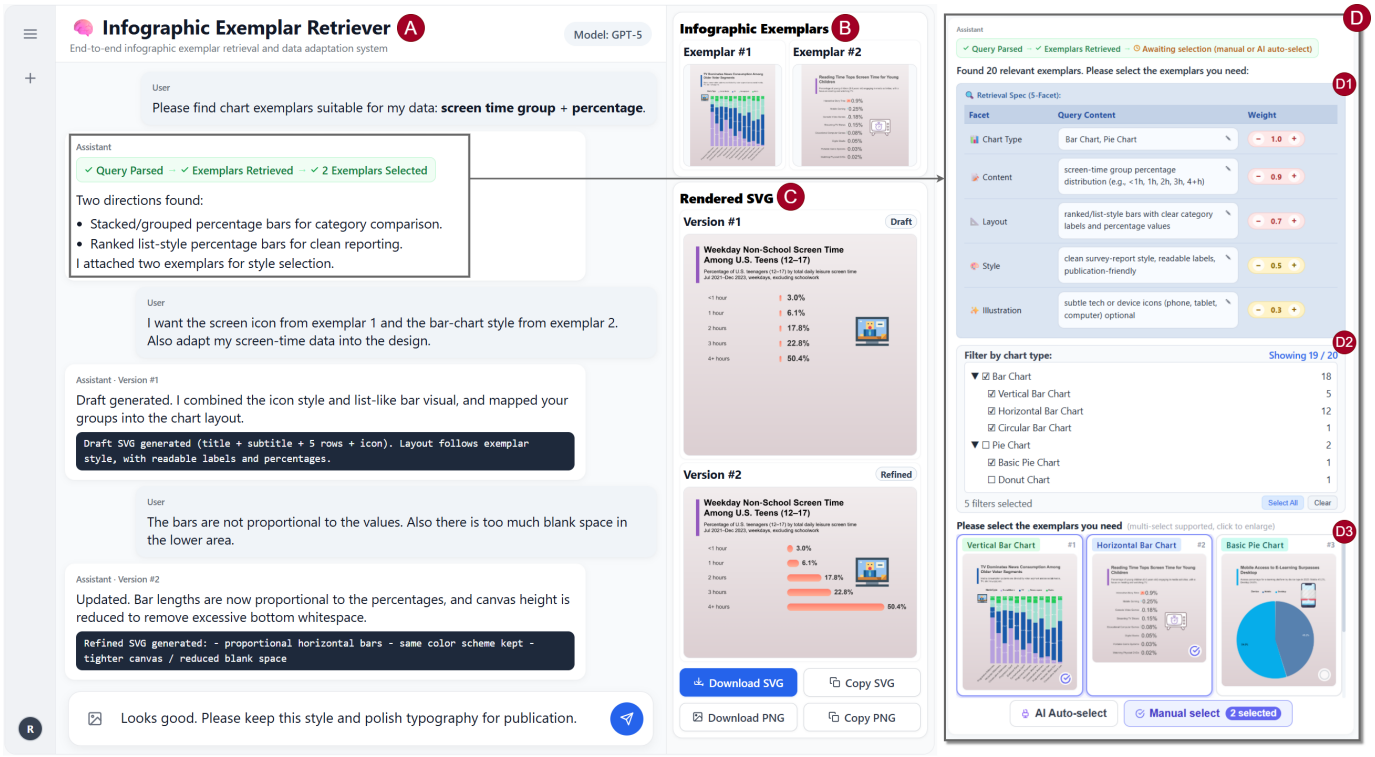

Fig. 5: **System UI for conversational exemplar retrieval and SVG-based adaptation.** (A) chat panel for intent articulation, assistant feedback, and iterative refinement; (B) committed infographic exemplars that persist as references during adaptation; (C) rendered SVG outputs with version history; and (D) retrieval panel for exemplar discovery and commitment.

5.2 Exemplar Retrieval

During the inspiration phase, the assistant invokes the retriever described in Section 4 to surface candidate infographic exemplars relevant to the user's request. The retrieval panel (Figure 5D) makes this process explicit by exposing an editable facet-based retrieval specification, together with hard chart-type filters and a browsable candidate gallery (D1–D3). When users are not satisfied with the retrieval results, they can directly revise multi-facet queries and their importance weights in the panel, or continue refining the request through the chat, in which case the system updates the query text and weights and reruns retrieval. After browsing the candidate gallery, users commit to a small set of infographic exemplars for downstream adaptation. The system supports two selection modes: manual selection and AI auto-selection. In AI auto-selection, a vision–language model re-ranks the top-10 retrieved candidates conditioned on the user's query and selects 1–3 exemplars, capturing fine-grained intent cues beyond similarity-based ranking alone while also promoting a more diverse set of references. Once committed, the selected infographic exemplars persist across subsequent turns and are displayed alongside the conversation (B), grounding iterative follow-up requests in the same exemplar set.

