



An Efficient System for Automatic Map Storytelling: A Case Study on Historical Maps

Ziyi Liu ^{1,*}, Claudio Affolter ^{1,*}, Sidi Wu^{1,†}, Yizi Chen¹, and Lorenz Hurni¹

¹Institute of Cartography and Geoinformation, ETH Zurich, Zurich, Switzerland

*These authors contributed equally to this work.

Correspondence: †Sidi Wu (sidiwu@ethz.ch)

Abstract. Historical maps provide valuable information and knowledge about the past. However, as they often feature non-standard projections, hand-drawn styles, and artistic elements, it is challenging for non-experts to identify and interpret them. While existing image captioning methods have achieved remarkable success on natural images, their performance on maps is suboptimal as maps are underrepresented in their pre-training process. Despite the recent advance of vision-enabled GPT models in text recognition and map captioning, they still have a limited understanding of maps, as their performance wanes when texts (e.g., titles and legends) in maps are missing or inaccurate. Besides, it is inefficient or even impractical to fine-tune these models with users' own datasets. To address these problems, we propose a novel and lightweight map-captioning counterpart. Specifically, we fine-tune the state-of-the-art vision-language model CLIP to generate captions relevant to historical maps and enrich the captions with GPT models to tell a brief story regarding *where*, *what*, *when* and *why* of a given map. We propose a novel decision tree architecture to only generate captions relevant to the specified map type. Our system shows invariance to text alterations in maps. The system can be easily adapted and extended to other map types and scaled to a larger map captioning system.

Submission Type. algorithm.

BoK Concepts. image processing and analysis → image understanding → visual interpretation

Keywords. image captioning, GPT, historical maps, map storytelling

1 Introduction

Historical maps allow us to learn more about a certain place's geography, economics, history, and culture. However, unlike modern maps, they often contain less

accurate geographic information, varying artistic or religious symbols and legends, non-standard projections, and hand-drawn styles. This challenges non-experts (i.e., non-cartographers) to correctly identify and capture the key information. Image captioning (Anderson et al., 2018; Chen and Zitnick, 2014; Stefanini et al., 2022; Zhou et al., 2020) provides descriptions for images in natural language and serves as a powerful tool in various situations, such as content understanding for individuals with visual impairments, image tagging for database management, and efficient search and retrieval of images. Typically, an image encoder is trained for visual cues, and a textual decoder is used to produce the final caption. CLIP (Contrastive Language-Image Pre-Training), recently proposed by Radford et al. (2021), learns the shared representations for images and text prompts. It was trained over a tremendous number of images for a good correlation between images and texts and has been widely used for downstream tasks (like image captioning) with little or no further training. For example, the ClipCap model (Mokady et al., 2021) uses the pre-trained CLIP prefix and fine-tunes a language model to generate image captions, which has achieved state-of-the-art performance. However, most image captioning methods generate descriptions limited to visual elements, which are not sufficient to tell a meaningful story about maps.

In this paper, we propose a map-specific captioning system equipped with a basic understanding of maps, which is not yet addressed by any image-captioning models. By fine-tuning CLIP models for map-relevant captions and using GPT (Generative Pre-trained Transformer) to combine and enrich them, our system could generate a comprehensive story. We choose a range of recent GPT models, including GPT-3.5-turbo, GPT-4o, and GPT-4o-mini for this task. Compared to GPT-4-turbo and o1, these models have comparable performances at a much lower cost. Given an input map, the story should answer the following questions:

- *Where* does the map depict about?

- *What* is the map type, style, and topic?
- *When* was the map created?
- *Why* was the map created?

We focus on two major map types: topographic maps, which provide detailed and accurate graphical representations of an area (Kent, 2009), and pictorial maps, which use illustrations to represent information (Schnürer et al., 2021). Since not every aspect is relevant for both map types — for example, the topic is usually the same for all the topographic maps (i.e., geography and elevation) — we propose a decision tree structure to generate the captions with respect to the map type. Moreover, we design a user interface for interactive map storytelling, where the user can choose which aspects to include in the story.

2 Related Work

Map information retrieval. Various studies have been carried out on automatically retrieving information from maps. In Zhou et al. (2018), state-of-the-art deep convolutional neural networks (CNNs) were used for automatic map-type classification. In Schnürer et al. (2021), the authors used CNNs to identify pictorial maps and further recognize objects on pictorial maps. Besides map types, the study Li and Xiao (2023) also recognized geographic regions and projections. The work Hu et al. (2022) used GIS-based augmentation to bootstrap the recognition of map extents and state names. In the paper Touya et al. (2020), the authors used CNNs to infer the object classes existing in the map and the spatial extent from French geography textbooks.

