



# Generative Text-to-Image Diffusion for Automated Map Production Based on Geosocial Media Data

Alexander Dunkel<sup>1</sup>  · Dirk Burghardt<sup>1</sup>  · Madalina Gugulica<sup>1</sup> 

Received: 27 October 2023 / Accepted: 20 December 2023 / Published online: 13 February 2024  
© The Author(s) 2024

## Abstract

The state of generative AI has taken a leap forward with the availability of open source diffusion models. Here, we demonstrate an integrated workflow that uses text-to-image stable diffusion at its core to automatically generate icon maps such as for the area of the Großer Garten, a tourist hotspot in Dresden, Germany. The workflow is based on the aggregation of geosocial media data from Twitter, Flickr, Instagram and iNaturalist. This data are used to create diffusion prompts to account for the collective attribution of meaning and importance by the population in map generation. Specifically, we contribute methods for simplifying the variety of contexts communicated on social media through spatial clustering and semantic filtering for use in prompts, and then demonstrate how this human-contributed baseline data can be used in prompt engineering to automatically generate icon maps. Replacing labels on maps with expressive graphics has the general advantage of reaching a broader audience, such as children and other illiterate groups. For example, the resulting maps can be used to inform tourists of all backgrounds about important activities, points of interest, and landmarks without the need for translation. Several challenges are identified and possible future optimizations are described for different steps of the process. The code and data are fully provided and shared in several Jupyter notebooks, allowing for transparent replication of the workflow and adoption to other domains or datasets.

**Keywords** Generative AI · Diffusion models · Text-to-image · Image-to-image · Cartography · Social media

## Automatische Kartenproduktion mithilfe generativer Text-zu-Bild Diffusion unter Nutzung von raumbezogenen Daten sozialer Medien

### Zusammenfassung

Der Stand von generativen KI hat mit der Verfügbarkeit von Open-Source-Diffusionsmodellen einen großen Entwicklungssprung vollzogen. Hier demonstrieren wir einen integrierten Arbeitsablauf, der im Kern Text-zu-Bild-Diffusion nutzt, um automatisch Icon-Karten zu generieren, z.B. für das Gebiet des Großen Gartens, einem touristischen Hotspot in Dresden. Der Arbeitsablauf basiert auf der Aggregation von raumbezogenen Daten aus den sozialen Medien Twitter, Flickr, Instagram und iNaturalist. Diese Daten werden verwendet, um textbasierte Eingabeaufforderungen („Prompts“) für den Diffusionsprozess zu erstellen. Ziel ist es, raumbezogene kollektive Bedeutungs- und Wertzuschreibungen der Bevölkerung bei der Kartenerstellung zu berücksichtigen. Insbesondere stellen wir Methoden zur Verfügung, welche helfen, die Vielfalt der in sozialen Medien kommunizierten Kontexte zu generalisieren und zu reduzieren. Durch räumliches Clustering und semantische Filterung erzeugen wir vereinfachte Zusammenfassungen und verwenden diese in Prompts. Der Prozess zeigt, wie von Menschen bereitgestellte Basisdaten mittels „Prompt-Engineering“ zur automatischen Generierung von Icon-Karten verwendet werden können. Das Ersetzen von Kartenbeschriftungen durch aussagekräftige Grafiken hat den grundsätzlichen Vorteil, dass ein breiteres Publikum, z. B. auch Kinder und Legastheniker, erreicht werden kann. Die daraus resultierenden Karten

---

✉ Alexander Dunkel  
alexander.dunkel@tu-dresden.de

<sup>1</sup> Institute of Cartography, TU Dresden, 01062 Dresden, Germany

können darüber hinaus verwendet werden, um Touristen jeglicher Herkunft über wichtige Aktivitäten, Sehenswürdigkeiten und Wahrzeichen zu informieren, ohne dass eine Übersetzung erforderlich ist. Es werden verschiedene Herausforderungen identifiziert und mögliche zukünftige Entwicklungen für verschiedene Schritte des Prozesses beschrieben. Der Quellcode und die Daten werden vollständig zur Verfügung gestellt und in mehreren Jupyter-Notebooks geteilt, was eine transparente Replikation des Arbeitsablaufs und die Übertragung auf andere Gebiete oder Datensätze ermöglicht.

## 1 Introduction

Generative artificial intelligence (Generative AI) describes the use of deep learning to create text, images, or other media by reproducing patterns, structures, and relationships from input training data. Notable examples of generative AI include large language models (LLMs), such as ChatGPT, Bard, and LLaMA, and diffusion-based text-to-image models (Rombach et al. 2022), such as Midjourney, DALL-E, and Stable Diffusion. All text-to-image and text-to-text models share the ability to be guided by natural language input prompts. Schetinger et al. (2023) highlight the possibilities of applying generative models to all stages of the creative content creation process and illustrate a number of challenges and opportunities that have emerged. Examples of opportunities include reducing the mental effort required to design appropriate data models, automating tedious or repetitive cleanup tasks, speeding up processes, shaking up entrenched beliefs, suggesting novel color mappings, counteracting user biases, and adapting visualizations and maps to user needs, preferences, and disabilities.

A key issue discussed is agency, which is used to describe the ability of analysts to modify the outcome of the content generation process (Epstein et al. 2023). For example, the study by Schetinger et al. (2023) found that unreliable results, misinformation, and limited control over the content generation process were perceived by users as the greatest barriers. In this paper, we consider graphical content creation from a cartographic perspective using text-to-image diffusion models. Conceptually, there are several levels at which diffusion models (DMs) can be applied to cartography. One possibility is the direct production of maps. This has been shown by Chen et al. (2021) for satellite imagery and in combination with image-to-image diffusion, or by Kang et al. (2023) for direct map generation from prompts using text-to-image. Both approaches are comparable to how neural style transfer (Ai 2022) can be used to generate maps with specific looks. What these methods have in common is a single-step generation process. This limits the ability of humans to influence and modify the generated result and can introduce unwanted inaccuracies and misleading information, an issue that may be particularly important in cartography (Kang et al. 2023).

In this work, we propose a component-based workflow that provides more control at different steps of the generation

process through the use of Stable Diffusion 1.5. This *latent* diffusion model (LDM), a specific model type that is particularly efficient and adaptable, has been trained on 5.85 billion images (Schuhmann et al. 2022) and published by the CompVis group at LMU Munich and Stability AI (Rombach et al. 2022). For prompt engineering, our approach uses human-based input data crowdsourced from people on the web (Instagram, Flickr, Twitter, iNaturalist). Based on a multi-step process, we generate prompts by aggregating geosocial media data through spatial and semantic clustering of (e.g.) tags and emoji. These prompts are then used in a text-to-image diffusion process to generate graphical icons. The icons are further processed before being scaled based on collective importance, placed on the map, and rendered together with other map elements such as background tiles. The resulting maps can be used to promote inclusive cartography (Holloway et al. 2019), for example by targeting children or illiterate groups. Our study highlights both the benefits of component-based generative cartography and the challenges that require further investigation.

## 2 Literature Review

The use of generative AI is currently undergoing a very dynamic development. Similar to García-Peñalvo and Vázquez-Ingelmo (2023), we use “generative AI” as an umbrella term to describe the emerging class of text-to-text, text-to-image, and image-to-image models. These models have great potential to augment and replace human creativity in many application domains, such as visualization (Schetinger et al. 2023) or as an aid to map generation (Juhász et al. 2023). So far, Juhász has been the only one to explore potential applications of text-to-text models in cartography. Current research in this area focuses mainly on image-to-image and text-to-image transformations, with the former accounting for the vast majority.