5.3 Infographic Adaptation

After users commit to a small set of infographic exemplars, the system helps them adapt those exemplars to their own data. Such adaptation typically involves directly fitting new tabular data into an existing design, reusing a specific visual style, or extracting particular visual elements (e.g., icons). However, supporting this naively is difficult because infographic SVGs are often extremely long, containing embedded base64 assets and repetitive low-level structures that can easily overwhelm an MLLM's context window. To address this, we design a progressive, tool-augmented adaptation pipeline driven by the user's data adaptation needs.

Default context: Compact structural summary. Whenever an exemplar is committed and the conversation enters the adaptation phase, the system does not feed the raw SVG to the model. Instead, it automatically converts the SVG into a compact tree-structured summary that preserves hierarchical grouping (e.g., nested <g> elements) and coarse content signatures (e.g., dominant element types and counts). Each node is assigned a stable `node_id`. This lightweight summary is provided to the assistant as the default context. It enables the model to globally reason about *where* a user's data should be mapped (e.g., marks, legends, or title regions) without incurring the cost of processing the full SVG source.

On-demand inspection: Sanitized code retrieval. When the user specifies a data adaptation request or a targeted styling change, the assistant determines which specific regions require modification based on the baseline structural summary. It then invokes a system tool, `show_full_svg(node_id)`, to retrieve the detailed, code-level representation for only those necessary nodes. To further mitigate context bloat, the backend sanitizes this retrieved code on the fly: large embedded payloads like base64-encoded images are removed and replaced with lightweight placeholders. This allows the assistant to zoom in and safely inspect layout-relevant SVG structure without processing massive binary strings.

Output generation and SVG reconstruction. With the necessary code snippets inspected, the assistant executes the adaptation—such as calculating new geometric attributes for data binding or updating text labels—by outputting the modified SVG code specifically for those queried `node_ids`. The system backend then automatically stitches these modified subtrees back into the overall SVG structure. Crucially, during this reconstruction, the backend restores the original high-fidelity payloads, such as the base64 images, that were temporarily held out by placeholders. The fully reconstructed SVG is then rendered in the output panel (Figure 5C), seamlessly completing the adaptation loop and allowing the user to provide subsequent natural language commands for further refinement.

Table 2: **Single-round query-to-image retrieval.** We report Recall@1/5 (%) and MRR@10 (fraction) on synthetic general queries, synthetic multi-facet queries, and human-written paired short/long queries. ↑ higher is better.

Method	Synthetic (General)			Synthetic (Multi-facet)			Human (Short)			Human (Long)		
	R@1↑	R@5↑	MRR@10↑	R@1↑	R@5↑	MRR@10↑	R@1↑	R@5↑	MRR@10↑	R@1↑	R@5↑	MRR@10↑
CLIP [26]	62.47	84.74	0.7203	46.42	70.83	0.5696	35.00	58.00	0.4375	41.00	67.33	0.5150
SigLIP2 [33]	19.38	38.05	0.2739	67.59	86.45	0.7566	33.00	50.00	0.4042	20.67	33.33	0.2632
MegaPairs [44]	67.42	88.18	0.7630	50.08	74.87	0.6075	39.33	65.67	0.5036	50.33	73.67	0.5913
OURS W/O IN-DOM. ALIGN.	70.22	87.43	0.7760	51.38	75.76	0.6168	30.00	47.00	0.3732	35.00	57.33	0.4424
OURS W/O FACETS	92.79	98.88	0.9545	89.67	97.60	0.9332	53.67	77.33	0.6345	65.00	82.33	0.7237
OURS W/O FACET WEIGHTS	**95.92**	**99.72**	**0.9760**	90.85	98.40	0.9422	54.33	81.67	0.6552	69.33	**86.67**	0.7689
OURS	95.72	99.68	0.9746	**91.29**	**98.48**	**0.9447**	**54.67**	**82.00**	**0.6591**	**70.33**	**86.67**	**0.7744**

6 EXPERIMENTS

We evaluate both our retriever and the end-to-end authoring system in three complementary settings: 1) single-round retrieval, 2) multi-round retrieval, and 3) retrieval-based authoring.

6.1 Single-Round Retrieval

6.1.1 Automatic Retrieval Evaluation

We randomly select 10,000 infographics from the ChartGalaxy to form a fixed corpus $\mathcal{X}$. Given each pair of a text query q and a unique ground-truth target infographic $x^* \in \mathcal{X}$, a retriever ranks all infographics in $\mathcal{X}$ and returns a top-K list. We report Recall@K and MRR@10 to evaluate the retrieval performance under a unique-target protocol.

The construction of query-target pairs. We construct three types of query-target pairs to cover different levels of query specificity.

- **Synthetic general queries.** For each target infographic $x_i \in \mathcal{X}$, we generate a rich natural-language description by prompting Qwen3-VL to summarize its visual content. This description is directly used as the retrieval query q_i. This serves as a natural synthetic baseline for measuring whether the retriever can align full-image descriptions with the corresponding visual instances.
- **Synthetic multi-facet queries.** Building on the rich description above, we further generate its facet-specific descriptions using the same pipeline described in Section 4. We then construct a composite query q_i by concatenating these facet-specific descriptions. This is designed to test the retriever's ability to leverage more structured and fine-grained design intent specifications.
- **Human-written queries.** To evaluate retrieval under more realistic user-facing conditions and avoid potential biases in the synthetic query generation, we randomly sample 300 infographics from $\mathcal{X}$ and collect human-written queries. Annotators were instructed to write both a short query and a long query for each infographic. The short query captures only the most salient cues required to identify the intended infographic, while the long query provides a more complete specification by describing additional aspects, such as layout and style, in free-form natural language. The detailed prompt and instructions to the annotators are provided in Supplementary Section 5.2.

