TopoGen: Training Generative AI to Produce Maps for Experiential Scenarios

Sophia Khan, Joel Walsh, & Benjamin Nye

University of Southern California Institute for Creative Technologies
Los Angeles, CA
skhan1@wildcat.fvsu.edu; jwalsh@ict.usc.edu; nye@ict.usc.edu

ABSTRACT

Military education programs emphasize scenario-based training, such as situational judgements, training simulations, and wargames.. Due to the importance of spatial relationships and movement for courses of action, maps must be either created or found that stimulate specific learning objectives (e.g., "a wet gap crossing approaching a large city"). Manually development of specialized maps is time-consuming, often contributing to the long development time for high-quality training scenarios. This work details the creation of TopoGen, a two-part dataset used to train generative models for this purpose. Popular diffusion models are trained on an internet-scale amount of (image, description) pairs and have emerged as a promising method for near-instantaneous image generation from natural language text descriptions.Unfortunately, most popular traditional generative models cannot efficiently perform domain-specific tasks like creating high quality maps for military educational purposes. To remedy this weakness we use an off-the-shelf image captioning model, along with maps taken from the internet, to create a bank of 110 (image, description) pairs needed to fine-tune the diffusion model. TopoGen also includes 7,867 (description, generated_image, bounding box) triplets for finetuning and prompting multimodal models, in order to generate better text test scenarios. Tests indicate that training methods like Google's Dreambooth can produce convincing maps based on prompts just these 110 labeled examples. The resultant finetuned diffusion model can create images using simple prompts describing geographic features that include the size of the map, relative locations of topography (e.g., hills, mountains), and water features (e.g., rivers, coasts). Once the images are generated, open vocabulary object detection models determine the location of these features using bounding boxes, to inform systems of where the specific key map features are located. TopoGen maps are relevant to a variety of training tasks, including practice items (e.g., route planning) and wargaming scenario documents. Future work will investigate empirical validation with subject-matter experts, extensions to generate terrain for interactive simulations, and comparisons with complementary generative pipelines (e.g., 3D-to-2D translation methods).

ABOUT THE AUTHORS

Sophia Khan received her Bachelor of Science degree in Computer Science from Fort Valley State University in December 2024. She is currently engaged in Android application development, contributing to innovative software projects that apply mobile technology to real-world problems. Her interests include mobile development, user interface design, and emerging technologies in computing.

Joel Walsh, Ph.D. is a postdoctoral fellow at the USC Institute for Creative Technologies (ICT). Dr. Walsh's research involves leveraging multimodal generative models for teaching and learning.

Benjamin Nye, Ph.D. is Director of Learning Science at the USC Institute for Creative Technologies (ICT). Dr. Nye's research has been recognized for excellence in intelligent tutoring systems, cognitive agents, realistic behavior in training simulations, and machine learning for adaptive learning. Nye's research is on scalable learning technologies and design principles that promote learning. His recent work emphasizes AI tools for instructors and content developers to use AI tools that enable them to rapidly update content and to create AI-enabled learning experiences (i.e., AI-human teams to generate AI tutoring)

INTRODUCTION

Military and emergency training scenarios often rely on detailed maps to convey key spatial relationships and support decision-making tasks. Creating or searching for these specialized maps manually is time-consuming and slows down the development of high-quality training materials. A fully automated pipeline capable of generating topographical maps from textual descriptions would significantly accelerate scenario creation, increase scalability, and enable generating scenarios which are aligned to training specific competencies and research objectives (Hans et al., forthcoming). Recent advances in generative AI, particularly diffusion-based image generation models, offer promise for producing maps from semantic building blocks like “roads”, or “river”. These models, trained on images and description data, can generate map images tailored to specific training needs. In addition, we adapt the widely used practice in adding bounding box coordinates as context to Large Language Models, which enables multimodal models to better understand and describe map features. This bundling of artifacts can enhance test items, and enable more flexible, text-guided development of training scenarios.

RELATED WORK

The pipeline used to create this dataset leverages a diverse set of existing research in multimodal computer vision, whereby text representations of images are used in a number of applications.

Text-to-image models allow a user to specify an image description in natural language, and then generate potential images that have this description. Modern text-to-image generation models were preceded by foundational work on Helmholtz machines (Dayan et al, 1995) and later Auto Encoding Variational Bayes (Kingma and Welling, 2013). Early attempts at image generation did not have a text-to-image component; they simply used images to build a distribution and then sample new images from this distribution. In the mid 2010s, researchers began to have limited success at generating images from natural language descriptions utilizing soft attention mechanisms (Mansimov et al, 2015) and Generative Adversarial Networks (Reed et al., 2016). In recent years, diffusion models have proven to be superior in terms of training ease and achieving text-to-image correspondence (Nichols et al., 2021). Diffusion models can also be finetuned with relatively small amounts of data to generate specific images subjects, such as Google’s Dreambooth model (Ruiz et al., 2022).

