GeoNarrative: Transforming Geospatial Textual Data into Narrative Storytelling Visualization

Ruixian Ma, Biru Cao, Winston Yap, Jiaxin Gao, Wenzhe Peng,
Takehiko Nagakura, Fábio Duarte, Paolo Santi, and Carlo Ratti

Abstract—Geographic narrative visualization integrates maps, charts, and explanatory text into cohesive spatial stories and plays an important role in journalism, urban planning, and public communication. However, creating these narratives still requires collaboration among GIS specialists, data analysts, and web developers, creating a high barrier to adoption. Existing visualization tools rely on static dashboards or rigid templates with steep learning curves, while large language models (LLMs) alone struggle to produce spatially grounded, visually coherent storytelling pages. To bridge this gap, we introduce GeoNarrative, an LLM-augmented system that transforms natural-language queries into multi-module narrative pages composed of text, maps, and charts. GeoNarrative includes three components: (i) an agent-based scheduling module for intent understanding and spatial data retrieval, (ii) a CLIP-based layout retriever that selects professionally inspired page templates, and (iii) a generative visualization engine that populates content while preserving narrative coherence. A built-in provenance mechanism lets users trace every data element to its origin, whether a named authoritative source or the generating model, supporting transparency and trust. We evaluate GeoNarrative through a controlled comparative experiment (N=16), an expert usability study (N=8), and a component-level ablation study. The results indicate that GeoNarrative produces narrative pages with stronger information flow and visual coherence than baseline methods, with large effect sizes, and that each architectural component contributes meaningfully to overall quality; given the modest sample sizes we regard these findings as promising rather than definitive. We also release an annotated dataset of 5,321 modules from 199 real-world narrative visualization pages to support future research.

Index Terms—Narrative visualization, geospatial visualization, storytelling, large language models, automatic visualization generation, data provenance.

1 Introduction

Geospatial narrative visualization plays a critical role in how societies understand spatially grounded information: journalists contextualize election results and disasters with map-driven stories, urban planners communicate zoning and transit proposals to stakeholders, and public-health agencies convey disease spread and resource allocation. Yet producing such narratives remains labor-intensive: a typical workflow requires GIS specialists, data scientists, graphic designers, and editorial staff working across incompatible software environments, a production bottleneck that limits the timeliness and reach of geospatial communication. Large language models (LLMs), with their capacity for natural-language understanding, structured reasoning, and multi-step task decomposition, offer an opportunity to automate much of this coordination, enabling a single analyst to go from a plain-language query to a publication-ready narrative visualization with minimal specialist intervention.

1.1 Problem Statement

Recent studies have extended LLMs' geospatial intelligence by combining map data (GeoLLM) [40], applying domain-specific fine-tuning (K2) [13], and introducing task decomposition strategies (UrbanLLM) [28]. These advances highlight the growing potential of LLMs for spatial reasoning and urban problem-solving. However, existing work primarily employ LLMs as *analytical agents* for processing and interpreting geospatial data. In contrast, relatively little attention has been given to spatial visualization as a *communication interface*. This gap separates LLM-driven geospatial analysis from the narrative visualizations that audiences increasingly expect. Geospatial narrative storytelling involves more than map-based analysis; it depends on the coordinated use of text, cartographic representations, statistical charts, and imagery that explain and reinforce one another. It is also important to distinguish between *storytelling visualization* (the design of individual visual encodings that convey a narrative message, such as an annotated choropleth map) and *storyboard creation* (the higher-level orchestration of multiple heterogeneous modules into a coherent, page-level narrative). Our work focuses on the latter: the end-to-end composition of multi-module storyboards from natural-language input.

1.2 Research Objectives

This study aims to automatically interpret the structure of *multi-module geospatial information* (structured compositions of maps, charts, narrative text, and images that jointly represent spatial data and relationships) and convert natural-language input into well-designed narrative visualizations, thereby reducing reliance on multi-specialist teams. We pursue four objectives: (i) *text-to-visualization conversion*, transforming queries into aesthetically coherent multi-module storyboards; (ii) *module harmonization*, aligning narratives with their visual elements for textual and text–visual coherence; (iii) a reusable *annotated dataset* capturing module types and narrative functions; and (iv) *data provenance*, so that every map layer, statistic, or claim can be traced to its source. The corresponding contributions are detailed below.

1.3 Contributions and Overview

This paper makes three primary contributions. First, we propose a three-component architecture consisting of *LLM Geo Agent*, *CLIP-based Layout Retrieval*, and *Generative Visual Representation*, which together enable the end-to-end transformation of natural-language queries into multi-module geospatial narrative pages. Second, we construct and release a manually annotated dataset of 5,321 modules from 199 real-world narrative visualization pages. We also introduce a functional labeling taxonomy to support downstream research in layout understanding and storytelling optimization. Third, through a controlled

- *Ruixian Ma, Winston Yap, Fábio Duarte, Paolo Santi, and Carlo Ratti are with the Senseable City Lab, Massachusetts Institute of Technology, Cambridge, MA, USA. E-mail: ruixian@mit.edu (corresponding author), winyap@mit.edu, fduarte@mit.edu, psanti@mit.edu, ratti@mit.edu.*
- *Biru Cao, Jiaxin Gao, and Wenzhe Peng are independent researchers, Cambridge, MA, USA. E-mail: birucao@mit.edu, giagao@qq.com, pwz@alum.mit.edu.*
- *Takehiko Nagakura is with the School of Architecture and Planning, Massachusetts Institute of Technology, Cambridge, MA, USA. E-mail: takehiko@mit.edu.*

Manuscript received xx xxx. 202x; accepted xx xxx. 202x. Date of Publication xx xxx. 202x; date of current version xx xxx. 202x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.202x.xxxxxxx

comparative experiment (N=16) and an expert usability study (N=8), we find that GeoNarrative produces narrative storyboards rated higher in information flow, readability, and visual coherence than baseline methods, with large effect sizes; we treat these results as promising evidence given the modest sample sizes rather than definitive proof. It also removes the need for specialized expertise in coding, design, or geospatial analysis. The remainder of the paper is organized as follows. Section 2 reviews related work on dashboard visualization, generative AI for visualization, and generative AI for geospatial applications. Section 3 presents the three-component methodology. Section 4 describes the dataset construction and annotation process. Section 5 presents case studies. Section 6 reports the evaluation results. Section 7 discusses broader implications. Section 8 addresses limitations and future directions.

2 Related Work

We position our work at the intersection of four research areas: data storytelling and narrative visualization, geospatial visualization and geo-narrative, dashboard tools and automated visualization generation, and generative AI for geospatial visualization.

2.1 Data Storytelling and Narrative Visualization

Segel and Heer [49] established the foundational design space for *narrative visualization*, identifying seven genres and dimensions such as visual narrative tactics and structure; Hullman and Diakopoulos [27] examined how editorial choices frame interpretation, and Stolper et al. [52] consolidated storytelling design patterns across journalism and analysis.

Building on these foundations, automated systems generate data stories from tabular data: Calliope [50] organizes facts into a coherent sequence with appropriate charts and layouts, DataShot [59] composes extracted facts into fact-sheet infographics, and Erato [53] supports cooperative story editing through fact interpolation.

Most of these automated systems, however, target chart-level generation or single-page layout arrangement rather than the *storyboard creation* task defined in Section 1.1; our work addresses the latter in a geospatial context, orchestrating maps, charts, text, and images into multi-module narrative pages.

2.2 Geospatial Visualization and Geo-Narrative

Maps have long supported storytelling. Caquard and Cartwright [7] introduced *narrative cartography* for the intersection of maps and narrative across oral, written, and audio-visual forms, and Roth [47] surveyed cartographic design as visual storytelling, proposing a design space of narrative elements, genres, and tropes adapted to geographic phenomena. These works show that cartographic storytelling is a design practice with its own rhetorical and compositional principles.

In practice, interactive geo-narrative platforms such as ESRI's ArcGIS StoryMaps [16] and Knight Lab's StoryMapJS [30] let authors combine maps, text, and media into scrollable stories. Latif et al. [32] studied the visualization–text interplay in geographic data-driven stories, highlighting the gap between geo-narratives' expressive potential and the authoring support available. These platforms still demand substantial manual effort (configuring map views, basemaps, annotations, and media by hand), so producing a coherent multi-module geo-narrative is time-consuming and requires both cartographic and web-development expertise, which motivates our work on automating end-to-end geo-narrative generation while preserving spatial and thematic coherence.