Baseline methods and ablations. We compare our method (OURS) with the following baselines:

- **CLIP** [26]: a standard and widely used CLIP ViT-B/32 dual-encoder baseline that well captures text–image similarity.
- **SigLIP2** [33]: a more recent and stronger dual-encoder baseline.
- **MegaPairs** [44]: a retrieval-oriented model trained with large-scale synthetic supervision.

We also consider the following ablations to examine the contribution of individual components

- **OURS W/O FACETS**: removes facet decomposition and query-dependent fusion. Specifically, it only rewrites each query and then retrieves infographics using the same model as our method.
- **OURS W/O FACET WEIGHTS**: uses uniform weights over present embedding facets.
- **OURS W/O IN-DOMAIN ALIGNMENT**: removes the in-domain alignment step in our training pipeline.

Metrics. Let $\mathrm{rank}(q)$ be the rank of the target infographic for query q, and $\mathcal{Q}$ be the set of all queries. We report Recall@K and MRR@10.

Recall@K (R@K) is the fraction of queries whose target appears in the top-K ranked list:

$$\mathrm{R@}K = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbf{1}\big[\mathrm{rank}(q) \le K\big], \quad K \in \{1,5\}. \tag{6}$$

MRR@10 is the mean reciprocal rank of the ground-truth target, truncated at the top 10 results:

$$\mathrm{MRR@10} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \begin{cases} \frac{1}{\mathrm{rank}(q)}, & \mathrm{rank}(q) \le 10, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Results. Table 2 summarizes single-round retrieval results. Overall, OURS consistently improves retrieval compared to strong off-the-shelf retrievers, with especially large gains on human-written queries. The in-domain alignment step is necessary for improving retrieval performance, and explicitly modeling multiple facets can further improve retrieval performance, especially for long human-written queries.

A notable pattern among the external baselines is SigLIP2 [33], which remains competitive on synthetic multi-facet queries, but drops significantly on synthetic general queries and long human-written queries. This suggests that a global text embedding is less effective when multiple facets are interleaved. Our method overcomes this by explicitly modeling multiple facets and then performing retrieval.

6.1.2 Human Judgment of Retrieved Results

Automatic metrics under a unique-target protocol may not fully reflect retrieval usefulness for infographic authoring, because a query can correspond to multiple highly relevant exemplars. To better capture practical usefulness, we complemented automatic evaluation with a human judgment study on the retrieved lists.

Evaluation setup. For each human-written query, we retrieve the top-5 results returned by MegaPairs and OURS, respectively. These two ranked lists are presented side by side in randomized order. Each query–method pair is scored independently by two raters, and we use the average of the two scores in all analyses. Raters score each list on a 1–5 Likert scale based on overall intent match and usefulness as an exemplar, where 1 indicates *irrelevant* and 5 indicates *near-perfect match*. To anchor the upper end of the scale, raters are instructed to assign a score of 5 when the target or intended exemplar already appears at rank 1, since the retrieval has then fully satisfied the user's need.

Results. We report the mean of the two raters' scores for each query–method pair, and summarize win/tie/loss using paired comparisons of these averaged scores. Inter-rater agreement was high, with a quadratic weighted Cohen's κ of 0.98 and 90.0% exact agreement. As shown in Table 3, OURS substantially outperforms MegaPairs in perceived exemplar usefulness, increasing the mean score from 3.09 to 4.20 for short queries, from 3.26 to 4.51 for long queries, and from

Table 3: **Human scoring of top-5 retrieved results.** ↑ higher is better.

	Mean score (1–5)↑			Preference vs. MegaPairs (%)		
Method	**Short**	**Long**	**Overall**	**Win**	**Tie**	**Loss**
MegaPairs	3.09	3.26	3.17	–	–	–
OURS	**4.20**	**4.51**	**4.35**	**57.3**	37.3	5.3

3.17 to 4.35 overall. The larger improvement on long queries suggests that OURS better exploits richer intent descriptions beyond topical similarity. Across 600 paired comparisons, OURS wins in 57.3% of cases, ties in 37.3%, and loses in only 5.3%.

6.2 Multi-Round Retrieval

Beyond the single-round evaluation, we conducted a user study to assess retrieval performance in an interactive multi-round setting.

6.2.1 Study Design

Participants. We recruited 12 participants (P1-P12, 9 male, 3 female), age 20–25 (M=22.67, SD=1.37). Four identified as design novices (33.3%), six reported some prior experience (50.0%), and two reported proficiency (16.7%). Participants reported moderate-to-high experience in graphic design or layout tools (M=4.08/5, SD=1.24).