Image captioning. Image captioning can be categorized into template-based, retrieval-based, and novel caption generation approaches (Hossain et al., 2019). The template-based approach detects elements such as objects, actions, and scenes in images and fits those elements to a pre-defined template for generating a grammatically correct descriptive caption (Farhadi et al., 2010; Li et al., 2011; Zeng et al., 2018). However, there are limitations to such a caption-generation system when it comes to generating captions from diverse scenes or scenes that fall outside the scope of the provided templates. To address this issue, retrieval-based methods utilize image similarity to annotate images without captions by comparing them to known captions (Hodosh et al., 2013; Ordonez et al., 2011; Sun et al., 2015). These methods are incapable of generating diverse captions and producing image-specific and distinctive captions. Other works focused on enhancing the ability of image and language understanding with state-of-the-art deep learning models (Kiros et al., 2014; Xu et al., 2015; Yao et al., 2017; You et al., 2016). They achieve more diverse and image-specific results than template- and retrieval-based methods. However, the models for image and language understanding are trained separately, leading to a lack of integration between the two

feature sets. To better study the correlation between images and captions, CLIP Radford et al. (2021) was proposed to jointly train the image encoder and text encoder to predict the most relevant text snippet as the label for an image. ClipCap (Mokady et al., 2021) combines the CLIP encoder and GPT models for more detailed and comprehensive captions with semantic understanding. It leverages the pre-trained CLIP and GPT and trains a lightweight transformer-based mapping network in between.

GPT models. GPT models (Brown et al., 2020), such as GPT-3.5 (ChatGPT), GPT-4, and GPT-4o, developed by OpenAI, are large-scale language models (LLMs) based on transformers, primarily designed for tasks like text generation. Given instructions (called *prompts*), the GPT models can generate human-like texts in natural language and conversationally answer questions. They can be used for answering questions, searching, text summarization, and content generation. In this paper, we use GPTs to tell stories about maps based on the input keywords and prompts.

We compare the performances of GPT-3.5-turbo, GPT-4o, and GPT-4o-mini. GPT-3.5-turbo is optimized for quick interactions such as chat applications and real-time language processing tasks. It suits applications requiring rapid response times with high language processing quality. GPT-4o and GPT-4o-mini are designed to efficiently manage complex multi-modal tasks. GPT-4o maintains the equivalent performance as GPT-4-turbo on English text while offering significant advancements in processing other languages, as well as in vision and audio understanding, over previous models. Its API has faster operations and lower cost than GPT-4-turbo. GPT-4o-mini, a compact model, is designed for cost-effective performance in resource-constrained environments. It outperforms GPT-3.5-turbo across various academic benchmarks, including those measuring textual intelligence. We also compare our proposed method against vision-enabled GPT models, including GPT-4-turbo, GPT-4o, and GPT-4o-mini — the more advanced LLMs that can process image inputs and generate captions directly. While the more recent GPT-4o and GPT-4o-mini models are recognized for their impressive multimodal capabilities, GPT-4-turbo remains highly effective, particularly in reasoning tasks, and is among the first GPT models to support image processing.

3 Methodology

An overview of our methods is presented in Figure 1. We first process maps and their metadata automatically from the online map repository to generate a training dataset with keyword captions regarding *where*, *what* and *when* and use this dataset to fine-tune different CLIP models. In the inference phase, we propose a decision tree architecture to structure the keyword captions with respect to the map type and use GPT to extend the context (*why*) and summarize the story. Furthermore, a web interface is

developed for interactive storytelling with the decision tree architecture and fine-tuned models loaded at the backend.

3.1 Preliminary: CLIP

CLIP (Radford et al., 2021) is a neural network designed for learning joint representations of images and texts in a way that enables efficient cross-modal understanding. As shown in Figure 2, it employs a vision transformer (ViT) for image processing and a transformer-based language model to process text. The training strategy of CLIP is based on a contrastive learning framework where the model is trained to maximize agreement between representations of positive pairs (correct image-text pairs) and minimize agreement between representations of negative pairs (mismatched image-text pairs). CLIP was pre-trained on a dataset of 400 million images and associated natural language descriptions and can be applied to various cross-modal tasks, such as image classification and object detection, without requiring task-specific adaptations. One notable feature of CLIP is its ability to generalize to zero-shot scenarios. The model can make predictions on classes that were not seen during training: as long as the names/descriptions of the classes from the target dataset are specified, a single linear classifier is applied to predict the class with the highest probability.

3.2 Dataset preparation

We collected data from the David Rumsey Historical Map Collection¹, an online map repository containing historical maps from all over the world complemented with detailed metadata. As we focus on topographic maps and pictorial maps, only the maps in the collection's categories *Classical* and *Pictorial map* were considered. In total, after manually filtering out poor-quality maps, 1,334 topographical and 3,183 pictorial maps were gathered. To create ground-truth captions answering the four questions introduced in Section 1, we extracted necessary information from the metadata associated with each map. We processed topographic maps and pictorial maps separately as different challenges occurred.

Where. For topographic maps, since the location attribute in the metadata is often ambiguous, incorrect, and imprecise, we also parsed the location information from map titles. For pictorial maps, a substantial class imbalance emerged, with 3,183 maps depicting 1,349 different locations. Consequently, we decided to only focus on the two largest classes – the world and the United States.

What. For topographic maps, there are a few style variations, such as *with/without relief*, *with/without decorative elements*, and *hand-colored/engraved*, often described in metadata. However, as this description is not well-structured and consistent, we have only extracted keywords from these descriptions. We calculated the frequen-

cies of each keyword and then reduced the number of style classes to focus only on the most frequent ones. As topographic maps mainly describe the geography and topography of an area, we omitted the map topic in the caption. Pictorial maps are less constrained in styles with diverse color schemes and artistic illustrations, making it challenging to summarize the style. Thus, we excluded styles when captioning pictorial maps. Similar to *where*, the topics of pictorial maps present a strong imbalance. For example, there are 29% flight network maps but only 2% military maps. We decided to focus only on the most frequent topics and manually merged some sub-categories into a more general class.