2.3 Dashboard Tools and Automated Visualization Generation

Dashboard tools such as Tableau [55] and Power BI [42] offer rich interactive environments for exploratory analysis. As Few [18] observed, dashboards are designed for parallel monitoring of key indicators rather than interpretive storytelling; they lack the semantic consistency and deliberate sequencing that Segel and Heer [49] identify as central to narrative visualization, as well as semantically driven template matching for composing coherent multi-module pages.

To address these limitations, generative AI has expanded visualization beyond the dashboard paradigm. Research in this area includes chart generation, natural language-driven visualization, recommendation, and layout synthesis. Early systems such as Vega-Lite [48] established a declarative foundation for automated chart specification. Data2Vis [15] and Table2Charts [64] explored mappings from structured data to chart types, while AutoCaption [35] demonstrated bidirectional generation between charts and textual descriptions.

Natural Language-Driven Generation. With the rise of large language models (LLMs), natural language has become a prominent input modality. Recent systems convert user queries into visualizations or executable code [12, 14, 38, 56]. NL4DV and NL2INTERFACE [9, 44] parse natural language into structured specifications and layouts supporting interactive generation, while domain-specific systems such as SNIL [11] leverage LLMs to generate sports news narratives from structured insights.

Visualization Recommendation and Exploration. A complementary direction focuses on automated recommendation. SeeDB [57] proposes suggestions based on database queries, CompassQL [60] supports semantically driven design-space exploration, VizML [26] uses machine learning to predict suitable chart types, and Draco [43] formalizes design knowledge into constraint models. More recently, InsightPilot [36] generates analysis steps and visualizations from natural language questions, and LLM4Vis [58] produces chart recommendations through prompt-based reasoning in few-shot and zero-shot scenarios.

Sketch, Image, and Layout-Based Generation. Beyond natural language, researchers have explored alternative input modalities. LADV [37] reconstructs dashboards from sketch and image input. At the narrative page level, LayoutDiffusion [63] and LayoutGPT [17] apply generative approaches to multi-module layout synthesis, demonstrating how semantic content can be translated into structured spatial arrangements. Recent surveys have also systematically evaluated the capabilities and limitations of LLMs for visualization generation [33] and reviewed the broader opportunities and challenges of generative AI for visualization [1]. Together, these studies highlight persistent issues in reliability, semantic understanding, and user trust.

However, these approaches rarely combine natural-language understanding with geospatial reasoning: they treat maps as one chart type rather than the narrative backbone, addressing fact sequencing or layout generation without embedding geospatial semantics or ensuring spatial narrative coherence.

2.4 Generative AI for Geospatial Visualization

Generative methods have advanced rapidly in the geospatial domain along two main trajectories.

Deep Generative Models for Map Creation. Deep learning and generative adversarial networks (GANs) [22] have been widely applied to map generation tasks. GANmapper [61] transforms geographic data into maps, GeoGAN [21] introduces conditional GANs with reconstruction and style losses to generate standard map layers from satellite imagery, and SMAPGAN [8] enables semi-supervised stylized map tile generation. More recent work such as BEVGen [54] generates street-view images from bird's-eye layouts. Although these approaches produce visually realistic maps, they generally lack semantic context and narrative structure.

LLM-Driven Geographic Visualization and Analysis. LLMs are increasingly integrated with mapping tools to support geographic visualization and spatial reasoning. MapGPT [62] proposes an autonomous mapping framework combining LLMs with mapping tools; ChatGeoAI [39] and GeoLLM [40] enable spatial question answering and geographic knowledge extraction from text. However, most of these systems are limited to presenting results from text queries in a single module (e.g., a standalone map). Google's Geospatial Reasoning framework [23] envisions combining multimodal foundation models with generative AI for spatial reasoning, but remains at a conceptual stage.

In summary, prior work has advanced narrative visualization theory, geo-narrative design, dashboard tools, and geospatial AI as separate

threads. No existing approach integrates them to automatically generate multi-module, semantically coherent geo-narrative pages: the gap GeoNarrative addresses by combining LLM semantics with structured composition of maps, text, charts, and images for both narrative expressiveness and spatial coherence.

3 METHODOLOGY

Our goal is to develop an LLM-based retrieval-driven visualization system for generating narrative visualizations. Given a user query and summary, the system automatically produces a visualization page with a coherent structure and storytelling flow. The page consists of multiple modules, including text, charts, maps, images, and interactive components, each serving a distinct communicative function.

The system design is grounded in both empirical analysis and prior theory. We first conducted a systematic review of 199 professionally produced narrative visualization pages, detailed in Section 4, from nine major outlets. Through this review, we identified recurring structural patterns in the composition of spatial stories. We also drew on the narrative visualization pipeline articulated by Segel and Heer [49] and the layered grammar of visual storytelling proposed by Stolper et al. [52]. This combined analysis revealed three core stages of spatial story construction: (i) *intent interpretation and data marshalling*, in which the communicative goal is parsed and requisite data are assembled; (ii) *layout selection*, in which a compositional template is chosen to match the narrative genre and visual style; and (iii) *content generation and rendering*, in which individual modules are populated and coordinated into a coherent page. These stages correspond directly to the three system components described below.

Our system consists of three main components: (1) LLMs Geo Agent Component interprets user intent, schedules content, and generates intermediate representations (IR); (2) CLIP-based Layout Retrieval model retrieves the optimal template through matching embeddings from the IR; (3) Generative Visual Representation Engine populates module content, optimizes narrative logic, and renders the final visualization page. The LLM Geo Agent module consists of the following three stages: (see Figure 1)

3.1 LLMs Geo Agent

3.1.1 Input Understanding & Agent-based Content Scheduling

Input Understanding In the first stage, prompt engineering structures the user input in three steps: (1) the LLM selects the most relevant topic from eight predefined categories (e.g., economy, society, health); (2) it extracts geographic entities and temporal references into a unified JSON structure; and (3) it chooses controlled style settings: visual style (flat-UI, minimalist, retro, or realistic), density (less-text, balanced, or data-dense), and color scheme (dark, light, or gradient). This structured output lets downstream modules consume consistent, machine-readable instructions.

Agent-based Content Scheduling At the same time, an agent-based mechanism handles data retrieval, API calls, and summary generation: for geographic or route queries it selects an appropriate spatial data source (e.g., the Google Directions API or the American Community Survey) and retrieves results such as polyline paths or census fields, then generates a natural-language description of their geospatial meaning. If no suitable API source is available, the system falls back to prompt-guided LLM inference to generate structured results with explanatory text.

3.1.2 Intermediate Representation Construction

The system integrates the extracted intent and retrieved data into a unified intermediate representation (IR), a structured JSON object that serves as input for layout retrieval and content generation. The IR encodes the semantic intent (`core_topic`, `zoom_strategy`), retrieved or generated data (`api_data`, `selected_fields`, `location_field_map`), user-preferred aesthetics (`visual_style`, `visual_density`, `color_palette`), and a ∼300-word LLM-generated narrative summary (`description`). For example, the city query *"compare income and education across Boston and New York"* yields an IR with `core_topic: economy`, `zoom_strategy: city`, `selected_fields: [B19013, B15003]` (ACS median income and degree attainment), `target_cities: [Boston, New_York]`, `visual_density: balanced`, and `color_palette: light`, a single machine-readable object that downstream modules consume without re-parsing the natural-language query.

Algorithm 1: CLIP-based Layout Retrieval

Input: key_query q, required map count m, parallel flag p, template embeddings E
Output: Top-k retrieved layouts R

$q_{emb} \leftarrow$ Encode(q) using CLIP;
$S \leftarrow$ CosineSimilarity(q_{emb}, E);
$M \leftarrow$ ApplyConstraints(E, map_count=m, parallel=p);
$S[\neg M] \leftarrow -\infty$; `// Filter out invalid templates`
$R \leftarrow$ TopK(S, k);
return R;

3.2 CLIP-based Layout Retrieval

To recommend layouts matching user intent, we built an embedding library from the 199 annotated template pages, each labeled with semantic attributes (topic category, visual style, information density, color scheme, narrative zoom level). Using CLIP [45] (ViT-B/32), we encode each attribute's top-three labels and combine them by weighted fusion (0.1 topic, 0.3 visual style, 0.2 density, 0.3 color, 0.1 zoom), weighting visual style and color most heavily as they most shape a page's appearance. A sensitivity check confirms robustness: re-ranking under ±50% per-attribute perturbations leaves the top-3 retrieved set 72% stable on average (67% under uniform weights), with the lowest-weighted attribute (narrative zoom) least influential; combined with the hard structural filters (Section 3.2), this shows the selection is not hypersensitive to the exact weights. The fused, normalized embedding is stored per template with metadata such as map count and parallel-display flag, and users may adjust the weights to reflect their design preferences. Because each template is represented solely by its semantic attribute labels and a CLIP embedding, the library is extensible: users can register their own templates by annotating new pages with the same attribute schema, without retraining any model.