One of the first approaches utilizing image-to-image transformations to generate map representations was proposed by Kang et al. (2019). They used generative adversarial networks (GANs) to transform GIS vector data into Google Map-style maps, mimicking different painting styles. While the results were encouraging, the study revealed problems with topology, point markers, and text labels. Similarly, Chen et al. (2021) used GANs for image-to-image

transformation to generate Google Map-like map tiles from high-resolution satellite imagery. The authors used semi-supervised learning with paired and unpaired training examples. In a follow-up study (Chen et al. 2022), this approach was extended to multiscale map generation by transforming already styled tiles. The method of neural style transfer based on convolutional neural networks (CNNs) was first applied to artistic map generation by Bogucka and Meng (2019). The authors transferred painting styles with different emotional expressiveness to a map of Berlin. Christophe et al. (2022) used GAN-based neural maps to transfer historical map styles (e.g., Cassini, Etat-Major) to orthoimages. Zhao et al. (2021) explored the reverse transformation, using Cycle-Consistent Adversarial Networks (CycleGAN) to transform CartoDB topographic basemaps into fictional satellite images, highlighting the potential for AI manipulation.

Diffusion models were introduced by Sohl-Dickstein et al. (2015), but did not gain traction until 2022, when OpenAI made DALL-E 2 available. This model was the first to produce images with a quality comparable to photographs or human-drawn art. At the time of this study, Imagen (Google), Midjourney, and Stable Diffusion are the three other text-to-image models that can achieve a similar level of quality (Zhang et al. 2023). These models contain pre-trained weights based on the training data. Models are also called *checkpoints*. They can be merged with other models or further trained and refined on a specific subject and set of images using various methods to modify weights (Ruiz et al. 2022). All text-to-image models have two components in common. The first is a language model that transforms an input text (the prompt) into a latent representation. This text representation is then used in the second component, the generative image model, to guide the image generation process towards the desired result. A denoising diffusion probabilistic model (DDPM) is often used here, as proposed in a seminal paper by Ho et al. (2020). The method is based on random noise that is progressively subtracted in discrete steps (e.g., 50) by the name-giving *denoising* process. The efficiency of the denoising process is evaluated on the training data. Since large diffusion models are typically trained on millions of annotated images, rich representations are available, such as low-level image knowledge (textures, colors, transitions) or high-level semantic relationships (Zhao et al. 2023), allowing for subtle evaluation of meaning and context.

Exactly how the prompt and parameters need to be defined for different image generation tasks remains to be tested and explored. Prompt engineering has emerged as a new area of research (Nori et al. 2023; Witteveen and Andrews 2022). The term describes the often iterative definition of text prompts as part of generative processes to achieve a desired output. Several mechanisms can be considered as part of prompt engineering. One example are

negative prompts (Rombach et al. 2022), which invert the diffusion evaluation so that the given text prompt representation is *not* produced. Prompt engineering also includes techniques and strategies that help to incrementally refine and improve the results. Critical to such incremental improvements is the initial *seed*, which describes the starting value of the noise. If the seed and all parameters are left unchanged, the diffusion model will deterministically produce the same image. In prompt engineering, this property can be exploited using an identical seed with a minimally changed prompt or parameter selection to evaluate subtle changes (Dehouche and Dehouche 2023; Tsai et al. 2023). For the actual text prompt, Oppenlaender (2023) suggests a taxonomy of repeatedly used prompt parts, such as “[...] subject terms, image prompts, style modifiers, quality boosters, repeating terms, and magic terms” (p. 14).

The usefulness of text-to-image diffusion models for cartography was first explored by Kang et al. (2023). They used DALL-E (a closed-source model from openai.com) to generate sample maps from user-defined prompts (e.g., “A choropleth map of [sic] United States with warm colors”, p. 3, *ibid.*). In their summary, the authors discuss ethical implications and note that inaccuracies, misleading information, unexpected features, and irreproducibility are major issues. Unexplored so far are workflows that use text-to-image diffusion models as components to generate only part of the map representation. This is surprising given that maps typically consist of multiple data sources that are processed individually and then arranged and assembled into different layers, symbols, and labels, etc. to communicate information (e.g., Wood and Fels 1986). The meaning of maps can vary considerably between different types. For example, topological maps reduce information and detail to a necessary minimum while maintaining topological relationships to support way-finding in (e.g.) public transportation systems. In contrast, more creative map types exist, such as concept maps (Cañas et al. 2005) or tourist symbol maps (Brown et al. 2001). In their review of cartographic challenges related to geospatial big data, Robinson et al. (2017) point to the particular challenge of scaling the output of artists in mapmaking, emphasizing that “a solution needs to be found for the digital generation of artworks that are both meaningful wholes and are somehow a true representation of a big data set” (p. 45).

Icons and symbols are particularly well suited to begin such an endeavor, as they are pictorial representations of largely independent concepts that can be visualized on maps. Lin et al. (2014) explored an automated icon generation process from 3D models for landmark representation on maps, but their approach is limited to buildings for which 3D models exist. In addition, pictorial icons can be used to convey meaning beyond known landmarks. Maps can illustrate activities, objects such as benches (etc.), neighborhoods of communities, points of interest, or ephemeral features that

recur at particular locations such as events, wildlife (etc.). Brown et al. (2001) attempt to conceptualize this process and propose to distinguish three types of information levels, (1) primary content (symbols, objects and other map focal points), (2) secondary content (base map, topographic information), and (3) supporting content (legend, grid, additional information). Icons and symbols, in particular, can support the main theme of the map and may be scaled according to perceived importance or arranged in a particular order and pattern by the mapmaker (Wood and Fels 1986).

Here, generative AI methods offer to fill the important gap of the limited supply of artistic production in cartography (Zhang et al. 2023) to draw maps of any scale and size. In addition, with the widespread availability of data shared on geosocial media, a data source exists to map and visualize collectively weighted subjective information of large groups of people (Dunkel 2015). We combine these two opportunities to automatically generate icon maps. Since many steps are non-trivial tasks on their own, such as icon offset placement, semantic similarity-based clustering to detect outliers in a group of words, or geosocial media data mapping, we limit our description to the novel parts of the process. In particular, we focus on the prompt engineering of the text-to-image process, the parameterization, and the compilation of the final map. To reproduce the results and graphics shared here, three Jupyter notebooks are made available in the Supplementary Information (S1–S3) and are provided along with the data and code in a separate data repository (Dunkel et al. 2023).

## 3 Methods

### 3.1 Prompt Engineering Overview

Prompt engineering is at the core of our workflow. We begin with an overview of this step and later describe the tasks that precede and follow it. Figure 1 illustrates the main parts of our prompt engineering approach in the context of map icon generation. We distinguish three steps. Input data (1) derived from geosocial media (highlighted in red in Fig. 1) is used to derive the subject prompt for the icons. This data are referenced as a placeholder in the base-positive prompt, which we consider, together with the base-negative prompt, to be the core of prompt engineering (2). The positive prompt also consists of a static pre-prompt, which is a leading placeholder that is used equally across all images to improve the ability to later place graphics without background on a map. Other parts of the base-positive prompt are used as style and quality modifiers. Similarly, the base-negative prompt is applied to all images. While stable diffusion is invariant to

commas, spaces, punctuation (etc.), we still use these characters to improve readability. To generate images (3), we use an application programming interface (API) to Stability AI's *diffusers* library,<sup>1</sup> which is provided by a Python package.<sup>2</sup> The package offers several special operators in text prompts to direct the attention of the diffusion process. For example, parentheses "(" and ")" can be used to increase the attention for selected words by a factor of 1.1, or a colon (:) can be used to define the attention factor directly, followed by the weight (see Fig. 1).

The remaining static prompt parts (Fig. 1) were constructed from sample texts shared by the stable diffusion community and our own empirical testing. This testing process is not deterministic. We make use of several specific parameters and models to fine-tune the results, such as a low-rank adaptation (LoRA) for style transformations (detail reduction) or a negative textual inversion embedding. These additional parameters and model-related parts are described in Sect. 3.3.

