S1. Supplementary

S1.1. Additional Details on MapVerse Collection and Annotation

To ensure diversity and realism, we curated maps from publicly available sources across the internet, including news portals, educational resources, weather forecast platforms, and other open forums. For each map type, targeted Google searches were performed using specific keywords to retrieve a large initial pool of candidate maps. The search queries included terms such as:

- **Network Maps:** , "Ferry route maps", "School bus routes map", "Human migration route maps", "MigrantMaps" , "High quality metro map"
- **Divisional Maps:** "Divisional maps for pincodes Africa", "Divisional maps for pincodes Australia", "Divisional maps for pincodes Europe", "DivisionMapsSchoolDistrict" , "Time zones map"
- **Choropleth Maps:** "Choropleth", "choropleth map USA", "choropleth map world", "choropleth maps others", "multivariate choropleth maps"
- **Layout Maps:** "Memorial hospital layout maps", "Stadium layout maps"
- **Marker Maps:** "Top hospitals in California marker maps","Dot_Maps", "World_Heritage",
- **Mixed Maps:** "Printable city tourist attractions map", "Sightseeing map",
- **Isopleth Maps:** "Contour maps", "Isopleth weather maps", "Isopleth" , "climate", "WeatherMapsIso"
- **Multiple Maps:** "Bivariate map", "ComparisonMaps"
- **Conflict Maps:** "Territory changes in conflict map"
- **Cartograms:** "Cartogram maps", "cartogram in statistics"

While scraping the data from the internet, we browsed through all the maps images and then the final set of maps were selected based on human readability and resolution. Also note that while these queries were made with the intention of retrieving those candidate maps of the search query categories, we have manually verified removed noisy, low resolution, blurry, cropped or irrelevant images in the process along with reclassifying the misclassified ones.

Additional Annotation Details. Maps could be independently annotated by multiple annotators, each generating their own questions and answers without access to others' work. This intentional redundancy was designed to capture variability in human interpretation and improve QA quality. Even when the same map was assigned to multiple annotators, each produced a distinct set of questions; consequently, we did not encounter cases where annotators generated identical questions. Therefore, the issue of disagreements in answers does not apply here.

S1.2. Additional MapVerse taxonomy

Table S1 provides a detailed breakdown of the number of images in MapVerse across various resolutions, highlighting the prevalence of each size. Table S2 further characterizes the dataset by showing how images are distributed according to different levels of geographic granularity, offering insight into the spatial coverage and diversity of the dataset.

Image Size	Count
>8MP	79
2–8MP	212
0.5–2MP	465
0.1–0.5MP	250
<0.1MP	19

Table S1. **Image Size Distribution by Megapixels (MP) in MapVerse.**

Geographic Level	Count
World	168
Continent	68
Subcontinent	49
Country	264
State	32
Region	121
County	14
City	147
District	28
Neighborhood	30
Campus	42
Building	62

Table S2. **Distribution of Maps Across Different Geographic Granularity**.

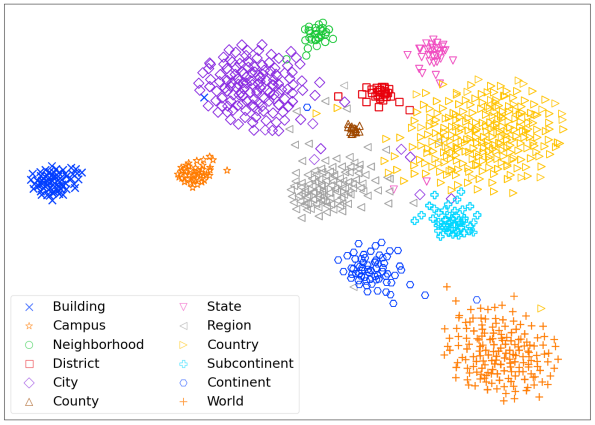
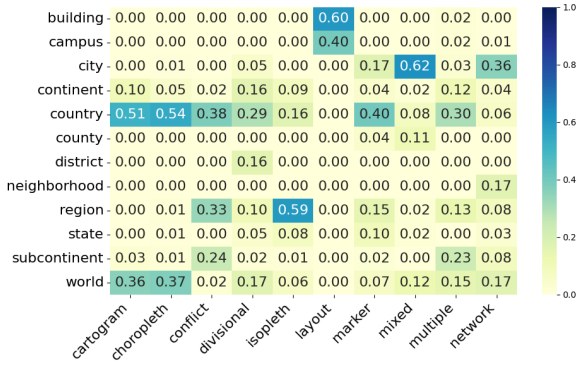
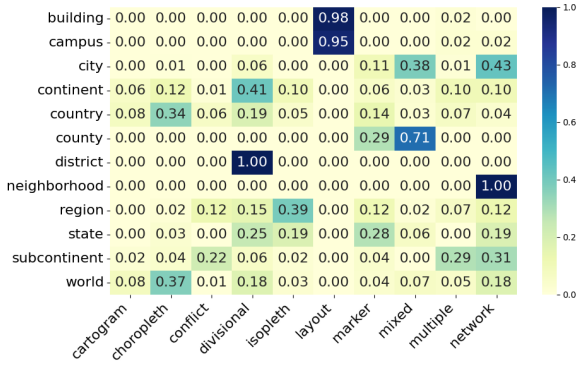


Figure S1. **T-SNE for images geographical granularity.**



(a) Relative heatmap normalized by map type.



(b) Relative heatmap normalized by geographical granularity.