In the years since diffusion models have become publicly available, text-to-image generative models have spread into all manner of consumer devices through tools like Dall-E (Ramesh et al., 2021) and Google Gemini (Anil et al., 2023). Since the adoption of these models, researchers have done limited work on using diffusion technology for generating maps or map elements. This includes generating map symbols from textual descriptions (Drews et al., 2024) and converting existing maps into a symbol-laden map (Dunkel et al., 2024). Diffusion models have also been used to generate map tiles from real aerial imagery (Sun et al, 2025).

The last part of our data pipeline is an object detection model that relies on the ability of vision transformers systems (Dosovitskiy, 2021) to form joint text and image representations. These open-set object detection models allow a user to specify an object name as text, and to receive bounding box coordinates for all instances of this object within an image. They also typically mark unknown objects for which there is no label. This is a no-shot process, meaning that it requires no labeled data. One of the most cited and implemented models is Grounding DINO (S. Liu et al., 2023). These bounding boxes allow Large Language Models to access deeper meaning and generative capabilities from images, a process known as visual instruction tuning (ViT; H. Liu et al., 2023).

We leverage these foundational approaches of diffusion, open-set object detection, and visual instruction in order to create a robust dataset of captioned map tiles, with bounding box labels for relevant features like rivers or mountains. We make the novel contribution of showing that this data is sufficiently large and fine-grained enough to train diffusion models to generate convincing images of maps, which can then be packaged with open-set object detection data and captions. This multimodal package can then serve as context for generating detailed and internally consistent assessments for military training scenarios.

APPROACH

The goal of TopoGen is to develop a system capable of:

- *Text-to-Map Output*: Generating maps from textual descriptions that capture key geographical features.

- *Terrain Feature Boxes*: Extracting bounding boxes to identify geographical features within generated maps.
- *Text-to-Map Dataset*: Constructing a dataset with 7,868 map tiles paired with captions to support future multimodal learning.
- *TopoMap Synthetic Map Corpus*: Creating a corpus of maps that can be used to train future models, ensuring adaptability for various applications.

Importance: This approach supports the creation of a domain-specific dataset where test descriptions are directly tied to map images. By including labeled features and spatial annotations, the dataset is well-suited for training models that understand both visual content and geographic structure.

Model: This approach is model-agnostic, ensuring compatibility with a range of machine learning models. While DreamBooth was selected for generative modeling and OpenVocab Dino (Wang at al., 2024) for object detection due to their demonstrated efficacy, alternative models can be employed based on application specific requirements. The flexibility of this methodology allows for adaptation to diverse domains, including military training.

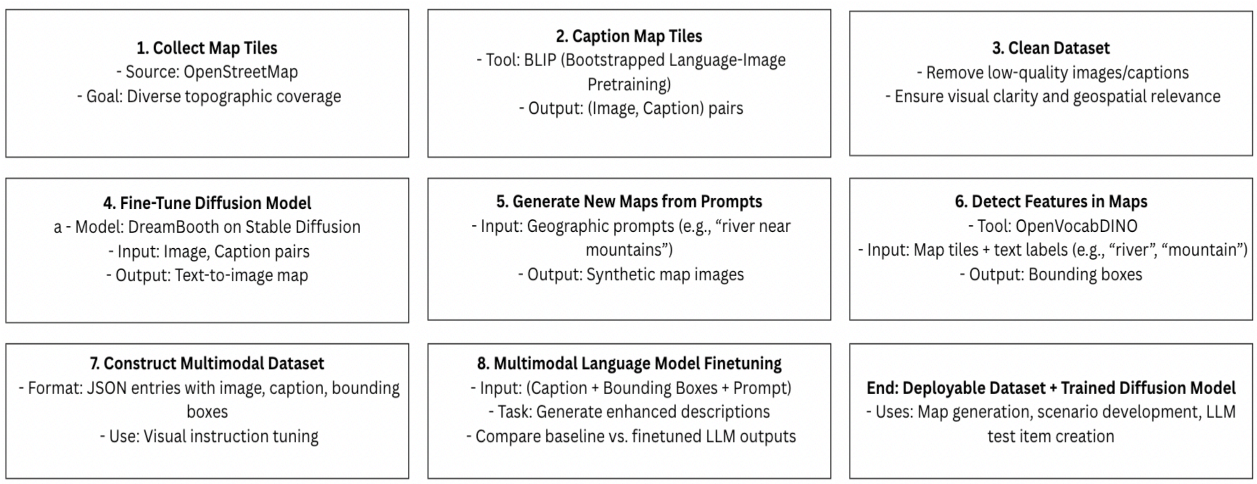


Figure 1. TopoGen pipeline from start (data collection) to end (trained model and dataset)

Defining Map Features

To ensure effective model training, we defined a vocabulary of features to help guide both the generation of map descriptions and the labeling of visual data. It was essential to include the main geographic elements most relevant to map-based scenarios.The table below lists examples of the key features selected. Many features can be specified on a map, with an associated bounding box.