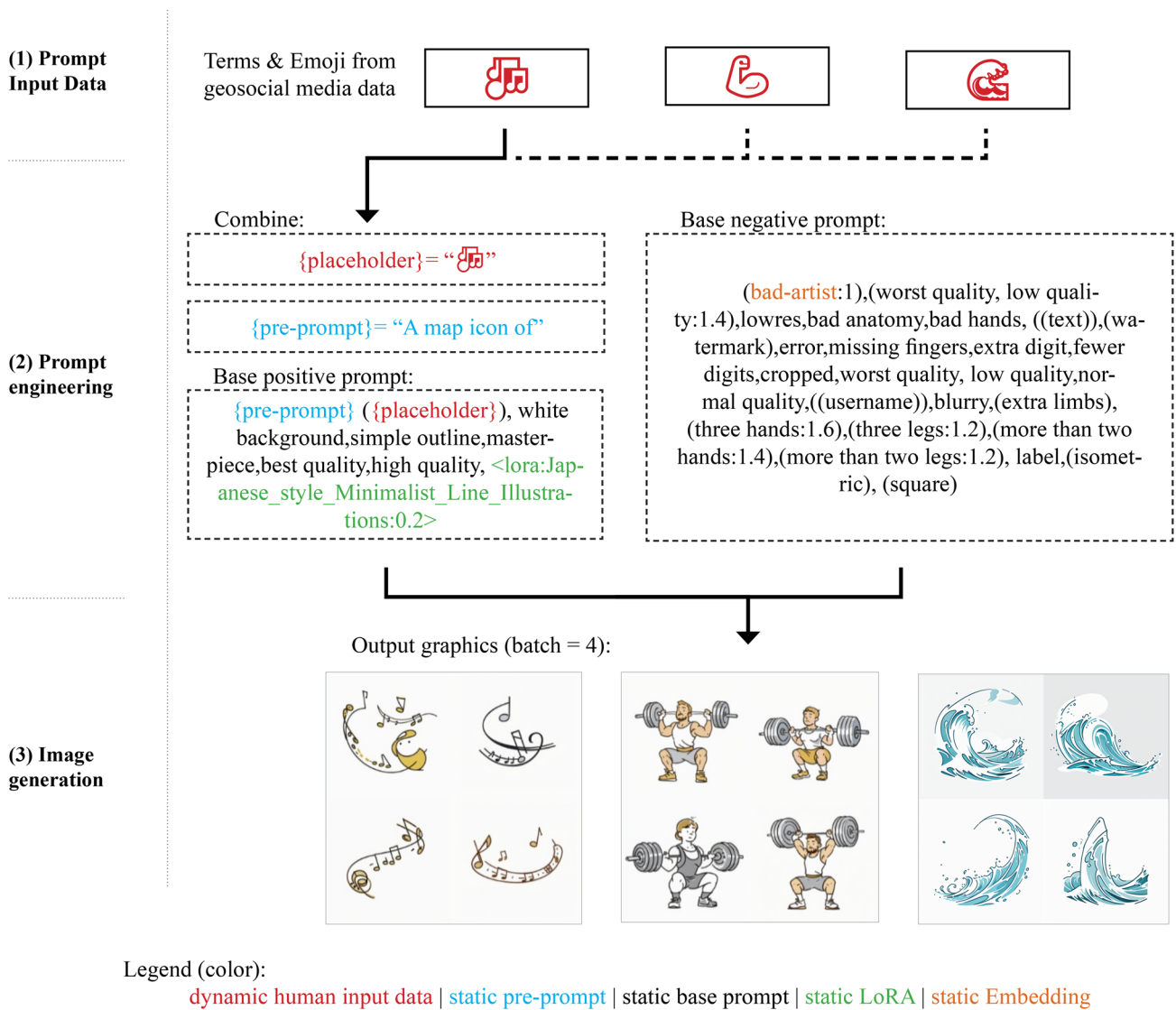
### 3.2 Map Production Workflow

To generate maps, several additional steps are required before and after the prompt engineering. Our entire workflow can be broken down into eight largely independent steps.

1. Subject prompt preparation
  - a. Area selection
  - b. Data collection
  - c. Geosocial media preprocessing and clustering
  - d. Spatial cluster merge
  - e. Semantic cluster split
2. Model selection
  - a. Checkpoint model,
  - b. LoRA (Low-rank adaptation), Hu et al. (2021); Textual Inversion, Gal et al. (2022)
  - c. Pretrained VAE (Variational Auto Encoder), Kingma and Welling (2022)
3. Parameter selection
4. Prompt engineering (see Fig. 1, Sect. 3.1)
5. Image generation
6. Background removal
7. Map compilation (image placement, scaling, composition mode, background map)
8. Optional image-to-image pass

<sup>1</sup> <https://github.com/Stability-AI/stablediffusion>.

<sup>2</sup> <https://github.com/AUTOMATIC111/stable-diffusion-webui>.



**Fig. 1** Prompt engineering: from input data, such as emoji shared on social media, to generated output icons. Different types of prompt parts are highlighted

The first step (1) is to prepare the subject prompt based on the aggregation of geosocial media data. The code for this process is provided in Supplementary Information S1 and S3. The data are provided in a separate repository (Dunkel et al. 2023). In this step, we first select an area and a scale for which the map is to be generated (1a). Next (1b), data are collected from public geosocial media APIs and transformed into a common data structure format (see Dunkel et al. 2020). The data (individual terms and emoji) are then (1c) spatially and thematically clustered using HDBSCAN (McInnes et al. 2017). From these clusters, alpha shapes for different terms and emojis are derived using the

tagmaps package.<sup>3</sup> Since we are interested in the use of multiple terms and emojis in prompts, we add a spatial cluster merging step (1d) to combine different terms and emoji that are used in the same region (e.g., “stadium”, “football”, “game”).

Sometimes several terms in the same spatial cluster group can refer to semantically different contexts (e.g., at different times). For example, the term “rammstein”, a reference to a music band that once performed at the Dynamo Dresden stadium, may appear as a semantic outlier in a group of terms related to soccer games (e.g., “stadium”, “football” and “game”). This is not surprising given that other types of events at the stadium, such as concerts, are

<sup>3</sup> <https://gitlab.vgiscience.de/ad/tagmaps>.

also communicated by social media users in addition to the prevalent football-related events. We wanted to cover these scenarios by creating different graphics for sufficiently different contexts. For this reason, we add a semantic cluster splitting step (1e) based on hierarchical clustering and cosine similarity computation. In this step, we identify the primary topic and split unrelated meanings (i.e., terms) into a second semantically different subject prompt. Later, we decided to still display these subordinate meanings as individual images, but to reduce their size (by scaling) on the map. To separate primary topic and unrelated outliers, we use a static threshold based on the normalized cosine similarity metric between word vectors (see S3 and Supplementary Materials, Dunkel et al. 2023).

Next, models (2) and parameters (3) for the diffusion process must be selected, as described in Sect. 3.3. Settings, data and prompt engineering (4) are brought together in the core image generation step (5). We generate images in batches of four, which makes it possible to compare output maps for different iterations, without having to regenerate images. Then, to place images on a map, we use Dichotomous Image Segmentation (DIS, Qin et al. 2022) together with the “isnet-general-use” model (Jin et al. 2021) to remove white backgrounds from images (6). The final map compilation (7) is implemented in Mapnik and includes several additional tasks, such as icon offsetting based on collisions, icon scaling based on weights (i.e., frequency of use on geosocial media), or adding a background map layer. Finally, we tested an optional ad hoc image-to-image pass (8) to unify and fuse the icons with the background map.

### 3.3 Model and Parameter Selection

The diffusion process can be influenced by the choice of base model and various parameters. The number of parameters and models available necessitated the use of several strategies to verify suitability and progress. Our overall goal was to produce minimalist graphics that could be placed on maps, since it is not practical to represent the full complexity of the contexts discussed by online users. We first selected parameters that had the greatest impact on the generated graphics (e.g., the checkpoint model), and then gradually added other models and parameters that had less impact. A primary distinction must be made between image content and style (Oppenlaender 2023). Style parameters affect all images and should therefore be tested to work across a range of prompts and objects. In contrast, the generated image content depends mainly on the dynamic prompt parts that change from image to image. The key strategic decisions here are the number of dynamic terms to use, the order of terms, or the selection and combination of operators. The evaluation of these content-related decisions is more difficult to do holistically, and we present individual results for

a selected area and a limited number of image examples in Sect. 4. In the following, we summarize strategic decisions that affect the style of all generated graphics. Here, the freedom of the generative process is progressively reduced during the prompt and parameter refinement process, making outliers and unwanted results less likely. To evaluate effects in isolation, we either left the seed unchanged to observe small changes in the prompt, or left the prompt unchanged while minimally changing a single parameter.