Task and stimuli. The task is to retrieve examples that are semantically and visually similar to the target infographic. We selected four representative infographics from the corpus $\mathcal{X}$ to cover diverse visual styles, including variation in chart type, layout structure, illustration density, and overall appearance. In each task, participants were shown one target infographic and used a web interface consisting of a target image panel, a text input box, a top-10 result grid, and a save panel for collecting candidate results. Participants can iteratively write and revise the query, save promising results throughout the search, rate saved results, and stop whenever they feel ready to choose a best result.

Conditions. We compared OURS with a BASELINE that uses the same LLM as our system (**GPT-5**) to rewrite the participant's current input into a single holistic caption-style query, followed by retrieval with MegaPairs [44]. In contrast to OURS, the baseline does not construct a five-facet retrieval specification and performs retrieval from only one rewritten query. The study used a counterbalanced within-subject design. Specifically, each participant completed four tasks, with two targets assigned to OURS and two to BASELINE.

Study procedure. Each session began with a brief introduction to the task and interface, followed by one practice task to familiarize participants with the multi-round retrieval workflow. Participants then completed four formal tasks. During each task, participants were required to save at least two retrieved infographics that they considered good matches or useful alternatives. For every saved infographic, they provided a satisfaction rating on a 1–7 Likert scale to indicate how well it matches the target infographic. When ready to stop, they selected one saved infographic as the **best** result for the task. We marked a task as **Found** if the participant saved the exact target infographic. The full session took approximately 30 minutes.

6.2.2 Metrics

We use four metrics to evaluate the performance of multi-round retrieval. **Rounds-to-stop** measures how many interaction rounds a participant completed before deciding to stop. **FoundRate** is the proportion of tasks in which the participant saved the exact target infographic. **dCRR@10** is a discounted cumulative reciprocal-rank measure adapted from CRR [2]. Specifically, it first computes the per-round reciprocal rank truncated at 10 and then discounts later rounds by a factor of γ:

$$\mathrm{dCRR@10} = \sum_{r=1}^{R} \gamma^{r-1}\,\mathrm{RR@10}_r,$$

where R is the rounds-to-stop, $\gamma = 0.9$ is the discount factor, and $\mathrm{RR@10}_r$ is the reciprocal rank truncated at 10 in round r. Thus, dCRR@10 rewards finding the target earlier and ranking it higher. **BestSatisfaction** is the 1–7 satisfaction rating assigned to the participant's final selected image.

Table 4: **Multi-round retrieval.** We report task-level mean±SD for continuous metrics and percentage (count) for binary metrics. ↓ lower is better; ↑ higher is better.

Category	**Metric**	**BASELINE**	**OURS**	p
Efficiency	Rounds-to-stop ↓	3.29±2.27	**1.42±0.50**	0.0029
Retrieval	FoundRate ↑	45.8% (11/24)	**91.7% (22/24)**	0.0018
	dCRR@10 ↑	0.21±0.34	**0.73±0.38**	5.7×10^{-5}
Utility (1–7)	BestSatisfaction ↑	5.50±1.62	**6.88±0.45**	0.0029

6.2.3 Statistical Analysis

Each participant completed four tasks, two under each condition, yielding 48 task observations in total. For inferential testing, we first aggregated task-level outcomes within each participant and condition, producing one paired value per participant for each metric: mean Rounds-to-stop, mean dCRR@10, mean BestSatisfaction, and the participant-level proportion for FoundRate. We then compared conditions using paired two-sided t-tests across participants. Reported p values are Holm–Bonferroni corrected across the four reported metrics. Descriptive statistics in Table 4 are shown at the task level.

6.2.4 Results

Table 4 summarizes the primary results. Overall, OURS improves exact-match retrieval and user utility while reducing interaction rounds. Compared to the single-query rewrite baseline, OURS reduces Rounds-to-stop by 1.87 rounds (3.29 vs. 1.42; p=0.0029). Exact-match success nearly doubles, increasing from 45.8% (11/24) to 91.7% (22/24), a gain of 45.9 percentage points (p=0.0018). We also observe substantially stronger cumulative exposure across rounds, with dCRR@10 increasing from 0.21 to 0.73 (+0.52; $p=5.7\times10^{-5}$), and higher satisfaction for the final selected result, with BestSatisfaction increasing from 5.50 to 6.88 (+1.38; p=0.0029).

6.3 Retrieval-based Authoring

Beyond evaluating retrieval quality alone, we further assess the full authoring workflow enabled by our system.

6.3.1 Study Design

Participants. The same 12 participants (P1-P12) from the multi-round search study (Section 6.2) completed this end-to-end evaluation.

Task and stimuli. Each participant completed two infographic-authoring tasks using two tabular datasets from different topical domains. Dataset A captured commuter mode share in Champaign County, Illinois, as percentages across seven travel modes. Dataset B captured weekday daily screen time among U.S. teenagers ages 12–17, grouped into five duration categories excluding schoolwork. For each task, participants were asked to retrieve infographic exemplars, adapt one or more selected exemplars to the data, and iteratively refine the generated SVG until they were satisfied or could no longer make further progress.