Table 1. Map Feature Set and Descriptions

Feature	**Description**
Mountain	Elevated landforms with steep slopes or peaks.
Forest	Densely wooded areas, typically with high tree coverage.
Lake	Inland water bodies, usually static and surrounded by land.
Desert	Dry, barren regions with minimal vegetation.

Ocean	Large saltwater bodies at the edges of landmasses.
Road	Pathways for vehicular travel, varying in size and type.
Grass	Flat or rolling terrain covered in grass or low vegetation.
Sand	Regions with sandy surfaces, such as beaches or deserts.
Reservoir	Man-made water storage areas, often behind dams.
Canal	Artificial waterways for transport or irrigation.
Waterbody	General term for visible water features.
Contour	Lines representing elevation changes and terrain shape.
Terrain	The surface structure of the land, including slopes, plains, and elevations.
Tree	Individual or grouped tree symbols representing wooded areas.
Water	Any visible representation of water, including rivers, lakes, and oceans.
Topographic	Features related to elevation and landform, often via contour lines.
Region	General geographic zones with shared features.
Elevation	Height above sea level, often implied through shading or markers.

Map Corpus Development

Before assembling our dataset, we investigated multiple online sources for map imagery, including ArcGIS. Although ArcGIS provides access to detailed geographic data, many of its map layers featured satellite imagery or stylizations that did not align with the clean topography style required for our generation task. These formats introduced visual noise and lacked the simple topographic features we aimed to recreate.

After exploring a number of different potential data sources, we selected OpenStreetMap as the basis for our dataset. OpenStreetMap (OSM) is a collaborative, open-source mapping platform that provides vector-based geographic data (OpenStreetMap contributors, 2017). Unlike satellite based maps, OSM tiles emphasize topography, land use and other symbolic features in a format well-suited for training a model. Its global coverage and consistent visual conventions made it an ideal source for collecting diverse and clearly structured map tiles.

The initial dataset was sourced from OpenStreetView, which visualizes OpenStreetMap data into map tiles. OpenStreetMap provides an open, editable geographic database that covers regions worldwide. To ensure diversity and represent various topographic features, we selected map tiles from multiple regions around the globe. This approach allowed us to incorporate a wide range of geographic elements including mountains, rivers, forests, deserts, etc., ensuring the dataset reflected the variety of topographies found in real-world maps.

Textual Descriptions for each map tile were generated using BLIP (Bootstrapped Language Image Pretraining). BLIP is a vision-language pre-training framework designed to improve both understanding and generation tasks, such as image captioning and image-text retrieval. It works by using a bootstrapping method, where a model generates captions for images and a filter removes poor-quality ones. In our project, BLIP was used to generate captions for the map tiles in our dataset. This helped create image-text pairs that capture both the visual features and context of the maps, which is key for training models that understand and generate maps effectively.

While BLIP is a strong model for generating captions, we still had to undergo a data cleaning process. We filtered out low-quality captions, duplicate images, or maps lacking identifiable geographic features. This process ensured that only high-quality, meaningful examples remained in the dataset, reinforcing the model's ability to learn from diverse and accurate map data. This produced the Text-to-Map Dataset which was used in later steps.

Image Diffusion Model: Fine-Tuning

Before selecting DreamBooth, we explored several diffusion-based approaches to generate accurate map images. Our initial experiments included Stable Diffusion, a latent text-to-image diffusion model that generates high-quality images through a compressed latent space. While stable diffusion performed well on general image generation tasks, its pretrained weights were not adapted for the map-based domain. As a result, the outputs displayed hallucinated map features, inaccurate layouts, and distorted or overly simplified visual elements.