This iterative testing and refinement process for model and parameter selection is shared in S2 and S3 (Dunkel et al. 2023). We selected a checkpoint model called “hellofunnycity”.<sup>4</sup> There has been a phenomenal growth of these models shared on platforms such as civitai.com or huggingface.co. However, the majority of models focus on human figures and improving the details of drawn humans by providing additional training such as for 5-finger hands. A few models include additional training data for objects or styles. Our selection above is one such model. It is based on the Stability AI 1.5 base model and is further trained to produce comic-style graphics. The generated graphics are less detailed than realistic images and therefore more suitable for our application. To further reduce complexity, we use a LoRA (Low-rank adaptation, Hu et al. 2021) pre-trained on Japanese minimalist line drawings.<sup>5</sup> For both the LoRA and the checkpoint, we were unable to contact the original authors and therefore cannot provide information on the training procedure or the underlying data used. The use of data of unknown origin makes it all the more important to systematically evaluate effects. Here, generating image series with a random seed greatly improved our evaluation, since a given combination of prompt and parameters may produce only a certain percentage of desired results.

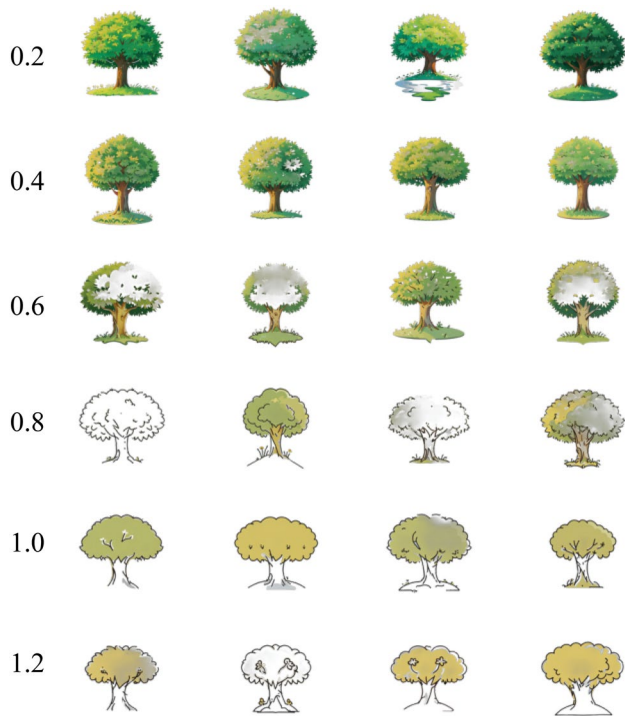
To better illustrate this process for the LoRa, compare the following text prompt with a tree emoji (🌳) surrounded by the attention operator “(” and “)”, and the LoRa reference with an explicit attention weight attached (“1.0”).

```
“(🌳) <lora:Japanese_style_Minimalist_Line_Illustrations:1.0>”
```

Our goal was to find an appropriate weight for the LoRA to be used in all images. Each row (Fig. 2) shows a series of four images. Each image was generated with a different random seed for the same attention weight per row, ranging from 0.2 (low weight) to 1.2 (high weight). Indeed, the six series allow to observe a gradual progression towards the minimalist Japanese line drawing style, with higher weights showing an increasing reduction of detail. At the same time, artifacts begin to appear at a weight of 0.6, such as blurred

<sup>4</sup> <https://civitai.com/models/124336/hellofunnycity>.

<sup>5</sup> <https://civitai.com/models/124933/japanesestyleminimalistlineillustrations>.



**Fig. 2** Incremental weight increase for the Japanese minimalist line drawings LoRa, with attention increasing from 0.2 (top row) to 1.2 (bottom row) for the subject prompt “(🌳)” (the tree emoji), see Supplementary Information S1 (Dunkel et al. 2023)

**Table 1** Overview of data sources and queried data

	Instagram	Flickr	Twitter	iNaturalist
Photos	106,385	19,729	7183	417
Users	50,628	1839	2152	120
Period	2010–2018	2007–2022	2016–2022	2010–2022

and white areas in the treetops. At higher weights, more and more tree images would be generated without color, which would be difficult to read on a map. We created several of these series for different emoji to test the suitability of the LoRa for different content. For some emoji (e.g., the water wave emoji 🌊), higher values tended to produce figure shapes (see S2). This is likely due to an overrepresentation of images depicting people in the LoRa training data. Based on these findings, we chose a relatively low weight of 0.2 for the LoRa across all generated images, which we found to be more robust and controllable.

Other settings had smaller but still noticeable effects. For example, we used a Variational Auto Encoder (VAE, Kingma and Welling 2022) to improve image quality (e.g., vivid colors,

sharper details) and reduce artifacts in the generated images.<sup>6</sup> The VAE also had the effect of converting light gray backgrounds to pure white, which is more suitable for background removal. Furthermore, in the base-negative prompt, we reference a textual inversion (Gal et al. 2022) that contains embeddings of negatively perceived vocabulary (“bad-artist”, Fig. 1).<sup>7</sup> These embeddings teach the base model new vocabulary about a particular concept, which can be referenced in the prompt using a *token* like the one above. The remaining settings are mostly defaults, such as a “clip skip” of 1 (no skipping of text embedding layers, for more specificity), a medium number of sampling steps of 28 (slightly increases quality while keeping resource requirements relatively low), a relatively high denoising strength of 0.75 (allows more creativity), and a Classifier-Free Guidance (CFG) scale of 7. The CFG scale defines how closely the model follows the input prompt, with lower values allowing for more creative interpretations. A CFG scale of 7 results in medium fidelity between the prompt and the output images.

Finally, a large number of samplers are available to influence the iterative denoising process. Sampling describes how noise is computed and used to incrementally improve (*denoise*) the results over many steps, given the model training data and prompt. Samplers range from generic sampling methods (e.g., Euler, Heun; Roberts 2012) to specific diffusion samplers such as Karras et al. (2022), DPM (Lu et al. 2022), or DDIM (Song et al. 2022). Samplers can be distinguished based on their behavior for improving images over steps. Generic samplers belong to the group of deterministic sampling methods, where the only source of randomness is the initial noise defined by the seed. This means that after a certain number of steps, the details of the image content do not improve any further, at which point the sampling is said to have *converged*. In contrast, the second group of stochastic samplers has been shown to produce better output quality by injecting fresh noise into the image at each step (Karras et al. 2022). These samplers still produce the same images after a certain number of steps given the same seed, but they do not converge. In other words, each image detail can be transformed into another as more sampling steps are used. The sampler we used (DPM++2M SDE Exponential) is such a non-deterministic sampler. The main aspects we considered were efficiency and quality, i.e., producing an image of high perceptual quality quickly. Our choice tended to produce high quality images with a relatively small number of sampling steps. A trade-off was that the sampler sometimes produced artifacts such as extra limbs or convoluted fusions of different contexts such as different animals combined into one. We found these artifacts acceptable because the map icons are small and the sampler was able to quickly produce

<sup>6</sup> <https://huggingface.co/stabilityai/sd-vae-ft-mse-original>.

<sup>7</sup> <https://huggingface.co/nick-x-hacker/bad-artist>.



**Fig. 3** Spatial cluster merging based on cluster area (alpha shape size) and cluster weight (user and post frequency). Left image: cluster boundaries for weights  $> 10$  and  $\leq 100$ . Right image: largest cluster

images of high perceptual quality given the large variability of contexts communicated in social media.