Conditions. **OURS** is our one-stop interface that integrates exemplar retrieval, exemplar selection, and iterative SVG adaptation with a live preview and persistent exemplars (Section 5). **Baseline** is a toolchain baseline that uses the same LLM as OURS (**GPT-5**) but exposes the workflow as separate tools: (i) a MegaPairs retriever [44] for finding exemplars, (ii) a plain chat interface for asking GPT-5 to adapt and edit SVG code, and (iii) an external SVG viewer for previewing intermediate results. This condition serves as a realistic toolchain baseline, where participants could use MegaPairs retriever and backbone model, but had to manually transfer exemplars and SVG code across tools. We counterbalanced the system order for each participant.

Table 5: **Retrieval-based authoring.** We report mean±SD for continuous metrics and percentage for preference rate. ↓ lower is better; ↑ higher is better. p values are from paired two-sided t-tests.

Category	Metric	BASELINE	OURS	p
Process	Dialogue turns	**4.83±1.70**	5.17±1.47	0.594
Workload (0–20, ↓)	Overall workload	11.23±4.13	**7.88±2.08**	0.034
	Mental demand	9.25±5.74	**6.00±3.33**	0.036
	Temporal demand	12.50±5.16	**7.75±3.62**	0.028
	Effort	13.50±3.48	**10.67±3.42**	0.055
	Frustration	9.67±6.14	**7.08±3.50**	0.160
Satisfaction (0–20, ↑)	Overall satisfaction	11.28±4.42	**14.36±3.21**	0.130
	Self performance	12.25±4.49	**14.00±3.19**	0.359
	Tool satisfaction	10.58±5.04	**14.58±3.58**	0.079
	Output satisfaction	11.00±4.47	**14.50±3.71**	0.121
Output quality	Blind preference rate ↑	35.4% (17/48)	**54.2% (26/48)**	–

Study procedure. At the beginning of the session, participants were introduced to the experimental setup and given instructions on how to interact with the two systems. During each session, participants completed two assigned tasks, using one system for each task. The experimenter observed the interaction process and maintained a structured record sheet to document process-level observations, including the number of dialogue turns, whether the final output satisfied the task requirements, and any notable breakdowns or recovery behaviors that occurred during task completion. After completing the tasks, participants filled out a post-study questionnaire and provided open-ended feedback. In addition, we conducted informal follow-up questions with four participants to clarify their experiences.

6.3.2 Measures

We measured perceived workload, satisfaction, and output quality.

Perceived workload was assessed using the NASA Task Load Index (NASA-TLX) [9]. Because the authoring tasks involved minimal physical demand, we included four NASA-TLX subscales: *mental demand*, *temporal demand*, *effort*, and *frustration*. Each subscale was rated on a 21-point scale from 0 (*very low*) to 20 (*extremely high*). We computed the mean of them as the primary workload outcome (RTLX).

Satisfaction was measured on the same 21-point scale across three dimensions: self-rated performance, tool satisfaction, and output satisfaction, where 0 indicated *not satisfied at all* and 20 indicated *extremely satisfied*. We also computed the mean of them (Satisfaction).

Output quality was assessed through a blind peer-review protocol among the participants. Each participant evaluated 4 pairs of final infographics produced by other participants. For each pair, the two infographics were generated using the exact same tabular dataset but with different systems by different authors. Participants were asked to judge which output was better overall or if they were of equal quality. This procedure resulted in $12 \times 4 = 48$ pairwise judgments.

6.3.3 Results

Numerical results. As shown in Table 5, OURS significantly reduced NASA-TLX workload relative to BASELINE (7.88±2.08 vs. 11.23±4.13; p=0.034), while satisfaction ratings also favored OURS but were not statistically significant (11.28±4.42 vs. 14.36±3.21; p=0.130). The number of dialogue turns was similar (5.17±1.47 vs. 4.83±1.70; p=0.594), suggesting that the lower workload reflected reduced coordination overhead rather than fewer iterations. In 48 blind pairwise judgments, outputs created with OURS were preferred more often than those from BASELINE (54.2% vs. 35.4%), with 10.4% ties. This suggests that the integrated workflow can reduce authoring friction without sacrificing output quality. Qualitative examples of the authoring results are provided in the Supplementary Section 7.

Qualitative feedback. Participants mainly described OURS as easier to work with because retrieval, preview, and iterative refinement remained in a single workspace. The most common positive comment was a *"one-stop"* experience (3 participants), followed by novice-friendly exemplar guidance (2) and better support for polishing simple aesthetics (2). In contrast, the dominant complaint about BASELINE was *multi-page operation and manual overhead* (4 participants), along with mismatches between retrieved exemplars and generated SVGs (2), and limited conversational memory (2). Figure 5 illustrates this contrast: in OURS, participants could keep multiple exemplars visible while iteratively inspecting and refining SVG drafts in place, whereas the baseline required manually moving artifacts across the retriever, chat interface, and external viewer. At the same time, several sessions revealed a *"debugging over designing"* pattern: later iterations often focused on repairing conversion issues such as missing pictograms, overwritten labels, or misaligned layouts rather than exploring richer alternatives. This suggests that current image-to-SVG generation already supports reasonably faithful exemplar adaptation, but higher-level creative exploration and aesthetic enrichment remain limited.