When: We derived the century of production from the *Date* attribute in the metadata, which is consistently complete. However, as most pictorial maps were created in the 20th century, it was no longer necessary to depict when they were created in the caption.

Why: The metadata provides no information about the purpose and functionality of a map. To fill this gap, we made use of GPT's generative capabilities. Instead of then using the generated caption as ground truth to fine-tune the model, which would take additional training effort and might lead to error propagation stemming from imperfect captions, we only made use of GPT in the inference step.

Eventually, we obtained separate datasets for each caption category. Each dataset contains maps (compressed, up to 768×768 pixels) and the corresponding captions. Note that the ground-truth captions are only comprised of keywords (or phrases) like *Italy* or *hand colored with pictorial relief* instead of a full sentence. Table 1 gives an overview of the final number of classes and maps for each of the six caption categories.

Table 1. Overview of the generated datasets for each caption category. Both the numbers of classes and map samples are shown. We differentiate the *location* for topographic maps and pictorial maps.

	Caption category	# of classes	# of maps
	Map type	2	4'517
Topographic	Location (topographic)	27	723
	Style	6	1'132
	Century	4	1'334
Pictorial	Location (pictorial)	2	290
	Topic	13	284

3.3 Fine-tuning CLIP

The visual information from maps is captured and transformed into textual information using CLIP models, each fine-tuned to generate keyword captions for a specific aspect. We utilize six CLIP models in total, generating keyword captions related to location, map type, topic, style, and century, as shown in Figure 1 A). The fine-tuning pro-

¹<https://www.davidrumsey.com/>

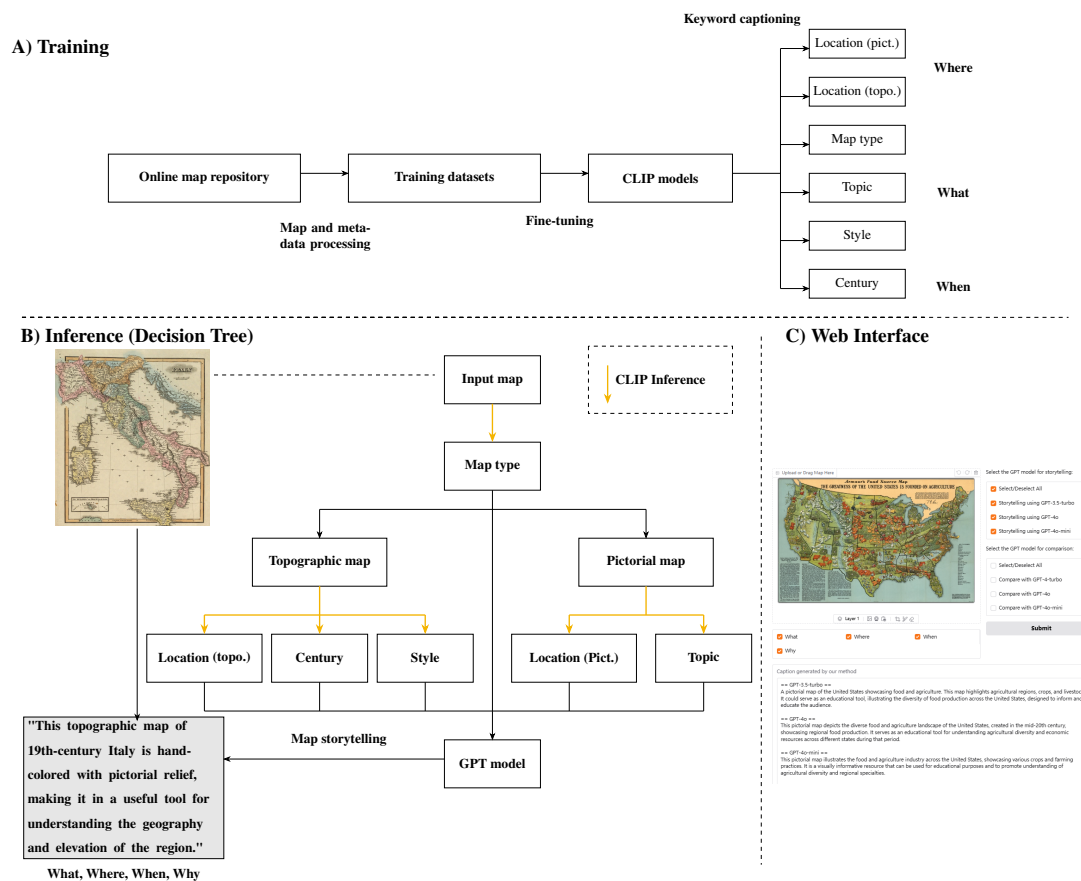


Figure 1. Overview of our proposed methodology.