### 3.4 Data Collection

We chose a test area in Dresden, Germany, which covers about 6.0 km<sup>2</sup> of the city and includes the largest public park, the *Großer Garten*. This area is popular among tourists because it offers a variety of places and points of interest for different activities. We queried available metadata (titles, descriptions, tags, emoji) for publicly shared and georeferenced images from the APIs of four social networks (Table 1). For iNaturalist, taxonomic species references were converted to emoji when a matching Unicode emoji character was available. The data contains a total of 347,191 tags and 44,427 emoji, which form the basis of our aggregation and visualization workflow.

### 3.5 Spatial and Semantic Clustering

Processing geosocial media from multiple sources is a non-trivial task, with challenges often referred to as the 5V's of big data (velocity, volume, value, variety, and veracity) (Laney 2001; Robinson et al. 2017). We first transform the data into a common cross-network geosocial data structure.<sup>8</sup> We use a number of preprocessing and filtering techniques to prepare the geosocial media data prior to analysis and clustering. Most of these preprocessing steps are applied during the ingestion step of the tagmaps package and are based on default parameters without the need for human

group covering more than 20% of the study area, with terms related to the Großer Garten park in Dresden, Germany

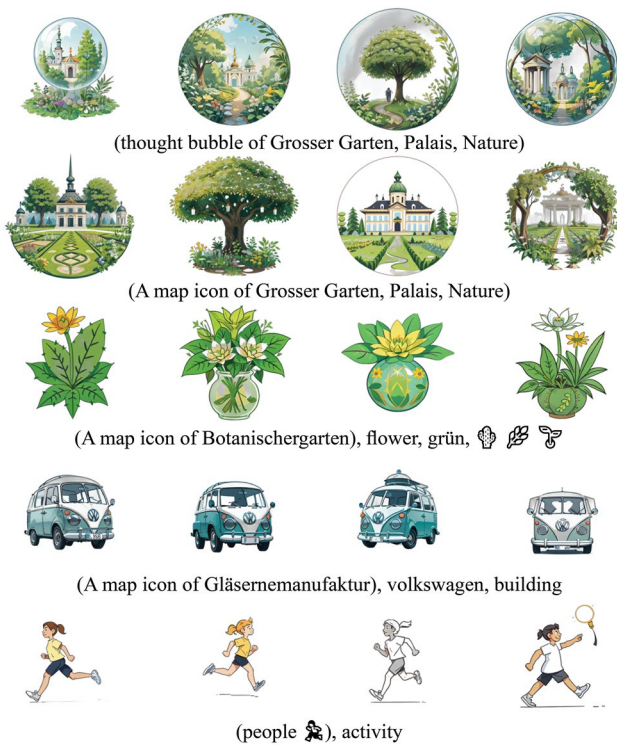
intervention. These steps include a broad stop list for filtering specific terms and tags (e.g., exclude only numbers, fill words with little lexical content, or camera brands, etc.). For emoji filtering, since some emoji are used very generically in many contexts, we use a broad positive filter list of 693 emojis (out of about 2000 available) to focus on activity and environment contexts (see S3).

To create prompts with a maximum of three terms for each graphic, we had to aggregate the data based on spatial and semantic similarities. We also needed to limit the number of graphics generated to reduce the overlap of icons on the map. This process must also be automated, as human supervision is difficult given the volume of data available. Spatial clustering is performed using HDBSCAN, which groups tags and emojis into spatially distinct point clouds. Area or *alpha shapes* are then created for these point clouds as concave and convex hulls based on Delaunay triangulation (the `scipy.spatial` implementation). Figure 3 shows a subset of these alpha shapes for the area of the Großer Garten. Each alpha shape represents a distinct footprint of a collectively used emoji or term communicated on geosocial media. Each shape is assigned a weight calculated based on the frequency of users and photos referencing that term or emoji.

Since each alpha shape references only one term or emoji, and we wanted to generate prompts with multiple terms, these individual term and emoji clusters are then combined into “cluster groups” (a spatial union operation) based on several criteria. First, the largest clusters by percentage of total area are selected to extract “background contexts” that cover the majority of the study area. We use a cutoff area percentage value of  $> 20\%$ , which extracts all terms referring

<sup>8</sup> <https://lbsn.vgiscience.org>.

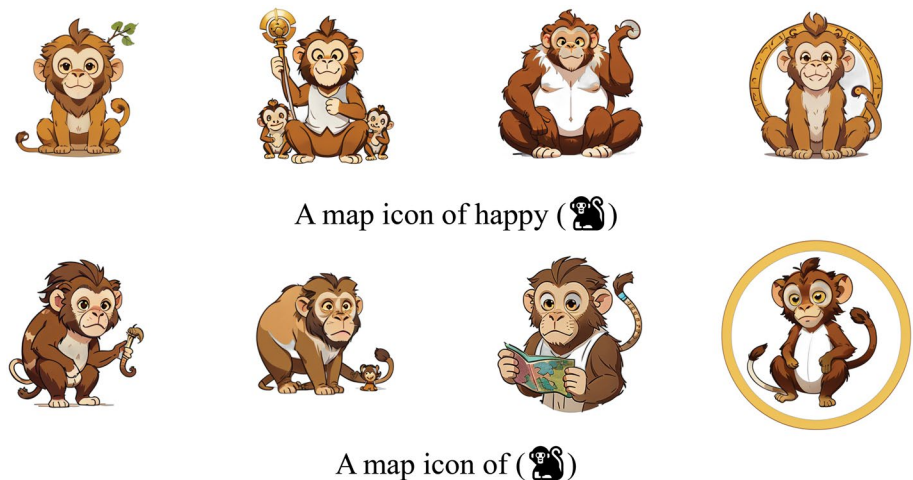




**Fig. 4** Examples of fine-tuning parameters and prompts. First two rows: pre-prompt “thought bubble of” vs. “A map icon of”, for white background, non-square image generation. Third and fourth row: attention splitting with brackets “(” and “)” for cluster terms and emojis in ascending importance, based on frequency of social media use. Last row: Prompt engineering based on cluster type classification, e.g., activities (as shown), verbs, objects, places (etc.). See Supplementary Material, S3 (Dunkel et al. 2023)

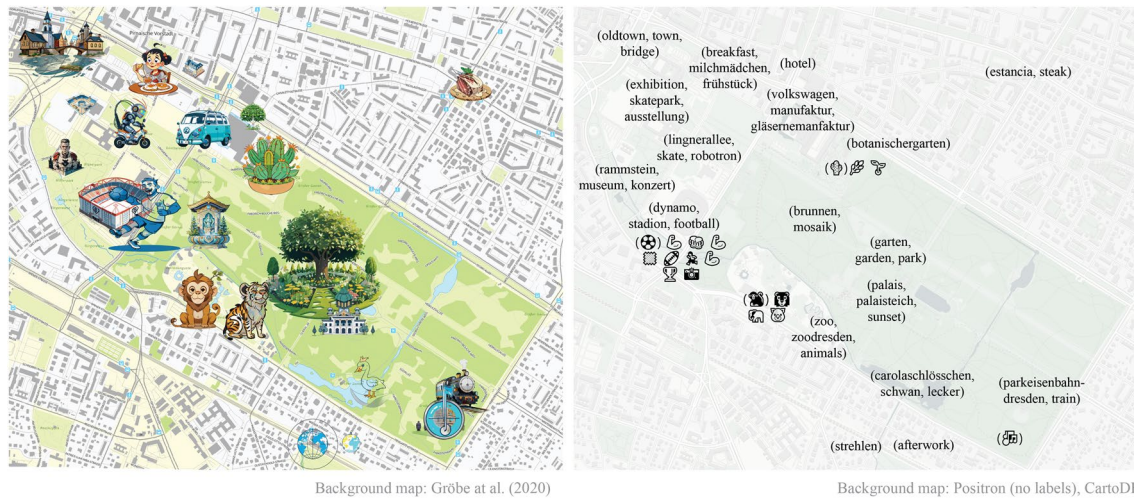
to the Großer Garten in our example (e.g., “grosergarten” [sic], “garden”, “garten”, “großer”, “park”; right image in Fig. 3). Then, all other clusters are processed sequentially and grouped based on spatial overlap until a fixed lower weight threshold is reached (e.g., weights  $\geq 10$ ). The terms

**Fig. 5** Applying quality modifiers such as “happy” (top row) to generate friendlier icons compared to images generated without modifiers (bottom row)



and emojis in each cluster group are then sorted in descending order based on usage frequency, and each cluster group is truncated to a maximum of three terms.