When processing a user query, the system concatenates the parsed semantic intent into a `key_query` string (e.g., `core_topic = economy`, `visual_style = flat_UI`, `visual_density = balanced`, `color_palette = dark`, `zoom_strategy = city`). It then encodes this query with CLIP into the same embedding space as the template library.

Before computing similarity, the system applies hard structural filters. Candidate templates must satisfy constraints such as the required number of maps and, optionally, the use of a parallel display. If no candidate satisfies both constraints, the system relaxes the parallel-display condition while preserving the map requirement.

The system then computes the cosine similarity between the query embedding and each candidate template embedding. It selects the top k candidates, with $k = 3$ by default, and returns the template identifier together with its metadata, including similarity score, map count, and parallel-display flag. By combining semantic alignment with structural constraints, this retrieval strategy recommends layouts that balance narrative coherence and visual logic.

The overall procedure is summarized in Algorithm 1, where q denotes the key_query string, m the required map count, p the parallel flag, E the templates embedding library, and R the final set of retrieved layouts.

3.3 Generative Visual Representation

3.3.1 Module Content Generation & Optimization

Once the system selects a layout, it generates the content for each module. Current modules include *headline*, *narrative*, *map*, and *chart*. The system will generate the required chart data and text content through

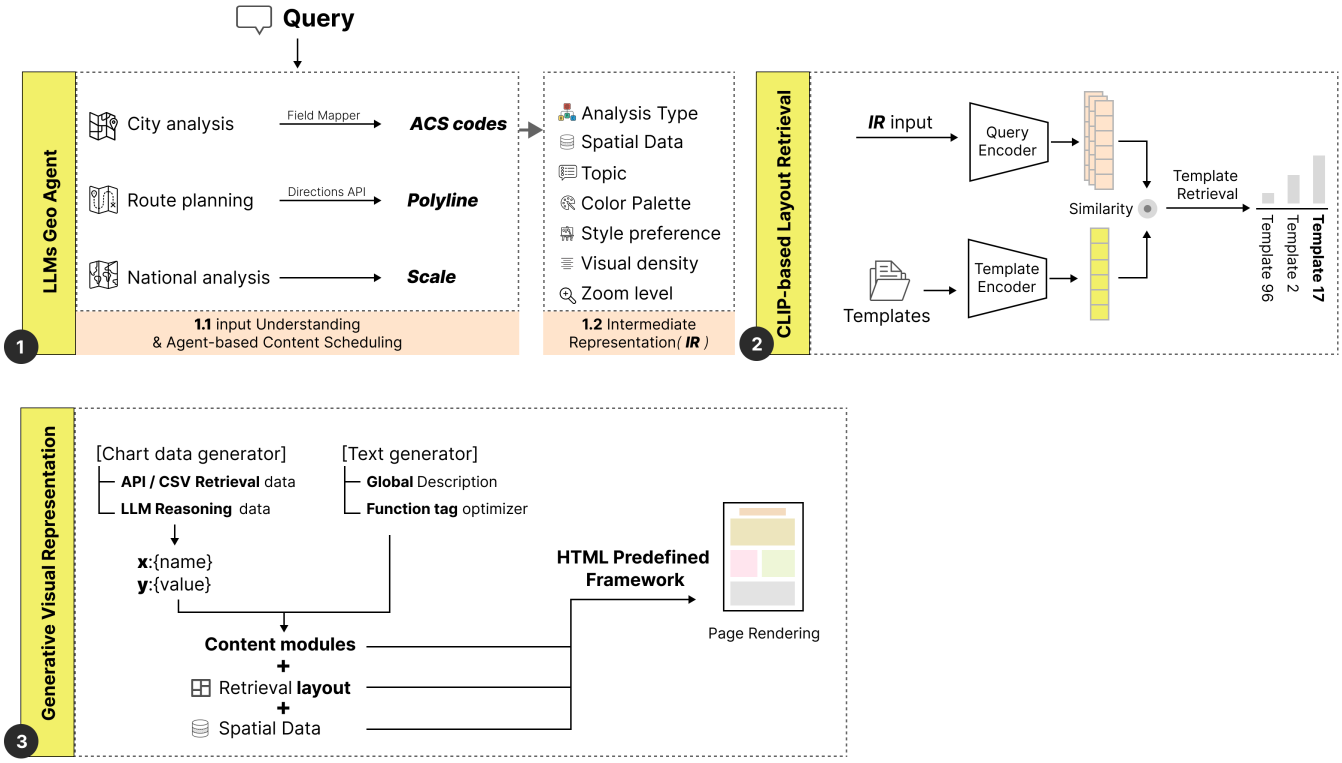


Figure 1: The pipeline consists of three steps: (1) LLMs Geo Agent, which parses queries and generates intermediate representations (IR); (2) CLIP-based Layout Retrieval, which encodes both IR and templates into a shared embedding space and retrieves the most semantically aligned layout; and (3) Visual Generation & Rendering, which populates content modules, optimizes narrative logic, and renders the final page.

prompt calls to LLM or external APIs based on the `function_tag` (such as Explanation, Highlight, Overview, etc.) and `chart_type` (such as scatter, bar, line, etc.) of each module. It also considers module position and overall page semantics.

We use D3 [3] as the front-end charting framework. The text generator produces global summaries and local narratives, the chart generator constructs visualizations through LLM inference or API calls, and the function-label optimizer adjusts each module's narrative role, together enabling tight text–chart coupling and a coherent narrative structure.

`Map Generation` : Map generation is based on the geographic data processed `api_data` field of the intermediate representation (IR). Geographic queries are drawn based on the American Community Survey (ACS) data in `location_field_map`. The system uses the Mapbox [41] basemap as the background and overlaying render layers such as path lines for routes or choropleth maps for attribute distributions.

`Chart Data Generation` : Chart data can be generated in two ways. When available, the system retrieves external datasets through structured prompts, such as ACS CSV tables. When such data is unavailable, the LLM generates semantically consistent synthetic data based on the user query and module function tag. In both cases, the results are organized into chart-specific formats, such as `"x": "resort elevation"` and `"y": "snow days"`.

`Text Generation and Cross-Module Coherence` : Text generation combines external data retrieval with LLM reasoning. At the global level, the LLM produces a narrative summary of the overall visualization. At the module level, it generates targeted explanations, such as trend interpretation and comparative analysis. To ensure coherence across modules, the system follows a two-step process. First, each module is assigned a function tag, such as **Overview**, **Highlight**, or **Explanation**, that defines its semantic role. Second, these tags are embedded into prompts that condition LLM outputs for textual narratives, maps, and charts. For example, an **Overview** tag prompts the model to generate a global summary, such as *Overall, the bar chart above shows . . .*", while a **Highlight** tag elicits contrastive statements, such as *Notably, Region A has the highest . . .*". This tag-conditioned prompting ensures that generated content aligns with each module's semantic role. Together, these mechanisms promote coupling and coherence across modules and ensure both textual-textual and text-visual consistency.

`Theoretical Grounding of Function Tags` Our six function tags derive from established narrative visualization theory. Segel and Heer [49] identify *narrative strategies* (establishing context, directing attention, and building progression), to which we map our tags: **Overview** to the *establishing shot*; **Highlight** to *annotation and emphasis*; **Progression** to *ordering*; **Explanation** to *messaging* and causal reasoning; **Multi_View** to *juxtaposition* and *small multiples*; and **Illustrate** to *visual embellishment*. This ensures each module serves a well-defined communicative role rather than a decorative one.

3.3.2 Module Optimization and Page Rendering

Finally, the system optimizes spatial placement, text length, and visual proportions (adjusting parameters such as font size and inter-module spacing), and renders the page into a front-end-ready format with coherent narrative flow and clear visual guidance. The full pipeline takes about 30 to 65 seconds per generation.