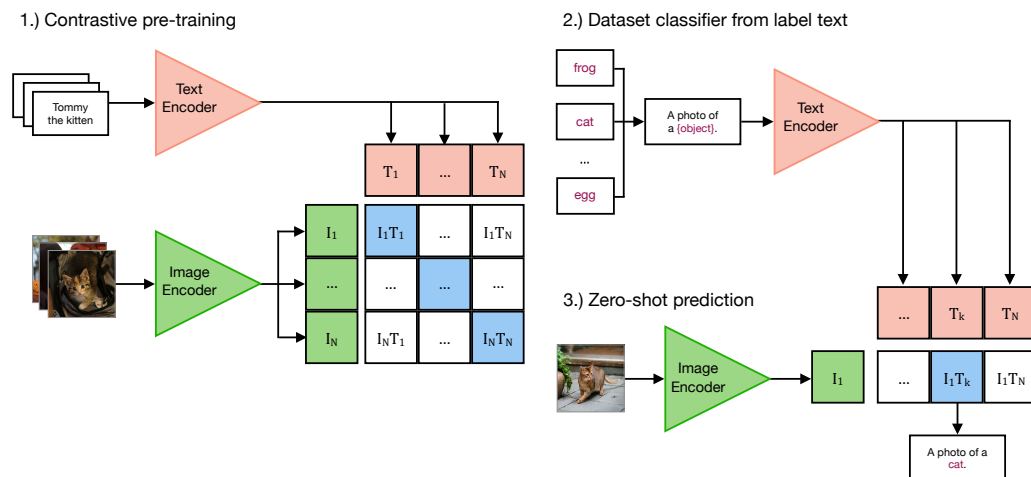


Figure 2. CLIP model architecture: an image encoder and a text encoder are jointly trained to predict the correct pairings of a batch (image, text) of training examples. During inference, the trained text encoder creates a zero-shot linear classifier by embedding the names or descriptions of the classes from the target dataset.

cess was adapted from Radford et al. (2021). We used a batch size of 10 and an initial learning rate of $1e-5$ with Adam optimizer Kingma and Ba (2014). All models were trained on a single 16 GB NVIDIA RTX A4000 GPU.

3.4 Decision tree for inference

As some aspects are only relevant to certain map types, we proposed a decision tree structure where our models first predict the map type at the root node and then predict other relevant keyword captions based on the identified map type. The inference process using the decision tree

is illustrated in Figure 1 B. For instance, given the map on the left in Table 4, the decision tree classifies it as a “pictorial map” (keyword 1), leading to the prediction of only the location “world” (keyword 2) and the topic “flight network” (keyword 3), while the style keyword is excluded as it is irrelevant in this context (see Section 3.2). At last, we use GPT to extend the story about *why* based on the generated keyword captions and to summarize the story by answering the questions in Section 1, with the prompt of the following structure: “Please create a concise sentence that encapsulates these keywords: {keywords}. Please also address the following aspects in a concise and coherent paragraph, in under 40 words, about: {questions}. Ensure the output is a single paragraph and must strictly no longer than 50 words. Do not include any generated information or fabricated details.”

3.5 Web Interface

An interactive web application has been developed for our map captioning system, where users can upload maps for caption generation. Users can select specific questions that they are interested in, and the application will generate captions with relevant information to address the selected questions. In addition, users can choose different GPT models as story generators, as well as various vision-enabled GPT models for captioning comparison. A screenshot of our web interface is shown in Figure 1 C. The core functionality of this application is built using Gradio², an open-source Python package designed for building web applications efficiently. To further enhance user experience, the Gradio application is integrated into a webpage that offers detailed descriptions and map examples.

4 Results

4.1 Fine-tuned CLIP Models



We compare the prediction accuracy of our fine-tuned CLIP models with the base CLIP model for each caption category. The base CLIP model can predict never-seen classes as long as the enumeration of class names is given. The similarity between the text encoding (class name) and the image encoding is then used to predict the most probable class. As shown in Table 2, based on 113 test maps (68 topographic maps and 45 pictorial maps), our fine-tuned CLIP models significantly outperform the base model in five out of six caption categories. The base model performed slightly better in the *location (pictorial)* caption category, likely due to its extensive training on illustrations of the United States and the world with significant graphic variations. In Table 3, two examples of keyword captions generated by the base CLIP model and our fine-tuned CLIP models are shown respectively.

²<https://www.gradio.app/>

Table 2. Comparison of prediction accuracies achieved per caption category with the base CLIP model and our fine-tuned CLIP models.

Caption category	Base CLIP	Fine-tuned CLIP
Map type	0.43	0.96
Location (topo.)	0.28	0.78
Style	0.29	0.75
Century	0.40	0.76
Location (pict.)	0.96	0.93
Topic	0.47	0.67
Average Accuracy	0.47	0.81

Table 3. Comparison of keyword captions generated by the base CLIP model and our fine-tuned CLIP models for the two test maps depicted in the left column. Falsely predicted caption is marked in red.

Test map	Caption category	Base CLIP	Fine-tuned
	Map type	pictorial map	pictorial map
	Location (pict.)	world	world
	Topic	world war 2	flight network
	Map type	pictorial map	topographic map
	Location (topo.)	eastern hemisphere	asia
	Style	hand colored with decorative elements and pictorial relief	hand colored
	Century	18th century	19th century

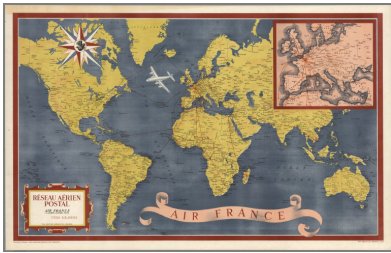
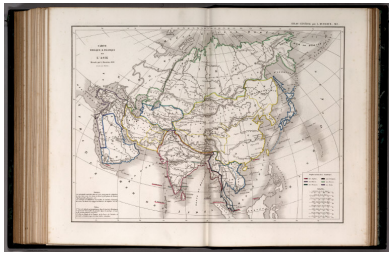
4.2 Map captioning

We compared our map captioning system with ClipCap model. Additionally, to assess the efficiency and stability of our system, we compared the performance of ours using different story generators, including GPT-3.5-turbo, GPT-4o, and GPT-4o-mini, and the recent vision-enabled GPT models, including GPT-4-turbo, GPT-4o, and GPT-4o-mini.