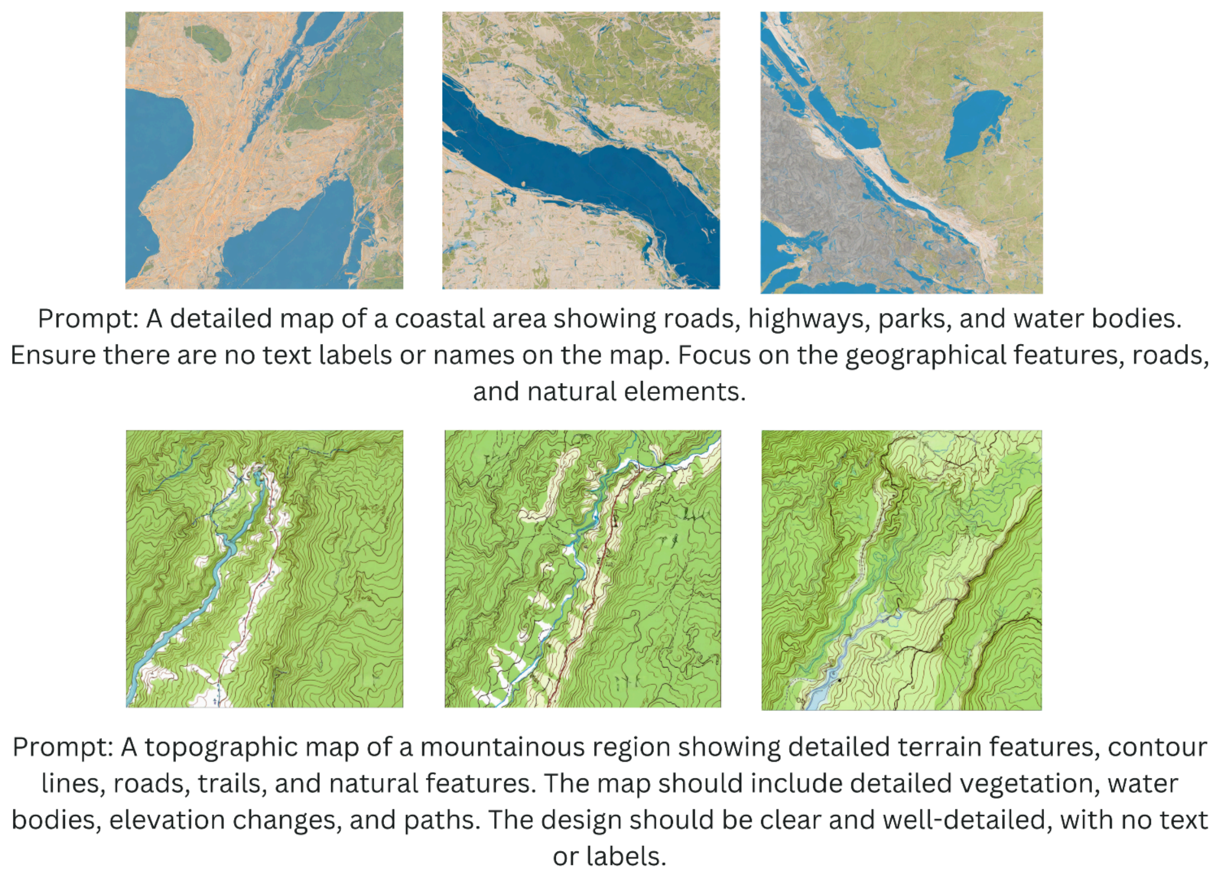


Figure 2: Examples of maps generated using DreamBooth and the TopoGen dataset

We also tested Imagen (Saharia et al., 2022), a text-to-image diffusion model developed by Google that emphasizes photorealism. This model seemed promising because it uses an approach similar to AI-upscaling, where a small image tile is generated, then additional steps expand the image size and generate more detail. This approach seemed appropriate for map generation. Imagen demonstrated impressive results on natural imagery tasks. However, when applied to our domain, Imagen struggled to capture the characteristics of a map. The generated outputs frequently contained blended features or visual clutter which misaligned with the structured forms commonly seen in topographic maps.

After extensive comparisons, we selected DreamBooth, a fine-tuning technique for Stable Diffusion that allows the model to learn new visual concepts with example images (Ruiz et al., 2022). A key strength of DreamBooth is its ability to adapt to specific visual domains. This is crucial because maps require a high level of spatial accuracy, meaning the model must place geographic features like mountains, rivers, trees, etc. in the correct relative positions and proportions, something that typical image generation models struggle to do without fine-tuning.

Dreambooth works well because it adds a rare, unique token during fine-tuning that links the training images to the new concept. This helps the model learn the visual traits without confusing them with existing ones. As a result, it avoids issues like overfitting, where the model memorizes the training images instead of learning general patterns. It also prevents mode collapse, where the model produces nearly identical outputs every time. This approach produces clear, consistent images that preserve the geographic accuracy and style of real map tiles (see Fig. 2).