3.3.3 Data Provenance and Source Tracing

Because GeoNarrative composes retrieved and model-generated content, every module carries a visible provenance marker so that readers can judge its trustworthiness. The mechanism operates at four levels. (1) *Source markers*: each module shows a colour-coded label in place, distinguishing a verified real source from AI-generated content.

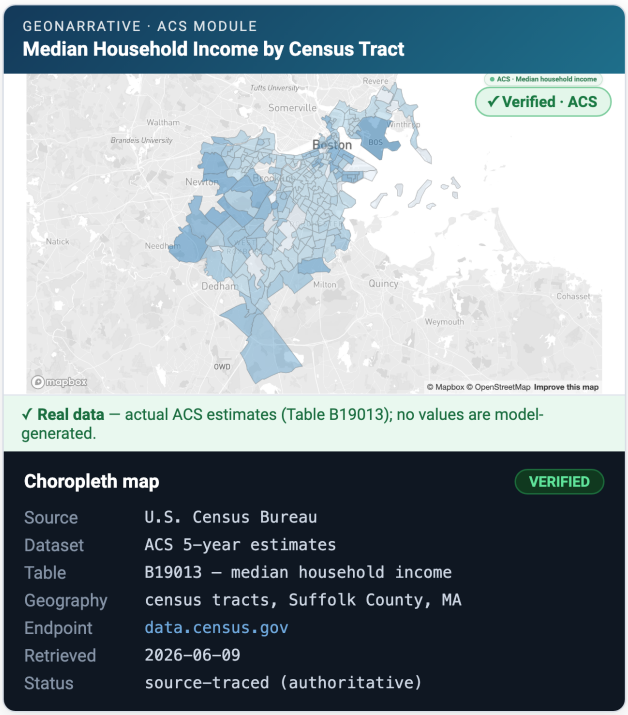


Figure 2: Data-provenance source panel for a generated city page. The income choropleth is traced to *real* U.S. Census ACS data; the panel names the dataset (ACS 5-year), table (B19013, median household income), geography (census tracts), endpoint (data.census.gov), and retrieval date, marking it *verified*. The page's model-composed narrative and chart modules are instead flagged as *AI-generated* and unverified, and a one-click toggle hides all markers. By naming authoritative sources and flagging generated content at the point of use, provenance becomes active, rather than passive, verification. The full generated page, with markers on every module, is in the supplemental materials.

(2) *Named origins*: real-data modules name their source (for example, a demographic choropleth reads *ACS: median household income, census-tract level*, and a route reads *Google Directions*), while model-generated narrative and chart modules are marked accordingly. (3) *Auditable detail*: hovering a module reveals its exact ACS fields or API endpoint, generating model, and retrieval timestamp. (4) *Non-intrusive integration*: a one-click toggle hides all markers, so provenance is available on demand without disrupting reading or evaluation (Figure 2). By naming real sources and flagging generated content at the point of use, the panel turns provenance from passive disclosure into legible, at-a-glance verification.

3.4 Implementation and Reproducibility

GeoNarrative is *model-agnostic*: the pipeline exposes typed interfaces (intent→IR→layout→content), so any instruction-following model can serve a given stage. Our reference implementation uses a local Llama-3-8B classifier (temperature 0) for intent routing among the route, city, and country pipelines; a frozen CLIP ViT-B/32 encoder for layout retrieval; and GPT-4o (temperature 0) to populate module narratives, headlines, and chart series. Real data is sourced rather than generated wherever possible: city demographic maps use real American Community Survey (ACS) statistics at the census-tract level (the agent selects the relevant ACS fields, such as median household income `B19013` and educational attainment `B15003`, and the choropleths display the actual values for the target cities), and route maps use polyline and waypoint coordinates from the Google Directions API. All generative LLM calls run at temperature 0, making the pipeline near-deterministic for a fixed query; we report exact model versions in place of random seeds. Real-data modules (ACS, Google Directions) are deterministic, and only model-generated narrative and synthetic chart series carry residual stochasticity, which temperature 0 minimizes.

4 Data and Label Design

To support our goal of understanding the configurational logic of visual representation in spatial narrative visualization, and its semantic and structural relationship to story content, we constructed a representative dataset of spatial narrative visualizations. This dataset provides the foundation for learning the form-content coupling mechanism in narrative visualization.

4.1 Source Selection

Prioritizing representativeness and narrative richness across design styles, themes, and media types (and following Cairo's observation [6] that outlets like *The New York Times*, *Financial Times*, and *Bloomberg* have distinct visual narrative styles), we collected and curated 199 unique pages from nine respected institutions: MIT Senseable City Lab, *The New York Times*, BBC, *Bloomberg*, FiveThirtyEight (538), *The Economist*, *Financial Times*, *The Washington Post*, and ArcGIS StoryMaps, spanning academic, news, business, public-platform, and data-journalism sources; duplicate and overlapping pages were removed during curation so that each page contributes a single template embedding.

4.2 Topic Selection

We focused on eight domains with strong geographic relevance: Politics, Society, Health, Economy, Environment, Technology, Education, Travel. These thematic categories were informed by the classification approach in *Data-Driven Storytelling* [46] and the narrative visualization framework proposed by Segel and Heer [49].

4.3 Annotation Process

We decomposed each page into discrete "visual semantic modules." Automatic detection methods (e.g., object detection on screenshots) proved unreliable due to heterogeneous layers such as semi-transparent overlays on map tiles, so we adopted a human-centered pipeline supported by GPT-4o pre-annotation.

The primary researcher completed all annotations following a detailed guidebook refined through a 20-page calibration phase. For pre-annotation, GPT-4o received each page's screenshot, extracted text, and structured prompt definitions for all module categories and function tags, outputting a JSON proposal that the annotator could accept, revise, or reject. This reduced annotation time from approximately 25 to 12 minutes per page.

The researcher then reviewed all 199 pages. For each page, the reviewer examined GPT-4o's proposal module by module against the guidebook, confirming or correcting each module's type and function tag and adjusting bounding-box extents where the proposal was imprecise. Because annotation began from a structured proposal rather than from scratch, this verification step, not labeling from a blank page, accounts for the reduction to about 12 minutes per page. Overall, 34.2% of GPT-4o labels were revised, primarily function tag reassignment (18.7%) and boundary adjustment (11.5%). To assess consistency, three researchers independently annotated 500 modules (∼10% of the dataset), achieving 97.2% raw agreement and Fleiss' $\kappa = 0.96$ [19], indicating almost perfect agreement [31].

4.4 Module Classification and Definition

We defined six module types: **Headline** (section titles), **Narrative** (main text driving story flow), **Annotation** (local explanations of charts or maps), **Map** (spatial data carriers), **Chart** (non-spatial charts), and **Image** (photographs or illustrations). In total, we identified 5,321 modules across 199 pages (Table 1). To capture each module's semantic role, we annotated all modules with six function tags: Overview, Highlight, Progression, Explanation, Multi-View, and Illustrate.

Among the map modules with zoom-level annotations, country-level views dominate (59.9%), followed by city (20.1%), state (14.0%), street (3.8%), and world (2.3%).

Table 1: Module-type distribution with dominant sub-categories (N = 5,321 modules across 199 pages).

Type	Dominant sub-categories (count)	Total (%)
Annotation	Map Title (350), Data Source (273), Map Legend (256)	1909 (35.9)
Narrative	Basic Describing (1201), External Links (40)	1310 (24.6)
Headline	Section Title (458), Main Title (197), Subtitle (88)	743 (14.0)
Map	2D Dimensions (635)	636 (12.0)
Chart	Line (67), Bar (53), Bar-Horiz. (46), Table (43)	362 (6.8)
Image	Photo (250), Map Images (70), Illustration (41)	361 (6.8)

Among the map modules with render-type annotations, choropleth maps are most common (41.7%), followed by dot maps (28.6%), heatmaps (9.7%), polylines (6.1%), satellite views (5.9%), and grids (5.3%).