Table 4 shows examples of the stories generated by our method and ClipCap. While the original ClipCap can recognize maps, there are wrong interpretations like “room” and “map cutter”. To fine-tune it, we combined *topic* and *location* for pictorial maps and *century* and *location* for topographic maps in a single sentence. In Table 4, while the fine-tuned ClipCap correctly detects the Air France global flight network in the first example, it falsely recognizes the production time (should be 19th instead of 17th century) in the second example. By comparison, our method can generate more accurate, comprehensive, and detailed captions, including *what*, *when*, *where*, and *why*.

In Tables 5 to 7 we compare the generated captions of our system with vision-enabled GPT models under the influence of missing or wrong map texts. We evaluate our model’s performance across various versions of GPT, as

Table 4. The stories of the same test maps generated by our method and ClipCap (Mokady et al., 2021). *: Fine-tuned ClipCap.

Test map		
Ours	This pictorial map illustrates the global flight network, showcasing worldwide destinations and travel routes. It is a visual representation of the world, providing information about flight connections, and can be used for planning and visualizing travel itineraries.	This hand-colored topographic map of Europe in the 19th century features pictorial relief. It shows the geographical features of Europe and can be used for geographical analysis.
ClipCap	A map of the world is on display in a room.	An old map with a map cutter on it.
ClipCap*	Map depicting Air France worldwide flight network.	Map depicting Europe in the 17th century.

our model can be seamlessly integrated with each for storytelling. We can see that the vision-enabled GPT models present superior capability in recognizing texts in maps (such as the map title “Gallia Vetus” in Table 5 and the year “1942” in Table 7) and enriching the contextual information with external knowledge. However, hallucinations like the map title “Baird North’s War Map” in Table 7 also occur. When we directly alter or modify texts on maps to simulate the scenarios where maps have missing or incorrect textual information, the GPTs struggle to identify the correct information from maps. For example, in Table 5, the GPT models failed to predict the depicted geography (mainly France). In Table 6 and Table 7, they were unable to predict the relevant century (19th century) and map topic (World War 2), respectively. In the three examples, the vision-enabled GPT models were generally ineffective when map texts were missing, incomplete, or incorrect. In contrast, our method identifies accurate information from maps without relying on textual data, showing robustness against text modifications and maintaining consistency across various GPT versions.

5 Discussion

The primary objective of our work is to develop a captioning system equipped with a foundational understanding of historical maps, which is not yet achieved by existing image-captioning models. By supervising the CLIP model to predict keywords related to “what”, “where”, and “when” on historical maps, we efficiently extract and represent essential map content. These keywords play a crucial role in our map captioning system, as they enable accurate comparison with ground truth data—keywords derived from map metadata—and ensure that the narratives generated by GPT are grounded in factual information.

As mentioned in 4.1, our fine-tuned CLIP models outperform the base model in five out of six categories. On



keywords: pictorial map, united states , transport routes

Figure 3. Keyword captions generated by the base CLIP model following our decision tree approach. As the base CLIP cannot correctly identify the map type (which is supposed to be *topographic map*), the errors propagate through the decision tree and generate wrong captions.

average, fine-tuning enhanced CLIP’s performance from 47% to 81%, representing a 72% improvement. The base model performed slightly better in the *location (pictorial)* caption category, possibly because the base CLIP has already been well-trained with a multitude of illustrations depicting the United States or the world with large graphic variations. On the other hand, we used only 290 maps to fine-tune the model for these two classes, which might lead to over-fitting. We assess the accuracy of each individual caption. If we evaluate the co-occurrence of all the relevant captions of a given map, the performance of base CLIP can be even worse. Moreover, as the base CLIP models cannot predict the map type well, the error will propagate through the whole decision tree. An example is shown in Figure 3.

Albeit not efficient enough to train an individual CLIP model for each keyword caption, our proposed architecture has the following advantages: 1) users can fine-tune specific categories independently from other keyword captions; 2) it can be easily extended to other map types and scaled to a bigger captioning system by incorporating additional tree branches and integrating additional fine-tuned CLIP models. Based on the predicted keywords, we leverage language models, i.e., GPTs, to narrate the story by

Table 5. Comparison of generated captions between our system integrating different GPT versions for storytelling and the vision-enabled GPTs that can generate captions directly, when texts on maps are missing. On the right, the title and production year (at bottom-left) are covered. Falsely predicted information is marked in red.