As social media users communicate different events from the same locations, it is not uncommon to find terms and emojis related to different topics within the same spatial cluster group. To mitigate the risk of icon clutter from unrelated topics, we propose a generalization approach that identifies the dominant topic for each cluster group. This step allows us to individually feed terms for different topics to the diffusion algorithm for separate icon generation. We implemented this semantic cluster splitting using hierarchical clustering based on word embeddings (word2vec, Mikolov et al. 2013) and cosine similarity computation. The embeddings were generated from a pre-trained Word2Vec model. These word embeddings capture the semantic properties of terms based on their co-occurrence and order in social media posts, allowing for the measurement of similarity. Since communication varies greatly across regions and languages, we specifically trained this model using geosocial media data collected for the city of Dresden (Gugulica and Burghardt 2023). We computed the cosine similarity matrix between the word embeddings of a cluster group. The cosine similarity metric is then calculated by dividing the dot product of the vectors by the product of their lengths, which indicates the strength of semantic similarity between words. The matrix is the basis for the hierarchical clustering algorithm. We use a fixed threshold (0.3) for the cosine distance. This threshold is transformed into a fixed value (0.7) for the similarity metric by subtracting the value from 1 ( $0.7 = 1 - 0.3$ ). The closer the value is to 1, the higher the semantic similarity. This threshold is used to identify the semantic cluster containing the majority of terms, which corresponds to the dominant topic. Words found in the remaining clusters are marked as semantic outliers.



**Fig. 6** Composite map (left) for the Großer Garten area in Dresden, Germany, based on generative text-to-image from geosocial media and a background tile map style from Gröbe et al. (2020). The emoji

and tags derived from the spatial clustering of geosocial data and used for the diffusion subject prompt are listed on the right

## 4 Results

Before creating the final map, we tested different combinations of pre-prompts, quality modifiers, operators, and geosocial media terms (Fig. 4). For example, to drive the diffusion process to produce non-square image icons with preferably white backgrounds, we tested “thought bubble of” (row 1, Fig. 4) and “a map icon of” (row 2, Fig. 4). We chose the latter because it produced a greater variety of icon shapes. Additional test runs focused on combining users’ emojis and tags into a single prompt and directing attention to only the first term/emoji based on ascending frequency of use (rows 3–4, Fig. 4).

Here, our results show that the diffusion process is poor for too complex combinations of terms and emojis. This is especially true for emojis, since they already reduce complex contexts to pictorial symbols. Therefore, we rejected this direction. We also found that subject terms should be limited to a maximum of three terms and that the attention operators “(” and “)” should include all three terms, instead of (e.g.) just the first one, because otherwise terms outside the attention modifiers would be largely ignored. Specifically for emoji image generation, we added a quality modifier (“happy”) because without it, the generated icons looked a bit sad and dull (bottom row, Fig. 5).

The final map (Fig. 6) consists of icons for 20 cluster groups (16 tag groups, 4 emoji groups). The icon for the “(garten, garden, park)” group, which was transformed into a tree with flower beds around it, received the highest weight and was therefore scaled the largest. The four emoji clusters mainly represent what is given as the first emoji in the cluster group, such as a cactus for the “(🌵 🌿 🌱)” group that can be found at the Botanical Garden location; a monkey for

the “(🐒 🐼 🐘)” group at the Dresden Zoo location; a soccer player for the “(⚽ 🏆 🏀)” group at the Dynamo Dresden stadium; and a somewhat hard-to-identify circular icon for the music emoji “(🎵)” at the Junge Garde, a popular open-air music venue in the southeast corner of the Großer Garten.

Similar observations to the emoji clusters can be made for tag clusters. Here, the diffusion process also tends to ignore terms for which there is little training data. An example of this is the cluster for the VW Gläserne Manufaktur (“Transparent Factory”), a brand showcase and production site, referenced by “(volkswagen, manufaktur, gläsernemanufaktur)”, which are the three most commonly used terms in this area by geosocial media users. All four images generated in the batch for these terms resulted in a VW minibus icon, ignoring the references to the production site (“manufaktur”) and its characteristic transparency (“gläserne”). This can also be observed for “(estancia, steak)”, as a reference to a restaurant in this area. The generated image shows a direct visual interpretation of the term “steak” because there is little reference data available for “estancia” as the proper name of the restaurant. For “(zoo, zoodresden, animals)”, the diffusion produced an image of a tiger, which we consider too specific for the given list of terms. Given these terms, we expected an image of several animals in a zoo environment. In other cases, such as for “(brunnen, mosaik)”, the generated image is able to convey basic characteristics (a fountain with mosaic stones), but fails to reproduce the actual unique visual appearance of the original. This is not surprising, given that Stable Diffusion 1.5 was trained on generic images scraped from the web. Creating separate graphics for semantic outliers in cluster groups resulted in a number of additional icons that are scaled down on the map. This

includes a graphical fusion of the terms “rammstein”, “konzert”, and “museum” (upper left corner of Fig. 6), which otherwise would have been merged with the terms found at the Dynamo Dresden stadium, due to spatial proximity. However, “rammstein/konzert” and “museum” (referring to the Hygiene-Museum Dresden) also refer to semantically different contexts. Solving this situation would require recursive semantic separation, which is not yet implemented in our workflow. A small robot driving a scooter is a pictorial representation of the term “robotron”, a proper name of the area, which is difficult to interpret as such on the map.

## 5 Discussion and Conclusions

Generative cartography is a new area of map making based on automated content production with unique requirements, challenges, and limitations. Our work highlights the benefits of breaking the generative process into multiple discrete steps. These components can be individually tuned and improved, helping to increase analysts' agency over the map production process. Specifically, we demonstrate how multifaceted contexts communicated in geosocial media can be aggregated and generalized for use in generative text-to-image prompts, and show how these prompts produce visual representations that can be placed on a map. Prompt engineering is at the core of our workflow. Here, human-contributed baseline data from geosocial media is fused with multiple static parts, such as base-positive prompt, base-negative prompt, modifiers, and models, to create a unique and consistently styled map. While human supervision is not required, it can help improve the readability of the produced map by incrementally calibrating and fine-tuning individual parameters. Based on our parameter fine-tuning, the map shown in Fig. 6 is already capable of illustrating complex collective meanings in the area of the Großer Garten in Dresden.

However, many settings, parameters, and model choices can be used to influence the diffusion process, and it is difficult to assess the effects a priori. We demonstrate approaches to fine-tune the effects of individual parameters, such as incrementally adjusting the attention weight for the Japanese line drawings LoRa (Fig. 2) or the composition of prompt and attention operators (Figs. 4 and 5). Many steps require a trade-off between control and creativity. More control often means less creativity, and vice versa. The same is true for the subdivision of steps in the workflow. More subdivision, such as allowing the generation of a larger number of icons for semantically different topics, will result in better pictorial representations because diffusion is easier for less ambiguous prompt terms. However, more icons will also lead to more icon clutter on the map, which is a challenge for the icon placement and scaling step. There are several solutions

that could help significantly improve these specific subproblems of cartography, such as automatic icon and label placement algorithms (Huang and Gartner 2012), which we do not currently use.