To understand how modules are sequenced in professional geo-narratives and inform our template-based ordering, we conducted an n-gram analysis of module sequences. The analysis reveals that *annotation* modules act as connective tissue between content modules: the three most frequent bigrams are annotation→annotation (607), annotation→narrative (535), and narrative→annotation (417). The recurring trigram annotation→annotation→annotation (163) reflects the clustering of metadata elements (e.g., title, legend, data source) that typically surround a map or chart. The pattern headline→narrative (337) confirms that professionally designed pages consistently open sections with a headline followed by explanatory text. These sequential regularities inform our template-based generation strategy by establishing expected module orderings.

5 Case Study

5.1 Scenario 1: Election Journalism

Consider a journalist who needs to produce an interactive page comparing state-level 2024 U.S. election outcomes with LGBTQ support rates. Traditionally this requires a graphics team and hours of work. With GeoNarrative, she submits a natural-language query; the system classifies it as *politics*, retrieves election and demographic data via APIs (or generates structured data via LLM reasoning), selects a matching layout through CLIP retrieval, fills each module with targeted content, and exports the result as HTML, all in under 65 seconds. Figure 3 shows the generated page.

5.2 Scenario 2: Route Planning

Consider a user planning a multi-stop trip through New England and New York City. She describes the journey in natural language; the system extracts origin, waypoints, and destination, calls the Google Directions API to retrieve polyline path data, and renders the route on a Mapbox basemap with stop-level annotations and contextual narrative modules; the resulting route page is provided in the supplemental materials. Unlike the journalism scenario, route queries trigger the spatial path retrieval pipeline rather than demographic data APIs, demonstrating the system's ability to adapt its data acquisition strategy to the query type.

Across scenarios, each module's source is shown in place (Figure 2): demographic choropleths are traced to real ACS census statistics and routes to the Google Directions API, while model-generated narrative is explicitly marked, making the page's epistemic status transparent.

6 Evaluation

Our study was reviewed and approved by the Science and Technology Ethics Committee for Research Involving Humans of Tongji University (expedited review; approval no. TjdXSr2025121). At the start of each session we obtained informed consent from every participant via a signed consent form, and all data were anonymized before analysis.

We defined three complementary evaluation metrics: **Information Flow** (coherent sequencing of modules), **Narrative Plausibility** (whether the page reads as a believable story rather than loosely assembled facts), and **Logic** (internal consistency between claims and data). Together, these capture sequencing, storytelling quality, and structural coherence.

6.1 Controlled Comparative Experiment

We recruited 16 participants (9 male, aged 28–35, $M = 31.44$, $SD = 2.25$) to evaluate the information flow of four generation methods, each rated on a 1–5 Likert scale: GeoNarrative, GPT-only (ChatGPT 4o), Map-only, and Random Order (GeoNarrative's module content in shuffled sequence). All had professional or research experience: eight in design (urban planning, UI, or visualization design), eight in engineering or computer science, and three spanning both. Each participant completed five evaluation tasks: *New York City Commute* (NYC-Commute), *Oakland Career Comparison* (Oakland-Career), *Boston–New York Three-Day Itinerary* (BOS-NY), *2024 Election and Education Comparison* (Election-Edu), and *Los Angeles One-Day Cultural Tour* (LA-Tour).

Statistical Analysis. We used a non-parametric repeated-measures design: a **Friedman test** [20] for overall differences, followed by pairwise **Wilcoxon signed-rank tests** with **Bonferroni correction** ($\alpha_{\text{adjusted}} = 0.0083$). We reported Kendall's W and rank-biserial r as effect sizes.

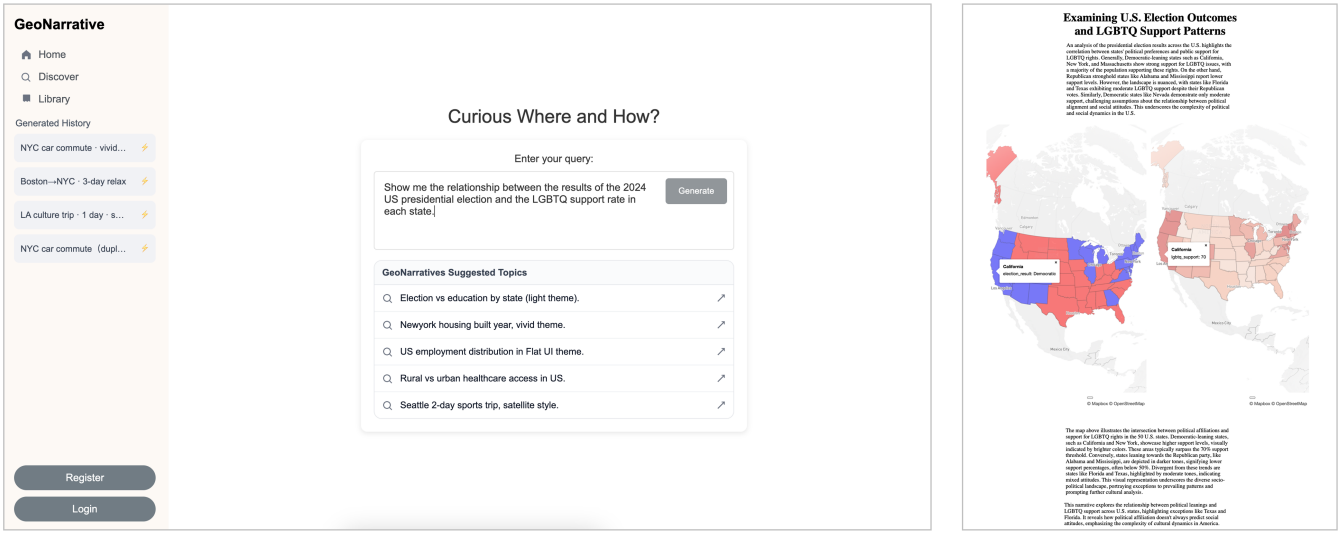
Statistical Results. A Friedman test on the participant-level mean ratings (averaged across five tasks) revealed a highly significant overall difference among the four generation methods, $\chi^2(3) = 31.43$, $p < .001$, with a large effect size (Kendall's $W = 0.655$). The Friedman test was also significant for each individual task: Election-Edu ($\chi^2(3) = 23.34$, $p < .001$, $W = 0.486$), LA-Tour ($\chi^2(3) = 29.85$, $p < .001$, $W = 0.622$), NYC-Commute ($\chi^2(3) = 18.88$, $p < .001$, $W = 0.393$), BOS-NY ($\chi^2(3) = 31.41$, $p < .001$, $W = 0.654$), and Oakland-Career ($\chi^2(3) = 28.59$, $p < .001$, $W = 0.596$).

Post-hoc pairwise Wilcoxon signed-rank tests with Bonferroni correction ($\alpha_{\text{adjusted}} = 0.0083$) showed that GeoNarrative ($Mdn = 4.60$) was rated significantly higher than all three alternatives: GPT-only ($Mdn = 2.95$; $Z = 3.52$, $p < .001$, $r = 0.88$), Map-only ($Mdn = 2.60$; $Z = 3.52$, $p < .001$, $r = 0.88$), and Random ($Mdn = 3.00$; $Z = 3.52$, $p < .001$, $r = 0.88$). All three comparisons involving GeoNarrative yielded large effect sizes ($r > 0.5$). In contrast, none of the pairwise comparisons among the three non-GeoNarrative conditions reached significance after Bonferroni correction: GPT-only vs. Map-only ($Z = 1.61$, $p = .108$, $r = 0.40$), GPT-only vs. Random ($Z = 0.91$, $p = .362$, $r = 0.23$), and Map-only vs. Random ($Z = 0.66$, $p = .509$, $r = 0.17$).

As shown in Figure 4, the statistical results revealed clear differences among the four generation methods.

GeoNarrative consistently received the highest ratings across all tasks ($M = 4.50$, $SD = 0.64$), demonstrating the superiority of its ordering in terms of logic and narrative plausibility. In contrast, Map-only ($M = 2.76$, $SD = 0.91$) and Random Order ($M = 2.88$, $SD = 1.05$) received the lowest mean ratings among all conditions. Random Order exhibited wide confidence intervals, indicating significant divergence across tasks. GPT-only ($M = 3.06$, $SD = 1.03$) fell between these extremes and also showed high divergence in user opinion across tasks.