Training the DreamBooth model was performed on a single NVIDIA Tesla A40 GPU with 16GB RAM. Comparable results could be achieved with modern consumer-grade GPUs, making this approach significantly less resource-intensive than fine-tuning large multimodal LLMs, which often require multiple high memory accelerators. Beyond efficiency, the pipeline is fully open source and can be trained on-premise. This enables organizations to fine-tune models on proprietary terrain data without exposure to external cloud services, a key consideration for scenarios requiring security or data sovereignty.

Bounding Box Detection

To extract structured visual information from each map tile, OpenVocabDINO (Wang at al., 2024) was used as the object detection framework. OpenVocabDINO is an open-vocabulary object detector built on top of the DINO architecture, that allows detection beyond a fixed set of objects. Unlike traditional detectors limited to a fixed set of classes, OpenVocabDINO can detect a wide range of objects based on text prompts, such as category names or descriptions. This makes it well-suited for recognizing various geographic features that may not belong to a predefined label set.

The model works by fusing vision and language through a multi-phase architecture that includes:

- A feature enhancer - employs mechanisms like bidirectional cross-attention to dynamically improve the text embedding's representation for better modality alignment.
- Language-guided query selection - selects the most relevant object embeddings by assessing the similarity. between the image features and the text embedding, which helps initialize and guide the decoder
- A cross-modality decoder - takes static learnable content queries and merges them with the dynamically selected, text-related object embeddings to produce the final classification and box regression predictions.

This embedding architecture allows the model to locate features in an image without prior exposure to them in the training data. OpenVocabDINO has shown strong performance across multiple benchmarks, including zero-shot detection tasks (i.e., detection with no labeled examples provided).

OpenVocabDINO was applied to detect and label key geographic features within each map tile. For every image, the model generated:

- *boxes*: spatial coordinates indicating the position and size of detected features.
- *labels*: map features such as “map,” “trail,” “river,” etc.
- *scores*: confidence values reflecting the model’s certainty in each detection.

Visual Instruction Tuning Data Set

To enable multimodal learning, each map tile was paired with descriptive captions and bounding box annotations, then compiled into a structured JSON format. This organization supports visual instruction tuning by aligning textual and spatial information. Each entry in the dataset included (Fig. 3):

- The map tile
- A caption (or multiple captions) generated by BLIP. describing the geographic features
- Bounding boxes with spatial coordinates and corresponding feature labels (river, terrain, etc.)

Caption: "A map of the Amazon River in Brazil showing topographical features, forested areas, various river branches, and surrounding terrain."

Bounding Boxes:

{
 "topographic": [0.2, 0.36, 511.97, 511.63],
 "waterbody": [0.14, 206.53, 250.43, 511.73],
 "grass": [0.16, 207.02, 249.75, 511.65]
}

Formatted Label:

"A map of the Amazon River in Brazil showing topographical features, forested areas, various river branches, and surrounding terrain.",
{"topographic": [0.2, 0.36, 511.97, 511.63], "waterbody": [0.14, 206.53, 250.43, 511.73], "grass": [0.16, 207.02, 249.75, 511.65]}

Figure 3: Example of Topogen dataset entry

This format helps models connect descriptive language with specific regions in the image, making it easier to follow instructions or identify features based on text input. By linking captions to bounding boxes and visual elements, the dataset supports more accurate spatial reasoning.

When an image and bounding box coordinates are added to a prompt as context, it allows for training scenario generation to impose a grid structure that would otherwise be impossible for any multimodal model. See Appendix A for an example of an-AI-generated map reading and land navigation test based on the data from Figure 3. This presents a set of navigation questions and a complete answer key.

DISCUSSION

This work builds on recent advances in multimodal learning, particularly in using diffusion models and open-vocabulary object detection for grounded reasoning tasks. While traditional map generation relies on manual design or GIS platforms, our approach leverages generative models to automate the process using natural language descriptions. Prior research in image generation focused on photorealism, with limited attention to structured geospatial content. By combining diffusion models (ex: DreamBooth) , captioning (ex. BLIP), and open vocabulary object detection (ex: OpenVocabDINO), we extend this field toward map-specific scenarios, where spatial accuracy and feature control are critical.

The TopoGen pipeline (Figure 1) creates artifacts that can then be used to generate customized training scenarios without relying on handcrafted maps. Military land navigation and emergency planning assessments often require maps tailored to specific conditions. For example, "a river separating two forested regions near a mountain." Our method enables scenario designers to create these maps quickly from textual descriptions, reducing the time needed for scenario prototyping.