7 DISCUSSION AND LIMITATIONS

Taken together, the retrieval experiments demonstrate that our method improves retrieval quality, and the authoring study suggests that these gains translate into more effective infographic authoring. Meanwhile, the results point to some limitations and opportunities for future work.

Adaptation fidelity. While the integrated workflow supports exemplar reuse, adaptation quality still depends heavily on the robustness of the downstream SVG editing pipeline. As observed in the authoring study, users sometimes shifted from designing to debugging, repairing issues such as missing icons, overwritten labels, misaligned layouts, or disproportionate marks. These failures suggest that the current adaptation pipeline is still fragile at the execution level. Even when the intended direction is clear, the system does not always apply edits faithfully and cleanly. As P9 commented, "Once I know what I want to change, I still need the system to execute that change more reliably. Right now, some of the work becomes checking whether the generated SVG broke something else." This suggests the need for more robust structural representations, stronger edit planning, and constraint-aware SVG editing to improve the reliability of exemplar-based adaptation.

Draft quality and downstream handoff. Beyond execution-level errors, the system is better positioned as a tool for ideation and first-draft creation than as a substitute for professional refinement. P2 remarked, "It gets me to a first draft very quickly, which is already valuable, but I would still want to polish the spacing, alignment, and typography myself before using it in a final design." P7 similarly observed, "This feels strongest when I am still exploring options. Once I know what I want, I start wanting much finer control over individual elements than the current workflow gives me." Together, these comments suggest that the main value of the current system lies in accelerating exploration, reference combination, and early draft construction, rather than in producing fully polished artifacts without further intervention. Accordingly, rather than treating SVG generation as the endpoint of the workflow, future work should support a more seamless handoff from generated drafts to professional editing environments, where designers can refine typography, spacing, alignment, and other details.

8 CONCLUSION

We presented an intent-aware infographic retrieval framework for design inspiration and exemplar-based authoring. Motivated by a formative study of how users describe desired infographics, our method represents queries with multiple intent facets, rewrites and weights these facets explicitly, and performs facet-aware matching over an infographic corpus. We further integrated retrieval with SVG-based adaptation in a conversational workflow that supports search, selection, reuse, and iterative refinement. The evaluation results demonstrate our method consistently outperformed strong retrieval baselines and reduced authoring workload while maintaining or improving output quality. We hope this work helps move infographic authoring support from generic search toward intent-aware, exemplar-centered systems that better align with how people actually seek and reuse visual inspiration.

SUPPLEMENTAL MATERIALS

The supplementary document includes (1) the query-parsing prompt templates and structured five-facet schema, (2) the chart-type soft-matching table and taxonomy overview, (3) additional details of the interactive interface and the SVG handling mechanism, (4) the human-written query and human-judgment protocols used in the evaluation, and (5) selected authoring-system case pairs discussed in the paper. The code will also be included in the supplementary package.

REFERENCES

[1] H. K. Bako, X. Liu, L. Battle, and Z. Liu. Understanding how designers find and use data visualization examples. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):1048–1058, 2023. doi: 10.1109/TVCG.2022.3209490 1, 2

[2] A. J. Chaney, D. M. Blei, and T. Eliassi-Rad. A probabilistic model for using social networks in personalized item recommendation. In *Proceedings of the 9th ACM Conference on Recommender Systems*, RecSys '15, 8 pages, p. 43–50. Association for Computing Machinery, New York, NY, USA, 2015. doi: 10.1145/2792838.2800193 8

[3] Q. Chen, Y. Chen, R. Zou, W. Shuai, Y. Guo, J. Wang, and N. Cao. Chart2Vec: A universal embedding of context-aware visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 31(4):2167–2181, 2025. doi: 10.1109/TVCG.2024.3383089 2

[4] W. Cui, J. Wang, H. Huang, Y. Wang, C.-Y. Lin, H. Zhang, and D. Zhang. A mixed-initiative approach to reusing infographic charts. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):173–183, 2022. doi: 10.1109/TVCG.2021.3114856 1, 2

[5] R. Davis, X. Pu, Y. Ding, B. D. Hall, K. Bonilla, M. Feng, M. Kay, and L. Harrison. The risks of ranking: Revisiting graphical perception to model individual differences in visualization performance. *IEEE Transactions on Visualization and Computer Graphics*, 30(3):1756–1771, 2024. doi: 10.1109/TVCG.2022.3226463 2

[6] V. Dibia and C. Demiralp. Data2vis: Automatic generation of data visualizations using sequence-to-sequence recurrent neural networks. *IEEE Computer Graphics and Applications*, 39(5):33–46, 2019. doi: 10.1109/MCG.2019.2924636 2

[7] S. Elaldi and T. Çifçi. The effectiveness of using infographics on academic achievement: A meta-analysis and a meta-thematic analysis. *Journal of Pedagogical Research*, 5(4):92–118, 2021. 1