To further examine order-structure consistency, inspired by Calliope [50], we computed Kendall's τ_b [29] and Spearman's ρ [51] between GeoNarrative and Random orderings. Tasks where Random diverged most from GeoNarrative (e.g., LA-Tour: $\tau_b = -0.022$, $\rho = 0.20$) received the lowest Random scores ($M = 2.31$), whereas tasks with more similar orderings (e.g., BOS-NY: $\tau_b = 0.20$, $\rho = 0.857$) received higher Random scores ($M = 3.47$). This confirms that user ratings are positively associated with order similarity to GeoNarrative's structured sequence.



(a) (b)

Figure 3: Generated from a user textual query. (a) GeoNarrative's query interface, where users specify relationships of interest (e.g., 2024 U.S. election results vs. LGBTQ support); the system processes the request in approximately 30–65 seconds. (b) The generated narrative page, showing multi-view maps and targeted textual explanations that integrate geospatial visualization with narrative text. *Note: the statistics in this example are simulated by a language model from the prompt and do not reflect real-world figures; in deployment such model-generated values are flagged by the provenance markers of Figure 2.*

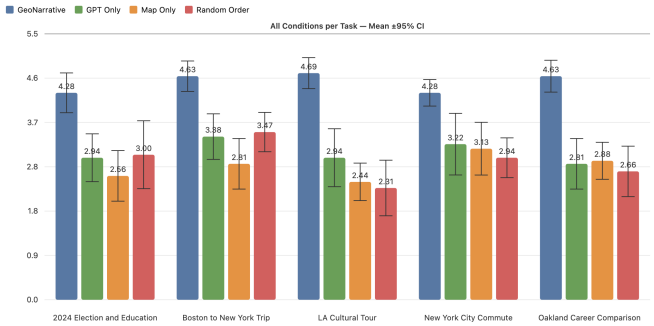


Figure 4: Ratings of information flow across five tasks on a 5-point Likert scale ($N = 16$), where 5 is best. Bars represent mean ratings for four generation methods: GeoNarrative, GPT-only (ChatGPT 4o), Map-only, and Random Order.

Ratings also varied with task complexity. Tasks with more modules (Oakland-Career, Election-Edu, LA-Tour) were rated lowest under Random ($M = 2.31$–3.00) and GPT-only ($M = 2.81$–2.94), while simpler tasks (NYC-Commute, BOS-NY) were less affected. This suggests that structured ordering provides the greatest benefit for complex, multi-module tasks.

6.2 Ablation Study

To disentangle the contributions of each major component, we conducted an ablation study with four conditions. The **Full** condition uses the complete pipeline. **–CLIP** replaces CLIP-based layout retrieval with random template selection (map-count constraints preserved). **–Agent** replaces the LLM Geo Agent with a single-pass LLM prompt without iterative planning or API calls. **–Template** removes template-based rendering entirely, using a single LLM call to generate the HTML page directly. We ran the same five evaluation queries under each condition and analyzed the outputs both quantitatively and qualitatively.

Automated Evaluation. We measured three automated metrics under each condition: CLIP Score (cosine similarity between the query embedding and the selected template embedding), Module Completeness (fraction of the four required module types present in the output, i.e., headline, map, chart, and narrative), and Generation Latency (wall-clock time from query to rendered page). Table 2 summarizes the results.

Table 2: Ablation study results (mean ± SD across five evaluation queries). CLIP Score measures semantic alignment between query and selected template. Module Completeness measures the proportion of required module types present. Latency is wall-clock generation time.

Condition	CLIP Score	Module Compl.	Latency (s)
Full	0.813 ± 0.054	0.95 ± 0.10	45.0 ± 8.5
–CLIP	0.806 ± 0.055	0.75 ± 0.00	69.9 ± 27.0
–Agent	0.813 ± 0.054	0.95 ± 0.10	29.0 ± 6.9
–Template	N/A	1.00 ± 0.00	12.2 ± 0.6

The small CLIP Score difference between Full (0.813) and –CLIP (0.806) does not measure CLIP's true contribution: the score is cosine similarity to the *selected* template, and even a random template passing the structural map-count filter is rarely far from the query in CLIP space, so the metric has a compressed dynamic range. CLIP's real effect appears in **Module Completeness**, which drops from 0.95 to 0.75 as random templates lack required module slots. We therefore read CLIP's contribution from module completeness and layout–content alignment rather than the saturated CLIP Score; the weight sensitivity analysis (Section 3.2) further shows the retrieval behaves stably and non-arbitrarily.

Qualitative Component Analysis. Examining the outputs reveals each component's role. *Removing CLIP retrieval (–CLIP)* preserves content but degrades layout–content alignment (a travel query may receive a multi-map political template with excessive map slots), and module completeness drops from 0.95 to 0.75 as random templates lack required module types. *Removing the LLM Geo Agent (–Agent)* eliminates structured data acquisition: the single-pass LLM produces plausible but generic text lacking API-retrieved data (ACS statistics, Google Directions polylines), and latency falls from ∼45 s to 29 s

but factual specificity suffers. *Removing template-based rendering (–Template)* is fastest (∼12 s) but yields simple linear layouts that lack the magazine-style composition of template-guided pages, even though the LLM can still generate all four module types.

Relationship to the Comparative Experiment. The ablation is consistent with the comparative experiment (Section 6.1): the GPT-only, Map-only, and Random-Order baselines are extreme forms of component removal (dropping the whole pipeline, all non-map modules, and structured sequencing, respectively), and the consistent degradation across both confirms that agent-based scheduling, semantic layout retrieval, and template-driven rendering each contribute meaningfully to overall quality.

6.3 Usability and Visual Performance Study

To evaluate usability and visual quality, we selected representative GeoNarrative pages covering political elections, cultural route planning, LGBTQ issues, and occupation distributions.

The study involved eight expert participants (four with doctoral degrees, all with more than 3.5 years of professional experience in urban studies or visualization). Participants rated ten items on a 5-point Likert scale (1 = *Strongly Disagree*, 5 = *Strongly Agree*) covering content quality (Q1–Q2), text–visual integration (Q3), narrative coherence (Q4), engagement (Q5), efficiency (Q6), and aesthetics (Q7–Q9).

Overall, participants evaluated the pages positively. Content quality was rated strongly (75% agreed the page answered their question, 88% found sufficient data), text–visual integration highest (63% strongly agreeing), and aesthetic items well (layout clarity, color comfort, and information accessibility all $\geq$63% positive); the main concern was generation latency (Q6), where 50% were neutral and 25% found it too long. Full per-item distributions are in the supplemental materials.

6.4 Qualitative Feedback Analysis

We analyzed qualitative feedback using reflexive thematic analysis [4]: three researchers independently coded transcripts and survey responses, then grouped codes into themes. Four emerged. **Data credibility:** generated content sometimes lacked context, and participants asked for explicit source references and an option to *view original data*: one journalist said *"this could make my work much easier"* but stressed the need for data-validity mechanisms. **Narrative structure:** the structure was clear, but participants wanted stronger narrative guidance and text–graphic coordination. **Interactive control:** users appreciated map comparison but requested module-level editing. **Application preferences:** interest centered on urban travel, policy analysis, and tourism planning, suggesting contextual factors (e.g., weather, transport modes) and personalized route-based stories.

7 Discussion

7.1 Design Implications for Automated Narrative Visualization

Our evaluation suggests several design implications. First, **modular decomposition is essential**: treating a page as a composition of typed modules (headlines, maps, charts, text) lets the system reason about sequencing structurally rather than relying on the language model's implicit sense of order, extending data-driven storytelling [10, 50] to the geospatial domain with its added ordering constraints (geographic proximity, route continuity). Second, **layout retrieval connects content and form**: CLIP-based retrieval aligns module arrangement with professionally designed pages, and without it (the –CLIP condition) module completeness drops from 0.95 to 0.75. Third, the benefit is **complexity-dependent**: structured ordering helps most for tasks with many modules or heterogeneous data, while GPT-only ordering suffices for simpler ones.

7.2 Trust and Transparency in AI-Generated Visualization

Participant feedback consistently highlighted concerns about data provenance: when the LLM Geo Agent cannot retrieve real-time data through APIs, it may generate plausible but unverified values, which creates a risk of misinformation. To address this, we implemented a **source-tracing module** that annotates each data-bearing module with provenance information (API source, dataset name, retrieval timestamp); our evaluation did not, however, directly test whether users noticed or used these annotations, which future controlled experiments should examine. More broadly, trust in narrative visualization [34] requires attention to both *data accuracy* (whether the numbers are correct) and *narrative framing* (whether the ordering or emphasis introduces bias).