The TopoGen dataset also advances the development of synthetic multimodal datasets by providing aligned sets of map tiles, map descriptions, and bounding box annotations. These examples are designed to support both assessment generation and potentially the training of multimodal models that incorporate finer-grained context. Unlike general-purpose datasets, TopoGen focuses specifically on geographic content, making it well-suited for applications such as terrain analysis and scenario generation for training environments.

The generative pipeline successfully produced maps in the desired topographic style, capturing features from minimal prompt descriptions. The fine tuned diffusion model was able to generate realistic terrain patterns, including rivers, elevation changes, etc. in a way suitable for scenario training and planning tasks. However, the addition of text labels to the generated maps proved to be challenging. When prompts included instructions to place text on the maps, the DreamBooth-generated outputs often hallucinated illegible or nonsensical text. In many cases, the model produced shapes that looked like letters but did not form actual words. This highlights a limitation of current diffusion image generation models; combining both structured spatial layouts and textual annotations in a single output is somewhat error-prone due to the architecture of the models (Zhang at al., 2025). While they can capture spatial structure and geographic features effectively, combining that with reliable, embedded textual elements in a single image is still a difficult task. However, this can be reasonably addressed by approaches such as generating bounding boxes and assigning random names to the features, i.e. affixing the text, “Mount Baldy” on the center a bounding box containing a mountain, or “River Tiber” on the center of bounding box containing a river. Ultimately, this approach is typically used for mapping software for maps, as it allows for translation of labels or showing/hiding labels when zooming in or out.

While TopoGen demonstrates the feasibility of generating map tiles aligned with textual prompts, several limitations must be acknowledged. First, the geographic accuracy of the generated maps is sufficient for general scenario visualization but not yet suitable for all advanced land navigation tasks. For example, contour lines occasionally merge, which hides the true steepness of the terrain. When this happens, it becomes hard to tell whether an area represents a passable slope or a dangerous cliff, which could make the maps less reliable for planning precision platoon-level routes. To address these uncertainties, we plan to involve land navigation subject-matter experts (SMEs) in future evaluations. Their assessments will help identify potential limitations in how well the generated maps align with instructional requirements for map-dependent training.

Additionally, while the feature vocabulary covers a broad set of map elements, some bounding box labels remain imprecise. For instance, the label “waterbody” groups together aquatic features, whereas in training contexts it would be useful to distinguish these more specifically. Similarly, terms like “topographic” require clearer definition to match common terminology in map-based instruction. Future dataset versions will refine these labels and provide greater specificity.

Finally, the current work focuses on direct 2D generation of map tiles from text prompts. An alternative approach would be to first generate a physics-informed 3D terrain model and the systematically convert it into a 2D topographic map. Such a pipeline might yield more coherent contour lines but would require greater computational resources. The choice of a lightweight 2D generation pipeline reflects a balance between computational efficiency and accessibility for researchers.

All data, including map tiles, captions, and bounding box annotations will be available to the public. The codebase supporting the data processing pipeline, model training and evaluation is available on Github enabling reproducibility for future research or applications (https://github.com/ICTLearningSciences/TopoMapScenGen).

CONCLUSIONS AND FUTURE DIRECTIONS

The process of combining real-world map tiles from OpenStreetView with automatically generated captions and bounding box annotations proved effective for creating a multimodal dataset focused on geographic features. The dataset effectively captures spatial relationships between important map elements such as terrain and water bodies.

Synthetic data, like the kind created in our project, helps models better understand and connect what they see on a map with what is described in text. This is especially helpful for AI models that combine images and language, which learn by studying lots of examples of images paired with detailed captions. Because synthetic maps include clear labels and precise locations for features such as rivers, forests, and mountains, the models learn how to recognize these features and understand how they relate to words. This kind of data is also useful for object detection models like OpenVocabDINO, which can detect objects of interest based on text instead of needing a fixed list of categories. By using synthetic data, these models get better at handling unusual map features and can be used for creating training scenarios.

Additionally, the integration of OpenStreetView's rich metadata layers could be used for further development. By incorporating contextual metadata, such as movement paths or tactical annotations tied to military course of action, it may be possible to generate high resolution maps that reflect complex spatial dynamics, This added realism could significantly enhance the relevance and instructional value of generated maps for simulation, planning, and decision-making tasks.

TopoGen lays the groundwork for scalable, machine-generated geospatial content that can be used to accelerate training scenario development and expand the capabilities of multimodal AI systems in understanding terrain-based data. In addition to the current process, this research might also be leveraged to generate other types of scenario topography, such as 3D terrain for virtual world or augmented reality training exercises (Boyce et al., 2016).