Test map		
Ours (GPT-3.5-turbo)	A hand-colored topographic map of France from the 19th century showcases the terrain and features for geographic analysis, military planning, and navigation purposes within the country during that era.	A 19th-century hand-colored topographic map of France. The map depicts the geographical features and terrain of France during that time and could be used for navigation, military planning, and understanding the country's landscape in the 19th century.
Ours (GPT-4o)	This hand-colored topographic map of France from the 19th century provides detailed elevation and landscape features of the region, offering insights into geographic terrains applicable to studies in historical geography, exploration, and cartography during that period.	This topographic map, hand-colored and depicting France in the 19th century, provides detailed geographical and elevation information, ideal for understanding the landscape and planning developments or travel strategies of that era across the nation's diverse regions.
Ours (GPT-4o-mini)	This hand-colored topographic map represents France in the 19th century. It portrays the country's geographical features, including elevation and terrain. Such maps could be used for historical research, education, and understanding of regional development during that period.	This 19th-century hand-colored topographic map illustrates the geographical features of France. It serves as a valuable resource for understanding the country's terrain, aiding in studies related to geography, history, and land use during that period.
GPT-4-turbo	This historical map, titled "Gallia Vetus," represents Ancient Gaul during Roman times, detailing its diverse tribal regions and main settlements. It serves educational and research purposes, offering insights into ancient geographic and political boundaries and the development of early European civilizations.	This historical map illustrates the railway networks across Spain and Portugal as of 1900 . It details rail lines, geographical features, and cities, making it a valuable tool for understanding regional connectivity and travel planning during that era.
GPT-4o	This map depicts Ancient Gaul, focusing on geographical and political regions in what is now France, during antiquity. It serves as a historical reference to understand the territorial organization and tribal distributions of that era.	This historical map depicts France and surrounding regions , focusing on political boundaries and geographical features of the time, around the 19th century. It serves as a tool for understanding historical territorial divisions and geographical knowledge during this period.
GPT-4o-mini	This map illustrates the region of Gaul as it appeared in 1820. It serves to depict administrative divisions, geographic features, and transportation routes, useful for historians, geographers, and those studying early 19th-century France.	This map depicts Romania , showcasing its administrative divisions and geographical features. It is from the 19th century and can be used for historical research and understanding regional developments during that time.

Table 6. Comparison of generated captions between our system integrating different GPT versions for storytelling and the vision-enabled GPTs that can generate captions directly, when texts on maps are incorrect. On the right, wrong title and production year are injected. Falsely predicted information is marked in red.


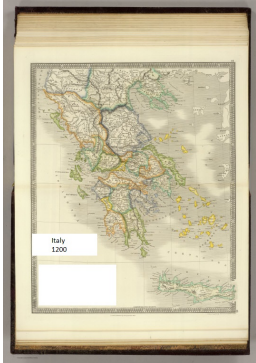


Test map		
Ours (GPT-3.5-turbo)	A 19th-century hand-colored topographic map of Greece could be utilized for geographical analysis, land surveying, and historical research, providing detailed visual information about the terrain features, elevations, and geographic characteristics of Greece as depicted in the 1800s.	A hand-colored topographic map of Greece from the 19th century depicts detailed physical features. It provides information on the country's terrain, elevations, and natural landscape, aiding in military planning, infrastructure development, and geographical studies of Greece during that era.
Ours (GPT-4o)	This hand-colored topographic map depicts the geographical features of Greece during the 19th century. It provides detailed elevation and terrain information, which could be used for historical research, educational purposes, or understanding the landscape and development of Greece in that era.	The hand-colored topographic map of 19th-century Greece illustrates the country's geographical features and elevations. Created during this period, it could be utilized for military planning, educational purposes, or enhancing understanding of Greece's diverse landscape in that era.
Ours (GPT-4o-mini)	This 19th-century topographic map of Greece is hand-colored, depicting the region's geographic features. It can be used for educational purposes, historical reference, or research into the landscape and topography of Greece during that era.	This 19th-century topographic map of Greece is hand-colored, showcasing the region's varied landscape. It can be used for geographical studies, historical research, or to understand the terrain and settlements of Greece during that period.
GPT-4-turbo	This map illustrates Ancient Greece along with surrounding island groups and inset maps of the Athenian Empire and the Environs of Athens. This historical map, rendered in detail, could serve educational and academic purposes, aiding in the study of Greek history and geography.	This map represents Italy in the year 1200 , detailing its geographical features and political divisions during the medieval period . It serves as an educational tool for understanding Italy's historical territorial boundaries and regional distinctions at that time.
GPT-4o	This map depicts Ancient Greece, focusing on its geographical regions and city-states. It provides a historical overview, likely from the classical period, and can be used for educational purposes or historical research on ancient Greek territories and their locations.	The map depicts Greece around the year 1200 , highlighting regional divisions and geographic features. This historical map can be used for educational purposes, illustrating political boundaries and geography during the period, aiding in studies of medieval Greek history and cartography.
GPT-4o-mini	This map depicts Ancient Greece, focusing on its geographic features and city-states. The map illustrates the region's landscape during classical antiquity, serving as a historical reference for educational and research purposes.	This map depicts Greece in the year 1200 . It illustrates territorial divisions and geographic features, valuable for historical research or education regarding medieval Greek topography and settlement patterns.

Table 7. Comparison of generated captions between our system integrating different GPT versions for storytelling and the vision-enabled GPTs that can generate captions directly, when texts on maps are altered. On the right, the title and map legend are erased using an AI tool. Falsely predicted information is marked in red.