7.3 Human–AI Collaboration and Geographic Challenges

Our qualitative findings reveal a tension between automation and user control. Participants valued the system's ability to produce publication-ready pages with little effort, but many wanted fine-grained editing capabilities such as module reordering, inline text editing, and constrained regeneration. This suggests that the optimal paradigm is **mixed-initiative collaboration** [24], in which the system provides a strong initial draft and the user refines it.

Geographic narrative visualization also introduces unique challenges. **Spatial coherence** requires managing two-dimensional relationships: a story comparing cities must organize by theme or by location, and our algorithm prioritizes thematic coherence; **scale sensitivity** demands appropriate zoom and aggregation from neighborhoods to nations; and **cultural variation** in place names and boundaries may limit generalization beyond U.S. contexts, since our implementation relies on U.S.-centric sources such as the ACS, though the data-retrieval stage is source-agnostic, so other national portals can be integrated without changing the pipeline. Adaptive ordering, systematic scale evaluation, and multilingual support are promising future directions.

8 Limitations and Future Work

Sample size. Our experiments ($N = 16$ and $N = 8$) follow HCI practice [5] but limit generalizability; future studies should recruit larger, more diverse samples. **Image and color.** Image modules are in our taxonomy and pipeline but disabled by default and omitted from the evaluation, owing to the latency and reliability of current image models; the system also uses fixed template colors, where generative methods such as GenColor [25] could improve quality and memorability [2]. **Template flexibility.** Users can already register templates by annotating new pages with the same attribute schema and adjust the fusion weights (Section 3.2), but authoring fully custom layouts from scratch remains future work. **Performance and interaction.** Generation latency (30–65 s) remains a barrier (driven largely by real-data API retrieval, as the –Agent ablation lowers it to ∼29 s), and future work will optimize the pipeline through parallel calls and cached embeddings and add module-level editing.

9 Conclusion

We presented GeoNarrative, a system that transforms natural-language geospatial queries into structured, multi-module narrative visualization pages through an end-to-end pipeline integrating an LLM-based planning agent, CLIP-based layout retrieval, and template-driven rendering, and we release an annotated dataset of 5,321 modules across 199 real-world pages labeled with functional role and module type. In a controlled experiment ($N = 16$) and an expert usability study ($N = 8$), GeoNarrative's structured information flow was rated higher than GPT-only, Map-only, and Random baselines in logic and narrative plausibility, with large effect sizes; given the modest sample sizes we read these as encouraging rather than conclusive. By lowering the barrier to rich geospatial narratives, GeoNarrative opens new possibilities for journalists, urban planners, and educators across geographic information science and narrative visualization.

10 Acknowledgments

We thank Nan Cao (Tongji University) for invaluable guidance on this work, Ye Zhang for technical advice, and Daniel Tsai for helpful discussions.

In accordance with IEEE policy, we disclose our use of AI-generated content. The figures that depict generated narrative pages (Figures 2 and 3 and the case-study pages) are produced by GeoNarrative's own

pipeline, which uses GPT-4o for module text and chart series, a local Llama-3-8B model for intent routing, and CLIP ViT-B/32 for layout retrieval, exactly as described in Sections 3.4 and 3.3.3; these AI-generated values are illustrative and, where simulated, are flagged as such. For the writing of this manuscript, the authors used AI-based tools (large language models) only for grammar checking and language editing; all ideas, system design, analyses, and conclusions are the authors' own.

11 Supplemental Materials

To support reproducibility and the claims made above, we provide the following supplemental materials. **(S1) Demonstration video** (<5 min): an end-to-end screen capture from a natural-language query to a generated multi-module page, including the always-visible provenance markers and the show/hide toggle. **(S2) Supplementary figures**: the full route-planning case-study page (Section 5), the per-item Likert distributions of the expert usability study (Section 6.3), the full provenance page with source markers on every module (Figure 2), and a gallery of additional generated pages spanning city-, route-, and country-scale topics. The annotated dataset of 5,321 modules across 199 pages (module-level annotations, schema, annotation guidebook, and the triple-annotated 500-module subset with the script used to compute Fleiss' κ) will be released on OSF upon publication.

References

[1] R. C. Basole and T. Major. Generative ai for visualization: Opportunities and challenges. *IEEE Computer Graphics and Applications*, 44(2):55–64, 2024. doi: 10.1109/mcg.2024.3362168 2

[2] M. A. Borkin, A. A. Vo, Z. Bylinskii, P. Isola, S. Sunkavalli, A. Oliva et al. What makes a visualization memorable? *IEEE transactions on visualization and computer graphics*, 19(12):2306–2315, 2013. doi: 10.1109/tvcg.2013.234 8

[3] M. Bostock, V. Ogievetsky, and J. Heer. D^3 data-driven documents. *IEEE transactions on visualization and computer graphics*, 17(12):2301–2309, 2011. doi: 10.1109/TVCG.2011.185 4

[4] V. Braun and V. Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006. doi: 10.1191/1478088706qp063oa 8

[5] K. Caine. Local standards for sample size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 981–992, 2016. doi: 10.1145/2858036.2858498 8

[6] A. Cairo. *The Functional Art: An introduction to information graphics and visualization*. New Riders, Berkeley, 2012. 5

[7] S. Caquard and W. Cartwright. Narrative cartography: From mapping stories to the narrative of maps and mapping. *The Cartographic Journal*, 51(2):101–106, 2014. doi: 10.1179/0008704114Z.000000000130 2

[8] X. Chen, S. Chen, T. Xu, B. Yin, J. Peng, X. Mei et al. Smapgan: Generative adversarial network-based semisupervised styled map tile generation method. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5):4388–4406, 2021. doi: 10.1109/tgrs.2020.3021819 2

[9] Y. Chen, R. Li, A. Mac, T. Xie, T. Yu, and E. Wu. Nl2interface: Interactive visualization interface generation from natural language queries. *arXiv preprint arXiv:2209.08834*, 2022. doi: 10.48550/arXiv.2209.08834 2

[10] Z. Chen, 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 8

[11] L. Cheng, D. Deng, X. Xie, R. Chen, and Y. Wu. SNIL: Generating sports news from insights with large language models. *IEEE Transactions on Visualization and Computer Graphics*, 31(7):3973–3986, 2025. doi: 10.1109/tvcg.2024.3392683 2

[12] W. Cui, X. Zhang, Y. Wang, H. Huang, B. Chen, L. Fang et al. Text-to-viz: Automatic generation of infographics from proportion-related natural language statements. *IEEE transactions on visualization and computer graphics*, 26(1):906–916, 2020. doi: 10.1109/tvcg.2019.2934785 2

[13] C. Deng, T. Zhang, Z. He, Q. Chen, Y. Shi, Y. Xu et al. K2: A foundation language model for geoscience knowledge understanding and utilization. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pp. 161–170, 2024. doi: 10.1145/3616855.3635772 1

[14] V. Dibia. Lida: A tool for automatic generation of grammar-agnostic visualizations and infographics using large language models. *arXiv preprint arXiv:2303.02927*, 2023. doi: 10.18653/v1/2023.acl-demo.11 2

[15] V. Dibia and Ç. 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

[16] ESRI. ArcGIS StoryMaps. https://storymaps.arcgis.com/, 2024. Accessed: 2025-03-01. 2

[17] W. Feng, W. Zhu, T.-j. Fu, V. Jampani, A. Akula, X. He et al. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems*, 36:18225–18250, 2023. doi: 10.52202/075280-0802 2

[18] S. Few. *Information dashboard design: The effective visual communication of data*. O'Reilly Media, Inc., Sebastopol, 2006. 2

[19] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971. doi: 10.1037/h0031619 5

[20] M. Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701, 1937. doi: 10.1080/01621459.1937.10503522 6

[21] S. Ganguli, P. Garzon, and N. Glaser. Geogan: A conditional gan with reconstruction and style loss to generate standard layer of maps from satellite images. *arXiv preprint arXiv:1902.05611*, 2019. doi: 10.48550/arXiv.1902.05611 2

[22] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair et al. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. doi: 10.48550/arXiv.1406.2661 2

[23] Google Research. Geospatial reasoning: Unlocking insights with generative AI and multiple foundation models. https://research.google/blog/geospatial-reasoning-unlocking-insights-with-generative-ai-and-multiple-foundation-models/, 2025. Accessed: 2025-08-25. 2