ACKNOWLEDGEMENTS

The project or effort depicted was or is sponsored by the U.S. Government under contract number W911NF-14-D-0005, as part of the USC ICT UARC and the AI Research Center of Excellence in Education (AIRCOEE). The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. We would also like to thank the National Defense Education Program (NDEP) and LMI for their support of the USC SMART Futures Contest program, which supported training and support for students engaged in this project.

REFERENCES

Boyce, M. W., Reyes, R. J., Cruz, D., Amburn, C. R., Goldberg, B., Moss, J. D., & Sottilare, R. A. (2016). *Effect of topography on learning military tactics–integration of generalized intelligent framework for tutoring (GIFT) and augmented reality sandtable (ARES)*. Army Research Laboratory.

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The Helmholtz machine. Neural Computation, 7(5), 889–904. https://doi.org/10.1162/neco.1995.7.5.889

Drews, J., Weißmann, M., Keil, J., Dickmann, F., & Edler, D. (2025). A New AI Tool for the Design of Cartographic Pictograms (PictoAI) and Its Potentials for Increasing Their Meaningfulness. KN - Journal of Cartography and Geographic Information, 75(1), 23–35. https://doi.org/10.1007/s42489-024-00183-9

Dunkel, A., Burghardt, D., & Gugulica, M. (2024). Generative Text-to-Image Diffusion for Automated Map Production Based on Geosocial Media Data. KN - Journal of Cartography and Geographic Information, 74(1), 3–15. https://doi.org/10.1007/s42489-024-00159-9

Gregor, K., Danihelka, I., Graves, A., Rezende, D. J., & Wierstra, D. (2015). DRAW: A Recurrent Neural Network For Image Generation (arXiv:1502.04623). arXiv. https://doi.org/10.48550/arXiv.1502.04623

Hans, S., Ustun, V., Core, M., Nye, B., Sterrett, J., & Green, M. (forthcoming). Towards AI-Assisted Generation of Military Training Scenarios. Submitted to *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) 2025.*

Li, J., Li, D., Xiong, C., & Hoi, S. (2022, February 15). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. arXiv.org. https://arxiv.org/abs/2201.12086

Mansimov, E., Parisotto, E., Ba, J. L., & Salakhutdinov, R. (2016). Generating Images from Captions with Attention (arXiv:1511.02793). arXiv. https://doi.org/10.48550/arXiv.1511.02793

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., & Chen, M. (2022). GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models (arXiv:2112.10741). arXiv. https://doi.org/10.48550/arXiv.2112.10741

OpenStreetMap contributors. (2017). *Planet dump retrieved from https://planet.osm.org* [Data set]. OpenStreetMap. https://www.openstreetmap.org

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative Adversarial Text to Image Synthesis (arXiv:1605.05396). arXiv. https://doi.org/10.48550/arXiv.1605.05396

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., & Aberman, K. (2023). DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation (arXiv:2208.12242). arXiv. https://doi.org/10.48550/arXiv.2208.12242

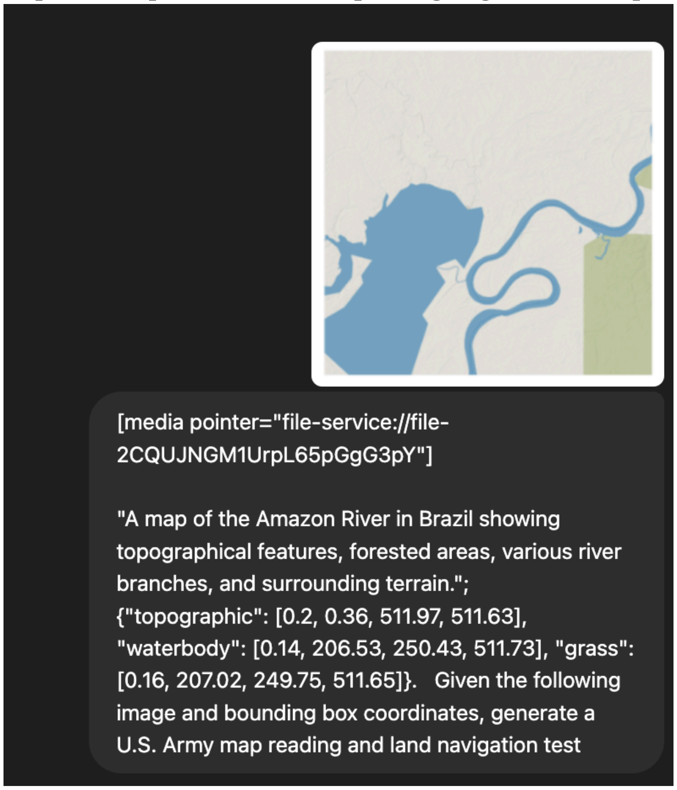
Sun, C., Xu, Y., Xu, X., Fan, X., Bai, J., Lu, X., & Chen, Z. (2025). Bridging Scales in Map Generation: A scale-aware cascaded generative mapping framework for seamless and consistent multi-scale cartographic representation (arXiv:2502.04991). arXiv. https://doi.org/10.48550/arXiv.2502.04991