Test map		
Ours (GPT-3.5-turbo)	A pictorial map of the world during World War 2. The map illustrates the global theater of the conflict, showcasing key locations, battles, and events from the war. It serves as a visual aid for understanding the scope and impact of World War 2.	A pictorial map of the world during World War 2, depicting key locations and events. It could be used for educational purposes or as a historical reference to visualize the global impact of the war.
Ours (GPT-4o)	This pictorial map of the world during World War 2 provides a visual representation of global events and territorial boundaries. It illustrates the geopolitical landscape of the era and could be used for educational purposes, understanding historical context, or visualizing wartime strategies.	This pictorial map illustrates global strategic locations and events of World War 2. It depicts the world during the period of World War 2 and can be used for educational purposes to understand the geopolitical dynamics and military campaigns of that era.
Ours (GPT-4o-mini)	The pictorial map of the world during World War II illustrates the geopolitical landscape and significant events of the time. It focuses on global movements and conflicts from 1939 to 1945, serving as an educational tool for understanding wartime dynamics.	This pictorial map of the world illustrates key events and themes during World War II. It focuses on various countries affected by the war, serving as a visual reference for historical education and analysis of global conflict from 1939 to 1945.
GPT-4-turbo	This map, titled “ Baird North's War Map, ” provides a comprehensive view of World War II status across the globe. It visually delineates territorial control and alliances as of 1942. It serves educational and historical analysis uses, detailing key regions, battles, and occupation zones worldwide.	This map depicts the Cold War era , highlighting the global political divisions between the Western Bloc, the Eastern Bloc, and Non-Aligned countries. It covers the entire world and is useful for understanding mid-20th-century geopolitical alignments and tensions.
GPT-4o	This map depicts global territorial control during World War II as of 1942. It highlights Axis, Allied, and neutral countries. Created for educational purposes, it provides a visual representation of geopolitical power dynamics and wartime alliances during this period.	This 1939 world map illustrates British Empire territories in red, showcasing colonial dominance across continents. It serves educational or political purposes, highlighting the global extent of British control at that time, reflecting geopolitical conditions shortly before World War II .
GPT-4o-mini	This map illustrates global military theaters during World War II, specifically from 1942. It serves to depict zones of conflict and military presence, aiding in geographical understanding of wartime strategies and alliances across various continents.	This map illustrates historical colonial empires and their territories around the world, focusing on the mid-20th century. It serves to highlight global colonial influences and could be used for educational purposes regarding imperial history and geographical changes.

completing the sentence and enriching the context. Besides, our designed web interface enables interactive storytelling, where the users can choose which aspects to include in the caption and which GPT model to use for storytelling.

Both our approach and ClipCap combine CLIP and GPT models. The difference lies in that no extra mapping network is needed between the CLIP and GPT models. Furthermore, we only fine-tune CLIP models while using GPT to extend and summarize the story at inference, which is more lightweight and efficient to train. Fine-tuning ClipCap does not lead to satisfactory results, possibly because: 1) the ClipCap model was trained on the natural images (Lin et al., 2014) and the size of our dataset is not sufficient to fine-tune the transformer-based mapping network; 2) CLIP models stay frozen in the training process of ClipCap, which can propagate errors to GPT if the visual prefix cannot be correctly obtained from maps.

Compared with vision-enabled GPT models that primarily rely on textual information to generate correct captions, our system leverages visual cues such as shapes and textures. This reliance on visual information significantly enhances the robustness of our system in cases where labels are illegible (e.g., due to the aging of printed maps or low resolution) or have been altered. Such modifications can be introduced for unethical reasons such as misrepresentation, propaganda, forgery or distorting historical context. Therefore, it is crucial for a captioning system to interpret maps without depending solely on textual information.

Our method also has limitations. First of all, GPTs may slightly hallucinate captions, especially for *why*, given that no ground truth is provided to be aligned with. Our work only evaluated the accuracy of keywords that can be compared with the ground truth, i.e., “what”, “where” and “when”. Since no ground truth is available for “why” (which is enriched by the GPT), and the quality of generated captions is aligned with the capability of the current language model per se, our work does not involve extra evaluation of the caption. The users can potentially assess whether the story is informative in future work. On the other hand, our proposed architecture requires enumerating all classes in the decision tree. The classes should be defined beforehand; thus, our proposed architecture cannot describe unseen categories/concepts.

While our system represents a first step toward automatic map storytelling, we acknowledge that the generated textual descriptions currently fall short of richer narrative depth. The outputs are intentionally concise and focus on summarizing the key spatial and temporal information; they do not yet capture historical context, causality, or thematic progression—key elements of storytelling in a broader sense. This limitation stems from both the nature of the input (often sparse and limited cartographic content) and the current capabilities of general-purpose language models when applied to niche, domain-specific visual inputs like historical maps. Future work will explore

ways to enrich the descriptive outputs with contextual historical knowledge, integrate multi-modal reasoning (e.g., combining text, map features, and external metadata), and move toward generating narratives that not only describe, but also interpret maps in ways that support educational and scholarly storytelling goals.