[8] G. Guo, S. Das, J. Zhao, and A. Endert. More like vis, less like vis: Comparing interactions for integrating user preferences into partial specification recommenders. *IEEE Transactions on Visualization and Computer Graphics*, 31(12):10328–10339, 2025. doi: 10.1109/TVCG.2025.3596541 2

[9] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati, eds., *Human Mental Workload*, vol. 52 of *Advances in Psychology*, pp. 139–183. North-Holland, 1988. doi: 10.1016/S0166-4115(08)62386-9 9

[10] K. Z. Hu, M. A. Bakker, S. Li, T. Kraska, and C. A. Hidalgo. VizML: A machine learning approach to visualization recommendation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, article no. 128, pp. 128:1–128:12. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3290605.3300358 2

[11] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In M. Meila and T. Zhang, eds., *Proceedings of the 38th International Conference on Machine Learning*, vol. 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 18–24 Jul 2021. 1, 2

[12] B. Kovacs, P. O'Donovan, K. Bala, and A. Hertzmann. Context-aware asset search for graphic design. *IEEE Transactions on Visualization and Computer Graphics*, 25(7):2419–2429, 2019. doi: 10.1109/TVCG.2018.2842734 2

[13] B. Lee, S. Srivastava, R. Kumar, R. Brafman, and S. R. Klemmer. Designing with interactive example galleries. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, 10 pages, p. 2257–2266. Association for Computing Machinery, New York, NY, USA, 2010. doi: 10.1145/1753326.1753667 1, 2

[14] H. Li, Y. Wang, A. Wu, H. Wei, and H. Qu. Structure-aware visualization retrieval. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, article no. 409, pp. 409:1–409:14. Association for Computing Machinery, New York, NY, USA, 2022. doi: 10.1145/3491102.3502048 2

[15] H. Li, Y. Wang, S. Zhang, Y. Song, and H. Qu. KG4Vis: A knowledge graph-based approach for visualization recommendation. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):195–205, 2022. doi: 10.1109/TVCG.2021.3114863 2

[16] Z. Li, D. Li, Y. Guo, X. Guo, B. Li, L. Xiao, S. Qiao, J. Chen, Z. Wu, H. Zhang, X. Shu, and S. Liu. ChartGalaxy: A dataset for infographic chart understanding and generation. In *The Fourteenth International Conference on Learning Representations*, 2026. 4, 5

[17] S. Liu, W. Yang, J. Wang, and J. Yuan. *Visualization for Artificial Intelligence*. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-031-75340-4 1

[18] Z. Liu, J. Thompson, A. Wilson, M. Dontcheva, J. Delorey, S. Grigg, B. Kerr, and J. Stasko. Data Illustrator: Augmenting vector design tools with lazy data binding for expressive visualization authoring. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, 13 pages, p. 1–13. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3173574.3173697 2

[19] J. Mackinlay, P. Hanrahan, and C. Stolte. Show Me: Automatic presentation for visual analysis. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1137–1144, 2007. doi: 10.1109/TVCG.2007.70594 2

[20] D. Moritz, C. Wang, G. L. Nelson, H. Lin, A. M. Smith, B. Howe, and J. Heer. Formalizing visualization design knowledge as constraints: Actionable and extensible models in draco. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):438–448, 2019. doi: 10.1109/TVCG.2018.2865240 2

[21] A. Narechania, A. Srinivasan, and J. Stasko. NL4DV: A toolkit for generating analytic specifications for data visualization from natural language queries. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):369–379, 2021. doi: 10.1109/TVCG.2020.3030378 2

[22] H. N. Nguyen and N. Gehlenborg. Safire: Similarity framework for visualization retrieval. In *2025 IEEE Visualization and Visual Analytics (VIS)*, pp. 246–250, 2025. doi: 10.1109/VIS60296.2025.00055 2

[23] A. Nowak, F. Piccinno, and Y. Altun. Multimodal chart retrieval: A comparison of text, table and image based approaches. In K. Duh, H. Gomez, and S. Bethard, eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5488–5505. Association for Computational Linguistics, Mexico City, Mexico, June 2024. doi: 10.18653/v1/2024.naacl-long.307 2

[24] M. Oppermann, R. Kincaid, and T. Munzner. VizCommender: Computing text-based similarity in visualization repositories for content-based recommendations. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):495–505, 2021. doi: 10.1109/TVCG.2020.3030387 2

[25] C. Qian, S. Sun, W. Cui, J.-G. Lou, H. Zhang, and D. Zhang. Retrieve-Then-Adapt: Example-based automatic generation for proportion-related infographics. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):443–452, 2021. doi: 10.1109/TVCG.2020.3030448 2

[26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In M. Meila and T. Zhang, eds., *Proceedings of the 38th International Conference on Machine Learning*, vol. 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. 1, 2, 7

[27] D. Ren, B. Lee, and M. Brehmer. Charticulator: Interactive construction of bespoke chart layouts. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):789–799, 2019. doi: 10.1109/TVCG.2018.2865158 2