[24] E. Horvitz. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 159–166, 1999. doi: 10.1145/302979.303030 8

[25] Y. Hou, X. Zeng, Y. Wang, M. Yang, X. Chen, and W. Zeng. Gencolor: Generative color-concept association in visual design. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2025. doi: 10.1145/3706598.3713418 8

[26] K. Hu, M. A. Bakker, S. Li, T. Kraska, and C. Hidalgo. Vizml: A machine learning approach to visualization recommendation. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–12, 2019. doi: 10.1145/3290605.3300358 2

[27] J. Hullman and N. Diakopoulos. Visualization rhetoric: Framing effects in narrative visualization. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2231–2240, 2011. doi: 10.1109/TVCG.2011.255 2

[28] Y. Jiang, Q. Chao, Y. Chen, X. Li, S. Liu, and G. Cong. Urbanllm: Autonomous urban activity planning and management with large language models. *arXiv preprint arXiv:2406.12360*, 2024. doi: 10.18653/v1/2024.findings-emnlp.98 1

[29] M. G. Kendall. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251, 1945. doi: 10.1093/biomet/33.3.239 6

[30] Knight Lab. StoryMapJS. https://storymap.knightlab.com/, 2024. Accessed: 2025-03-01. 2

[31] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977. doi: 10.2307/2529310 5

[32] S. Latif, S. Chen, and F. Beck. A deeper understanding of visualization-text interplay in geographic data-driven stories. *Computer Graphics Forum*, 40(3):311–322, 2021. doi: 10.1111/cgf.14309 2

[33] G. Li, X. Wang, G. Aodeng, S. Zheng, Y. Zhang, C. Ou et al. Visualization generation with large language models: An evaluation. *arXiv preprint arXiv:2401.11255*, 2024. doi: 10.48550/arXiv.2401.11255 2

[34] Q. V. Liao and J. W. Vaughan. AI transparency in the age of LLMs: A human-centered research roadmap. *Harvard Data Science Review*, Special Issue 5, 2024. doi: 10.1162/99608f92.8036d03b 8

[35] C. Liu, L. Xie, Y. Han, D. Wei, and X. Yuan. Autocaption: An approach to generate natural language description from visualization automatically. In *2020 IEEE Pacific visualization symposium (PacificVis)*, pp. 191–195. IEEE, 2020. doi: 10.1109/pacificvis48177.2020.1043 2

[36] P. Ma, R. Ding, S. Wang, S. Han, and D. Zhang. Demonstration of insightpilot: An llm-empowered automated data exploration system. *arXiv preprint arXiv:2304.00477*, 2023. doi: 10.18653/v1/2023.emnlp-demo.31 2

[37] R. Ma, H. Mei, H. Guan, W. Huang, F. Zhang, C. Xin et al. Ladv: Deep learning assisted authoring of dashboard visualizations from images and sketches. *IEEE Transactions on Visualization and Computer Graphics*, 27(9):3717–3732, 2021. doi: 10.1109/tvcg.2020.2980227 2

[38] P. Maddigan and T. Susnjak. Chat2vis: Generating data visualizations via natural language using chatgpt, codex and gpt-3 large language models. *Ieee Access*, 11:45181–45193, 2023. doi: 10.1109/access.2023.3274199 2

[39] A. Mansourian and R. Oucheikh. Chatgeoai: Enabling geospatial analysis for public through natural language, with large language models. *ISPRS International Journal of Geo-Information*, 13(10):348, 2024. doi: 10.3390/ijgi13100348 2

[40] R. Manvi, S. Khanna, G. Mai, M. Burke, D. Lobell, and S. Ermon. Geollm: Extracting geospatial knowledge from large language models. *arXiv preprint arXiv:2310.06213*, 2023. doi: 10.48550/arXiv.2310.06213 1, 2

[41] Mapbox. Mapbox – maps, navigation, and search for everywhere. https://www.mapbox.com/, 2025. Accessed: 2025-08-25. 4

[42] Microsoft Corporation. Power bi. https://www.microsoft.com/en-us/power-platform/products/power-bi, 2025. Accessed: 2025-08-18. 2

[43] D. Moritz, C. Wang, G. L. Nelson, H. Lin, A. M. Smith, B. Howe et al. 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

[44] 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

[45] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021. doi: 10.48550/arXiv.2103.00020 3

[46] N. H. Riche, C. Hurter, N. Diakopoulos, and S. Carpendale. *Data-driven storytelling*. CRC Press, 2018. doi: 10.1201/9781315281575 5

[47] R. E. Roth. Cartographic design as visual storytelling: Synthesis and review of map-based narratives, genres, and tropes. *The Cartographic Journal*, 58(1):83–114, 2021. doi: 10.1080/00087041.2019.1633103 2

[48] 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

[49] E. Segel and J. Heer. Narrative visualization: Telling stories with data. *IEEE transactions on visualization and computer graphics*, 16(6):1139–1148, 2010. doi: 10.1109/TVCG.2010.179 2, 3, 4, 5

[50] D. Shi, X. Xu, F. Sun, Y. Shi, and N. Cao. Calliope: Automatic visual data story generation from a spreadsheet. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):453–463, 2021. doi: 10.1109/TVCG.2020.3030403 2, 6, 8

[51] C. Spearman. The proof and measurement of association between two things. 1961. doi: 10.1037/11491-005 6

[52] C. D. Stolper, B. Lee, N. H. Riche, and J. Stasko. Emerging and recurring data-driven storytelling techniques: Analysis of a curated collection of recent stories. In *Data-Driven Storytelling*, pp. 85–105. A K Peters/CRC Press, 2018. doi: 10.1201/9781315281575-4 2, 3

[53] M. Sun, L. Cai, W. Cui, Y. Wu, Y. Shi, and N. Cao. Erato: Cooperative data story editing via fact interpolation. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):983–993, 2023. doi: 10.1109/TVCG.2022.3209428 2

[54] A. Swerdlow, R. Xu, and B. Zhou. Street-view image generation from a bird's-eye view layout. *IEEE Robotics and Automation Letters*, 9(4):3578–3585, 2024. doi: 10.1109/lra.2024.3368234 2

[55] Tableau Software. What is a dashboard? https://www.tableau.com/dashboard/what-is-dashboard, 2025. Accessed: 2025-08-18. 2

[56] Y. Tian, W. Cui, D. Deng, X. Yi, Y. Yang, H. Zhang et al. ChartGPT: Leveraging LLMs to generate charts from abstract natural language. *IEEE Transactions on Visualization and Computer Graphics*, 31(3):1731–1745, 2025. doi: 10.1109/tvcg.2024.3368621 2

[57] M. Vartak, S. Madden, A. Parameswaran, and N. Polyzotis. Seedb: automatically generating query visualizations. 2014. doi: 10.14778/2733004.2733035 2

[58] L. Wang, S. Zhang, Y. Wang, E.-P. Lim, and Y. Wang. Llm4vis: Explainable visualization recommendation using chatgpt. *arXiv preprint arXiv:2310.07652*, 2023. doi: 10.18653/v1/2023.emnlp-industry.64 2

[59] Y. Wang, Z. Sun, H. Zhang, W. Cui, K. Xu, X. Ma et al. Datashot: Automatic generation of fact sheets from tabular data. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):895–905, 2020. doi: 10.1109/TVCG.2019.2934398 2

[60] 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*, pp. 1–6, 2016. doi: 10.1145/2939502.2939506 2

[61] A. N. Wu and F. Biljecki. Ganmapper: geographical data translation. *International Journal of Geographical Information Science*, 36(7):1394–1422, 2022. doi: 10.1080/13658816.2022.2041643 2

[62] Y. Zhang, Z. He, J. Li, J. Lin, Q. Guan, and W. Yu. Mapgpt: an autonomous framework for mapping by integrating large language model and cartographic tools. *Cartography and Geographic Information Science*, 51(6):717–743, 2024. doi: 10.1080/15230406.2024.2404868 2

[63] G. Zheng, X. Zhou, X. Li, Z. Qi, Y. Shan, and X. Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22490–22499, 2023. doi: 10.1109/cvpr52729.2023.02154 2

[64] M. Zhou, Q. Li, X. He, Y. Li, Y. Liu, W. Ji et al. Table2charts: recommending charts by learning shared table representations. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 2389–2399, 2021. doi: 10.1145/3447548.3467279 2