6 Conclusion and outlook

While existing image captioning methods show promising results on natural images, their performances for maps remain suboptimal in terms of caption accuracy and granularity. Our proposed method outperforms ClipCap in map storytelling and is more stable than other vision-enabled GPT models when missing or altered text in maps. Compared with vision-enabled GPT models, our proposed lightweight method can be easily used to fine-tune map captioning with users’ private or proprietary datasets. Moreover, our system has a scalable decision tree architecture that is flexible to adapt and extend. However, there are also limitations. The current system focuses on broad periods (e.g., centuries) for identifying *when*, which can fail to capture significant historical nuances. Additionally, the caption quality depends on the current language model’s capabilities, which may lack depth in explaining the *why* behind a map. In the future, more efficient ways can be explored to automatically generate a larger and more diverse map dataset. Moreover, the generated narratives should be further enriched and deepened by integrating other knowledge bases for storytelling in a broader sense. The caption quality can be further strengthened and evaluated via user study. Combined with our decision tree approach, it would allow the development of a more powerful (historical) map captioning system.

Data and Software Availability

The data and code to reproduce our results on the test set with our approach are available at <https://github.com/claudaff/automatic-map-storytelling>.

Declaration of Generative AI in writing

The authors declare that they have used Generative AI tools only to a minor degree in the preparation of this manuscript. Specifically, the AI tools were utilized for improving grammar and language editing, but not for generating scientific content, research data, or substantive conclusions. All intellectual and creative work, including the analysis and interpretation of data, is original and has been conducted by the authors without AI assistance.

References

- Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6077–6086, 2018.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners, *Advances in neural information processing systems*, 33, 1877–1901, 2020.
- Chen, X. and Zitnick, C. L.: Learning a recurrent visual representation for image caption generation, *arXiv preprint arXiv:1411.5654*, 2014.
- Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., and Forsyth, D.: Every picture tells a story: Generating sentences from images, in: *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision*, Heraklion, Crete, Greece, September 5–11, 2010, *Proceedings, Part IV* 11, pp. 15–29, Springer, 2010.
- Hodosh, M., Young, P., and Hockenmaier, J.: Framing image description as a ranking task: Data, models and evaluation metrics, *Journal of Artificial Intelligence Research*, 47, 853–899, 2013.
- Hossain, M. Z., Sohel, F., Shiratuddin, M. F., and Laga, H.: A comprehensive survey of deep learning for image captioning, *ACM Computing Surveys (CSUR)*, 51, 1–36, 2019.
- Hu, Y., Gui, Z., Wang, J., and Li, M.: Enriching the metadata of map images: a deep learning approach with GIS-based data augmentation, *International Journal of Geographical Information Science*, 36, 799–821, 2022.
- Kent, A.: Topographic maps: methodological approaches for analyzing cartographic style, *Journal of Map & Geography Libraries*, 5, 131–156, 2009.
- Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980*, 2014.
- Kiros, R., Salakhutdinov, R., and Zemel, R. S.: Unifying visual-semantic embeddings with multimodal neural language models, *arXiv preprint arXiv:1411.2539*, 2014.
- Li, J. and Xiao, N.: Computational Cartographic Recognition: Identifying Maps, Geographic Regions, and Projections from Images Using Machine Learning, *Annals of the American Association of Geographers*, 113, 1243–1267, 2023.
- Li, S., Kulkarni, G., Berg, T., Berg, A., and Choi, Y.: Composing simple image descriptions using web-scale n-grams, in: *Proceedings of the fifteenth conference on computational natural language learning*, pp. 220–228, 2011.
- Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L.: Microsoft COCO: Common Objects in Context, <http://arxiv.org/abs/1405.0312>, retrieved January 02, 2024, from <https://cocodataset.org/#home>, 2014.
- Mokady, R., Hertz, A., and Bermano, A. H.: Clipcap: Clip prefix for image captioning, *arXiv preprint arXiv:2111.09734*, 2021.
- Ordonez, V., Kulkarni, G., and Berg, T.: Im2text: Describing images using 1 million captioned photographs, *Advances in neural information processing systems*, 24, 2011.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision, 2021.
- Schnürer, R., Sieber, R., Schmid-Lanter, J., Öztireli, A. C., and Hurni, L.: Detection of pictorial map objects with convolutional neural networks, *The Cartographic Journal*, 58, 50–68, 2021.
- Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., and Cucchiara, R.: From show to tell: A survey on deep learning-based image captioning, *IEEE transactions on pattern analysis and machine intelligence*, 45, 539–559, 2022.
- Sun, C., Gan, C., and Nevatia, R.: Automatic concept discovery from parallel text and visual corpora, in: *Proceedings of the IEEE international conference on computer vision*, pp. 2596–2604, 2015.
- Touya, G., Brisebard, F., Quinton, F., and Courtial, A.: Inferring the scale and content of a map using deep learning, *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43, 17–24, 2020.
- Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention, in: *International conference on machine learning*, pp. 2048–2057, PMLR, 2015.
- Yao, T., Pan, Y., Li, Y., Qiu, Z., and Mei, T.: Boosting image captioning with attributes, in: *Proceedings of the IEEE international conference on computer vision*, pp. 4894–4902, 2017.
- You, Q., Jin, H., Wang, Z., Fang, C., and Luo, J.: Image captioning with semantic attention, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4651–4659, 2016.
- Zeng, X.-H., Liu, B.-G., and Zhou, M.: Understanding and generating ultrasound image description, *Journal of Computer Science and Technology*, 33, 1086–1100, 2018.
- Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., and Gao, J.: Unified vision-language pre-training for image captioning and vqa, in: *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 13 041–13 049, 2020.
- Zhou, X., Li, W., Arundel, S. T., and Liu, J.: Deep convolutional neural networks for map-type classification, *arXiv preprint arXiv:1805.10402*, 2018.