Wang, H., Ren, P., Jie, Z., Dong, X., Feng, C., Qian, Y., Ma, L., Jiang, D., Wang, Y., Lan, X., & Liang, X. (2024, July 10). OV-dino: Unified open-vocabulary detection with language-aware selective fusion. arXiv.org. https://arxiv.org/abs/2407.07844

Zhang, T., Wang, X., Tai, Z., Li, L., Chi, J., Tian, J., He, H., & Wang, S. (2025). STRICT: Stress Test of Rendering Images Containing Text (arXiv:2505.18985). arXiv. https://doi.org/10.48550/arXiv.2505.18985

APPENDIX A. AI-Generated Draft for a Map Reading and Land Navigation Test (June 4, 2025 - GPT 4o)

Step 1: Prompt to the LLM, requesting to generate the quiz:



Step 2: LLM system asks if user would like PDFs

Test Page 1

U.S. Army Map Reading and Land Navigation Test

Map Description:

This map shows a section of the Amazon River in Brazil, featuring waterbodies, topographical elevation, and forested (grassy) areas.

Bounding Box Coordinates:

- Topographic Area: (0.2, 0.36) to (511.97, 511.63)
- Waterbody Area: (0.14, 206.53) to (250.43, 511.73)
- Grass/Forested Area: (0.16, 207.02) to (249.75, 511.65)

Test Page 2

Candidate Worksheet

Part I: Map Interpretation

1. Identify and Describe Boundaries
   - Q1. Using the bounding box coordinates, describe the location of the waterbody relative to the topographic and grass areas.

2. Relative Positioning
   - Q2. What is the orientation of the grassy area relative to the waterbody (e.g., north, south, east, west)?

3. Area Estimation
   - Q3. Assuming 1 pixel equals 1 meter, calculate the area in square meters of:
     - a) The waterbody area.
     - b) The grassy area.

4. Distance Measurement
   - Q4. Calculate the approximate center points of the waterbody and grass bounding boxes.
   - Q5. Determine the straight-line distance between these two center points.

5. Azimuth Calculation
   - Q6. From the center of the grass area to the center of the waterbody, calculate the azimuth (bearing angle) in degrees.

Test Page 3

Part II: Land Navigation

6. Route Planning
   - Q7. Starting from the center of the grassy area, plan a movement route to the edge of the waterbody area while avoiding crossing the topographic region. Sketch your route below.

7. Tactical Consideration
   - Q8. Given the size and shape of the grass and waterbody areas, which would be more suitable for concealing a platoon-sized element? Justify your answer based on cover and concealment principles.

Sketch Box (for Q7):

Answer Key:

Question	Expected Response
Q1	Waterbody lies southwest of the topographic area; grassy area lies adjacent east/northeast.
Q2	Grass area is northeast relative to the waterbody.
Q3a	Waterbody area = 250.29m x 305.20m ~ 76,418.8 m2.
Q3b	Grass area = 249.59m x 304.63m ~ 76,035.7 m2.
Q4	Waterbody center ~ (125.285, 359.13); Grass center ~ (124.955, 359.335).
Q5	Distance ~ sqrt((125.285-124.955)^2 + (359.13-359.335)^2) ~ 0.36 meters.
Q6	Azimuth ~ arctangent((359.13-359.335)/(125.285-124.955)) ~ 33 degrees (NE direction).
Q7	Suggested route: Move northeast around the topographic area, staying in the open grassland to avoid water and elevation.
Q8	Grassy area; offers natural concealment without reflective surfaces, flatter terrain for easier movement, and better overhead cover.

Notes for Instructors:

- Allow a margin of error of ±5% in distance and area calculations.
- Azimuth calculations should be within ±5 degrees.
- Sketches should show logical and tactical routing around obstacles.
- Tactical justifications should reference basic cover and concealment principles.

Test Page 4

Instructions:

- Use a protractor and a ruler for accurate azimuth and distance calculations.
- Round answers to two decimal places unless otherwise noted.
- Ensure your route sketch is clearly labeled and includes a legend if necessary.

End of Test