[28] B. Saleh, M. Dontcheva, A. Hertzmann, and Z. Liu. Learning style similarity for searching infographics. In *Proceedings of the 41st Graphics Interface Conference*, GI '15, 6 pages, p. 59–64. Canadian Information Processing Society, CAN, 2015. 1, 2

[29] A. Satyanarayan, D. Moritz, K. Wongsuphasawat, and J. Heer. Vega-lite: A grammar of interactive graphics. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):341–350, 2017. doi: 10.1109/TVCG.2016.2599030 2

[30] Z. Shuai, B. Li, S. Yan, Y. Luo, and W. Yang. Deepvis: Bridging natural language and data visualization through step-wise reasoning. *IEEE Transactions on Visualization and Computer Graphics*, 32(1):868–878, 2026. doi: 10.1109/TVCG.2025.3634645 2

[31] K. Son, D. Choi, T. S. Kim, Y.-H. Kim, and J. Kim. GenQuery: Supporting expressive visual search with generative models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, article

no. 180, pp. 180:1–180:19. Association for Computing Machinery, New York, NY, USA, 2024. doi: 10.1145/3613904.3642847 2

[32] Y. Song, X. Zhao, R. C.-W. Wong, and D. Jiang. RGVisNet: A hybrid retrieval-generation neural framework towards automatic data visualization generation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, 10 pages, p. 1646–1655. Association for Computing Machinery, New York, NY, USA, 2022. doi: 10.1145/3534678.3539330 2

[33] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, O. Hénaff, J. Harmsen, A. Steiner, and X. Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 2, 7

[34] A. Tyagi, J. Zhao, P. Patel, S. Khurana, and K. Mueller. Infographics Wizard: Flexible Infographics Authoring and Design Exploration. *Computer Graphics Forum*, 2022. doi: 10.1111/cgf.14527 2

[35] C. Wang, J. Thompson, and B. Lee. Data Formulator: Ai-powered concept-driven visualization authoring. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):1128–1138, 2024. doi: 10.1109/TVCG.2023.3326585 2

[36] Y. Wang, Z. Hou, L. Shen, T. Wu, J. Wang, H. Huang, H. Zhang, and D. Zhang. Towards natural language-based visualization authoring. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):1222–1232, 2023. doi: 10.1109/TVCG.2022.3209357 2

[37] Y. Wang, H. Zhang, H. Huang, X. Chen, Q. Yin, Z. Hou, D. Zhang, Q. Luo, and H. Qu. InfoNice: Easy creation of information graphics. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, 12 pages, p. 1–12. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3173574.3173909 2

[38] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Towards a general-purpose query language for visualization recommendation. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, HILDA '16, article no. 4, 6 pages. Association for Computing Machinery, New York, NY, USA, 2016. doi: 10.1145/2939502.2939506 2

[39] K. Wongsuphasawat, D. Moritz, A. Anand, J. D. Mackinlay, B. Howe, and J. Heer. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):649–658, 2016. doi: 10.1109/TVCG.2015.2467191 2

[40] K. Wongsuphasawat, Z. Qu, D. Moritz, R. Chang, F. Ouk, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Voyager 2: Augmenting visual analysis with partial view specifications. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 2648–2659. Association for Computing Machinery, New York, NY, USA, 2017. doi: 10.1145/3025453.3025768 2

[41] W. Yang, M. Liu, Z. Wang, and S. Liu. Foundation models meet visualizations: Challenges and opportunities. *Computational Visual Media*, 10(3):399–424, May 2024. doi: 10.1007/s41095-023-0393-x 2

[42] L.-P. Yuan, Z. Zhou, J. Zhao, Y. Guo, F. Du, and H. Qu. InfoColorizer: Interactive recommendation of color palettes for infographics. *IEEE Transactions on Visualization and Computer Graphics*, 28(12):4252–4266, 2022. doi: 10.1109/TVCG.2021.3085327 2

[43] S. Zhang, H. Li, H. Qu, and Y. Wang. AdaVis: Adaptive and explainable visualization recommendation for tabular data. *IEEE Transactions on Visualization and Computer Graphics*, 30(9):5923–5938, 2024. doi: 10.1109/TVCG.2023.3316469 2

[44] J. Zhou, Y. Xiong, Z. Liu, Z. Liu, S. Xiao, Y. Wang, B. Zhao, C. J. Zhang, and D. Lian. MegaPairs: Massive data synthesis for universal multimodal retrieval. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19076–19095. Association for Computational Linguistics, Vienna, Austria, July 2025. doi: 10.18653/v1/2025.acl-long.935 2, 5, 7, 8

[45] T. Zhou, J. Huang, and G. Y.-Y. Chan. Epigraphics: Message-driven infographics authoring. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, article no. 200, 18 pages. Association for Computing Machinery, New York, NY, USA, 2024. doi: 10.1145/3613904.3642172 2

[46] C. Zhu-Tian, Y. Wang, Q. Wang, Y. Wang, and H. Qu. Towards automated infographic design: Deep learning-based auto-extraction of extensible timeline. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):917–926, 2020. doi: 10.1109/TVCG.2019.2934810 2