We tested four iterations based on random seeds to generate the map icons for the Großer Garten (see S3). All iterations produced icons of similar quality to the examples shown in Fig. 6. Interestingly, in contrast to text-to-text models, where larger and more specific prompts typically lead to better quality (Nori et al. 2023), our text-to-image workflow produced the most appropriate images for prompts that were limited to a few terms or concepts. In particular, prompt engineering helped improve robustness and reduce the likelihood of unwanted results, although a clear methodology is lacking in the literature. We use several options to divide prompt parts into static and dynamic components with different tasks. For example, using a static pre-prompt (“A map icon of”, Sect. 4), we were able to produce non-square icons that could be better placed on maps. Or, by adding “isometric” to the static negative prompt, we effectively prevented the generation of isometric icons. The greatest variance in output quality was introduced by the dynamic prompt parts responsible for icon content, where we also had limited control due to the use of geosocial media data. In the future, automatic quality evaluation routines based on image recognition techniques could help select the most appropriate images from a set of images. Another example with great potential for image generation, currently used only to improve LLM prompts, are “micro-agents” that fine-tune results by self-editing prompts.<sup>9</sup>

Our map production workflow, based on multiple layers that are combined at the end, helps maintain topographic order, but results in relatively little fusion between background tiles and foreground symbols. This is evident when comparing our map to human-drawn symbol maps (e.g., Antoniou et al. 2015; Child 1956), where background and symbols are often better integrated into a single coherent artistic product. Here, our use of custom-styled tiles, such as those of Gröbe et al. (2020) in Fig. 6, can be seen as a first step in improving fusion. Further work is needed, with specific aggregation routines that help generalize the map background or move details into the generated icons. We tested a posterior image-to-image step to better fuse the final map shown in Fig. 6 with the background, but the result was unsuitable due to distortions (see S3). Here, stable diffusion extensions such as ControlNet<sup>10</sup> (Zhang and Agrawala 2023) or tile-based processing<sup>11</sup> could be used in the future to incrementally construct a map with smooth transitions between different image representations for different areas.

<sup>9</sup> <https://github.com/aymenfurter/microagents>.

<sup>10</sup> <https://github.com/llyasviel/ControlNet>.

<sup>11</sup> <https://github.com/Coyote-A/ultimate-upscale-for-automatic1111>.

How to preserve topographic order in such an approach is an open question.

Our work shows an increasing reliance on user-contributed data, such as publicly shared geosocial media data or crowdsourced images, to train LDMs. We do not see this as a drawback. Maps that are automatically updated based on user input can be a fundamental tool for incorporating the collective attribution of meaning and importance by the population into map generation. They can also be used to motivate community input and enable participation in city and neighborhood planning. Community efforts can also be directed toward increasing the specificity of models through the use of embeddings (Ruiz et al. 2022) based on local image training data, which are better suited to replicate the unique appearance of certain landmarks, neighborhood styles, or contexts known only to locals. Our work also highlights the challenges of generative AI as a “black box” that is difficult to control without decomposing individual steps. By fully sharing the code and data for this study (Dunkel et al. 2023), we hope to encourage further research. Despite impressive practical progress, the theory and systematization of generative AI research has lagged behind development. Several parts of our workflow rely on publicly shared content whose exact creation is unknown. Here, reproducibility, transparency, and soundness are critical areas that require further contributions from the research community.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s42489-024-00159-9>.

**Acknowledgements** This work was supported by the German Research Foundation as part of the priority programme “Volunteered Geographic Information: Interpretation, Visualisation and Social Computing” (VGIScience, priority programme 1894). The authors gratefully acknowledge the support of the Gemeinsame Wissenschaftskonferenz (GWK) for funding this project by providing computing time through the Center for Information Services and HPC (ZIH) at TU Dresden. A total of 139 graphics processing unit (GPU) hours were used for this research.

**Author Contributions** Conceptualization AD, DB and M.G.; Methodology, AD and MG; Software (visualization), AD; Validation, AD and DB; Formal analysis, AD and MG; Investigation, AD and MG; Resources, DB; Data curation, AD; Writing—original draft preparation, AD and MG; Writing—review and editing, DB and MG; Visualization, AD; Supervision, DB; Project administration, DB; Funding acquisition, DB and AD. All authors have read and approved the published version of the manuscript.

**Funding** Open Access funding enabled and organized by Projekt DEAL.

**Data availability** All data used to produce figures and results in this work (see code Supporting information S1–S3) is made available in a public data repository <https://doi.org/10.25532/OPARA-253>.

## Declarations

**Conflict of interest** The authors declare no conflict of interest.

**Ethical statement** Data were retrieved using the public Twitter, Flickr, Instagram (Netlytic) and iNaturalist APIs and generalized to anonymized data. No personally identifiable information (PII) is required to reproduce this study using the material shared in Dunkel et al. (2023).

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Ai T (2022) Some thoughts on deep learning empowering cartography. *J Geogr Cartogr* 5(2):25. <https://doi.org/10.24294/jgc.v5i2.1670>
- Antoniou A, Ehmann S, Klanten R (eds) (2015) *Mind the map: illustrated maps and cartography*. Die Gestalten Verlag, Berlin
- Bogucka EP, Meng L (2019) Projecting emotions from artworks to maps using neural style transfer. *Proc ICA* 2:1–8. <https://doi.org/10.5194/ica-proc-2-9-2019>
- Brown A, Emmer N, Van Den Worm J (2001) Cartographic design and production in the internet era: the example of tourist web maps. *Cartogr J* 38(1):61–72. <https://doi.org/10.1179/caj.2001.38.1.61>
- Cañas AJ, Carff R, Hill G, Carvalho M, Arguedas M, Eskridge TC, Lott J, Carvajal R (2005) Concept maps: integrating knowledge and information visualization. In: Tergan S-O, Keller T (eds) *Knowledge and information visualization: searching for synergies*. Springer, Berlin, pp 205–219. [https://doi.org/10.1007/11510154\\_11](https://doi.org/10.1007/11510154_11)
- Chen X, Chen S, Xu T, Yin B, Peng J, Mei X, Li H (2021) SMAP-GAN: generative adversarial network-based semisupervised styled map tile generation method. *IEEE Trans Geosci Remote Sens* 59(5):4388–4406. <https://doi.org/10.1109/TGRS.2020.3021819>
- Chen X, Yin B, Chen S, Li H, Xu T (2022) Generating multiscale maps from satellite images via series generative adversarial networks. *IEEE Geosci Remote Sens Lett* 19:1–5. <https://doi.org/10.1109/LGRS.2021.3129285>
- Child H (1956) *Decorative maps, the ‘how to do it’ series*. Studio Publications, London and New York
- Christophe S, Mermet S, Laurent M, Touya G (2022) Neural map style transfer exploration with GANs. *Int J Cartogr* 8(1):18–36. <https://doi.org/10.1080/23729333.2022.2031554>
- Dehouche N, Dehouche K (2023) What’s in a text-to-image prompt? The potential of stable diffusion in visual arts education. *Heliyon* 9(6):e16757. <https://doi.org/10.1016/j.heliyon.2023.e16757>
- Dunkel A (2015) Visualizing the perceived environment using crowdsourced photo geodata. *Landsc Urban Plan* 142:173–186. <https://doi.org/10.1016/j.landurbplan.2015.02.022>
- Dunkel A, Löchner M, Burghardt D (2020) Privacy-aware visualization of volunteered geographic information (VGI) to analyze spatial activity: a benchmark implementation. *ISPRS Int J Geo Inf* 9(10):607. <https://doi.org/10.3390/ijgi9100607>
- Dunkel A, Burghardt D, Gugulica M (2023) Supplementary materials for the publication. Generative text-to-image diffusion for automated map production based on geosocial media data. <https://doi.org/10.25532/OPARA-253>

- Epstein Z, Hertzmann A, Herman L, Mahari R, Frank MR, Groh M, Schroeder H, Smith A, Akten M, Fjeld J, Farid H, Leach N, Pentland A, Russakovsky O (2023) Art and the science of generative AI: a deeper dive. *Science* 380(6650):1110–1111. <https://doi.org/10.1126/science.adh4451>
- Gal R, Alaluf Y, Atzmon Y, Patashnik O, Bermano AH, Chechik G, Cohen-Or D (2022) An image is worth one word: personalizing text-to-image generation using textual inversion. arXiv: <http://arxiv.org/abs/2208.01618>
- García-Peñalvo F, Vázquez-Ingelmo A (2023) What do we mean by GenAI? A systematic mapping of the evolution, trends, and techniques involved in generative AI. *Int J Interact Multimedia Artif Intell*. <https://doi.org/10.9781/ijimai.2023.07.006>. (in press)
- Gröbe M, Dunkel A, Burghardt D (2020) A new web map for the MeinGrün app. *Abstr ICA 2:1–2*. <https://doi.org/10.5194/ica-abs-2-8-2020>
- Gugulica M, Burghardt D (2023) Mapping indicators of cultural ecosystem services use in urban green spaces based on text classification of geosocial media data. *Ecosyst Serv* 60:101508. <https://doi.org/10.1016/j.ecoser.2022.101508>
- Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds) *Advances in neural information processing systems*, vol 33. pp 6840–6851
- Holloway L, Marriott K, Butler M, Reinders S (2019) 3D printed maps and icons for inclusion: testing in the wild by people who are blind or have low vision. In: *The 21st international ACM SIGACCESS conference on computers and accessibility*, pp 183–195. <https://doi.org/10.1145/3308561.3353790>
- Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W (2021) LoRA: low-rank adaptation of large language models. arXiv <http://arxiv.org/abs/2106.09685>
- Huang H, Gartner G (2012) A technical survey on decluttering of icons in online map-based mashups. In: Peterson MP (ed) *Online maps with APIs and WEBSERVICES*. Springer, Berlin, Heidelberg, pp 157–175. [https://doi.org/10.1007/978-3-642-27485-5\\_11](https://doi.org/10.1007/978-3-642-27485-5_11)
- Jin Z, B Liu, Q Chu, N Yu (2021) ISNet: integrate image-level and semantic-level context for semantic segmentation. In: *2021 IEEE/CVF international conference on computer vision (ICCV)*. IEEE, Montreal, QC, Canada. pp 7169–78. <https://doi.org/10.1109/ICCV48922.2021.00710>
- Juhász L, Mooney P, Hochmair HH, Guan B (2023) ChatGPT as a mapping assistant: a novel method to enrich maps with generative AI and content derived from street-level photographs. In: *Spatial data science symposium, 2023*. <https://doi.org/10.25436/E2ZW27>
- Kang Y, Gao S, Roth RE (2019) Transferring multiscale map styles using generative adversarial networks. *Int J Cartogr* 5(2–3):115–141. <https://doi.org/10.1080/23729333.2019.1615729>
- Kang Y, Zhang Q, Roth R (2023) The ethics of AI-generated maps: a study of DALLE 2 and implications for cartography. arXiv <http://arxiv.org/abs/2304.10743>
- Karras T, Aittala M, Aila T, Laine S (2022) Elucidating the design space of diffusion-based generative models. arXiv <http://arxiv.org/abs/2206.00364>
- Kingma DP, Welling M (2022) Auto-encoding variational bayes. arXiv <http://arxiv.org/abs/1312.6114>
- Laney D (2001) 3D data management: controlling data volume, velocity, and variety. META Group
- Lin C-H, Chen J-Y, Hsu S-S, Chung Y-H (2014) Automatic tourist attraction and representative icon determination for tourist map generation. *Inf vis* 13(1):18–28. <https://doi.org/10.1177/1473871612472177>
- Lu C, Zhou Y, Bao F, Chen J, Li C, Zhu J (2022) DPM-solver: a fast ODE solver for diffusion probabilistic model sampling in around 10 steps. arXiv <http://arxiv.org/abs/2206.00927>
- McInnes L, Healy J, Astels S (2017) HDBSCAN: hierarchical density based clustering. *J Open Source Softw* 2(11):205. <https://doi.org/10.21105/joss.00205>
- Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. <http://arxiv.org/abs/1301.3781>
- Nori H, Lee YT, Zhang S, Carignan D, Edgar R, Fusi N, King N, Larson J, Li Y, Liu W, Luo R, McKinney SM, Ness RO, Poon H, Qin T, Usuyama N, White C, Horvitz E (2023) Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. arXiv <http://arxiv.org/abs/2311.16452>
- Oppenlaender J (2023) A taxonomy of prompt modifiers for text-to-image generation. arXiv <http://arxiv.org/abs/2204.13988>
- Qin X, Dai H, Hu X, Fan D-P, Shao L, Van Gool L (2022) Highly accurate dichotomous image segmentation. arXiv <http://arxiv.org/abs/2203.03041>
- Roberts AJ (2012) Modify the Improved Euler scheme to integrate stochastic differential equations. arXiv <http://arxiv.org/abs/1210.0933>
- Robinson AC, Demšar U, Moore AB, Buckley A, Jiang B, Field K, Kraak M-J, Camboim SP, Sluter CR (2017) Geospatial big data and cartography: research challenges and opportunities for making maps that matter. *Int J Cartogr* 3(sup1):32–60. <https://doi.org/10.1080/23729333.2016.1278151>
- Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B (2022) High-resolution image synthesis with latent diffusion models. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp 10684–10695
- Ruiz N, Li Y, Jampani V, Pritch Y, Rubinstein M, Aberman K (2022) DreamBooth: fine tuning text-to-image diffusion models for subject-driven generation. ArXiv Preprint [arXiv:2208.12242](http://arxiv.org/abs/2208.12242)
- Schettering V, Bartolomeo SD, El-Assady M, McNutt A, Miller M, Passos JPA, Adams JL (2023) Doom or deliciousness. challenges and opportunities for visualization in the age of generative models. <https://doi.org/10.31219/osf.io/3jr3m>
- Schuhmann C, Beaumont R, Vencu R, Gordon C, Wightman R, Cherti M, Coombes T, Katta A, Mullis C, Wortsman M, Schramowski P, Kundurthy S, Crowson K, Schmidt L, Kaczmarczyk R, Jitsev J (2022). LAION-5B: an open large-scale dataset for training next generation image-text models. arXiv <http://arxiv.org/abs/2210.08402>
- Sohl-Dickstein J, Weiss E, Maheswaranathan N, Ganguli S (2015) Deep unsupervised learning using nonequilibrium thermodynamics. In: Bach F, Blei D (eds) *PMLR*, vol 37. pp 2256–2265
- Song J, Meng C, Ermon S (2022) Denoising diffusion implicit models. arXiv <http://arxiv.org/abs/2010.02502>
- Tsai Y-L, Hsu C-Y, Xie C, Lin C-H, Chen J-Y, Li B, Chen P-Y, Yu C-M, Huang C-Y (2023) Ring-a-bell! How reliable are concept removal methods for diffusion models? arXiv <http://arxiv.org/abs/2310.10012>
- Witteveen S, Andrews M (2022) Investigating prompt engineering in diffusion models. arXiv <http://arxiv.org/abs/2211.15462>
- Wood D, Fels J (1986) Designs on signs/myth and meaning in maps. *Cartographica Int J Geogr Inf Geovisual* 23(3):54–103
- Zhang L, Agrawala M (2023) Adding conditional control to text-to-image diffusion models. arXiv <http://arxiv.org/abs/2302.05543>
- Zhang C, Zhang C, Zhang M, Kweon IS (2023) Text-to-image diffusion models in generative AI: a survey. arXiv <http://arxiv.org/abs/2303.07909>
- Zhao B, Zhang S, Xu C, Sun Y, Deng C (2021) Deep fake geography? When geospatial data encounter Artificial Intelligence. *Cartogr Geogr Inf Sci* 48(4):338–352. <https://doi.org/10.1080/15230406.2021.1910075>
- Zhao W, Rao Y, Liu Z, Liu B, Zhou J, Lu J (2023) Unleashing text-to-image diffusion models for visual perception. arXiv <http://arxiv.org/abs/2303.02153>