Figure S2. Correlation heatmaps showing the proportional relationship between map type (X-axis) and geographic level (Y-axis)

Figure S2 shows two heatmaps that provide a comprehensive analysis of the distribution of geographical granularity across different map types in our dataset. This analysis highlights significant structural biases and specialization within the data.

We normalize the heatmap in Figure S2a by map type and show the proportional distribution of geographic levels within each map category. The data reveals a strong trend of specialization. For instance, layout maps are almost exclusively associated with the building and campus levels, while choropleth and isopleth maps have been frequently found at the country level. In contrast, map types like mixed and network show a more diverse spread across various geographic levels, indicating their versatility for representing a wider range of spatial scales.

The heatmap in Figure S2a, normalized by geographical granularity, offers a complementary perspective by showing which map types are most prevalent within a specific geographic level. The data highlights a clear dominance of certain map types at particular granularity. For example, layout maps constitute nearly all content at the building and campus levels, confirming their conventional use for these finer granularities. Similarly, divisional and network maps dominate at district and neighborhood levels respectively. However, some geographic granularities, such as country and state, are more diverse, with significant representation from multiple map types. This suggests a richer variety of mapping tasks and conventions are employed at these scales.

S1.3. Evaluation Metrics

Rank-Wise Precision (RWP) evaluates a model's ranking ability by considering not only the order of elements but also instances where the model may add or delete elements from the set. This differs from simple rank-based metrics which assume a fixed set of elements and only measure misalignment. It is defined as the average of the $Precision@K$ for each valid value of K. The formula is given by:

$$\text{RWP} = \frac{1}{M} \sum_{k=1}^{M} \text{Precision@k} \tag{1}$$

where L_{pred} is the model-predicted ranked list, $|L_{ref}|$ is the correct ranked list and $M = max(|L_{pred}|, |L_{ref}|)$ with $|L|$ denoting length of list L.

S1.4. Ablation Results

Table S3 presents the results of our ablation study examining how various types of image corruption affect the performance of VLMs, highlighting the model's sensitivity to degraded visual input.

Similarly, Table S4 illustrates the impact of reduced image resolution on VLM performance, demonstrating how lower-quality visual data can lead to substantial drops in accuracy across different tasks, especially for reasoning questions. Although the dataset includes map images with varying native resolutions, we did not apply any manual resizing or preprocessing. All images were loaded as PIL objects and passed directly to the corresponding Hugging Face AutoProcessor, which automatically performs any required transformations—such as resizing or padding—to match the vision encoder's input specifications. Thus, while the models accept images of arbitrary native resolution, all rescaling is handled internally by the models' official preprocessing pipelines.

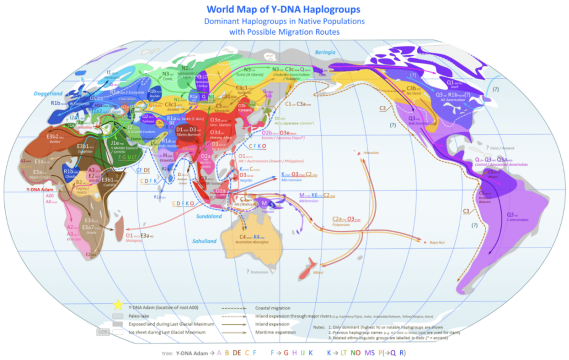
	Bool	**SE**	**Count**	**List (P/R)**	**Rank**	**Rsn**
Metrics	*EM*	*EM*	*EM*	*Prec / Rec*	*RWP*	*EM*
Overall	72.2	26.8	23.0	30.4 / 21.3	49.6	30.7
RN	71.5	26.4	22.7	29.7 / 21.1	50.0	29.1
RPN	70.9	21.6	20.9	24.5 / 17.0	48.5	24.0
RBR	71.5	24.9	21.8	29.0 / 19.7	49.2	27.0

Table S3. **QWEN 2.5 VL – Perturbation Ablation.** RN = Random Noise, RPN = Random Pepper Noise, RBR = Random Black Rect,EM = Exact Match, RWP = Rank-wise Precision. List column shows Precision / Recall.

	Bool	SE	Count	List (P/R)	Rank	Rsn
Metrics	*EM*	*EM*	*EM*	*Prec / Rec*	*RWP*	*EM*
Orig. Res	72.2	26.8	23.0	30.4 / 21.3	49.6	30.7
50% drop	70.5	21.9	20.3	25.3 / 17.3	48.3	25.3
75% drop	69.1	16.1	17.9	19.2 / 13.0	45.4	15.2

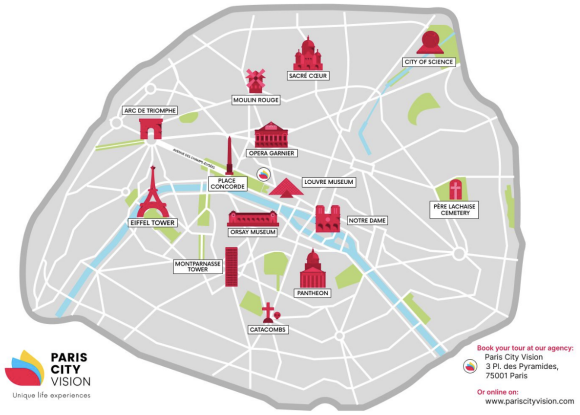
Table S4. **QWEN 2.5 VL – Resolution Ablation.** EM = Exact Match, RWP = Rank-wise Precision. List column shows Precision / Recall.

S1.5. Additional MAPVERSE question answers Samples



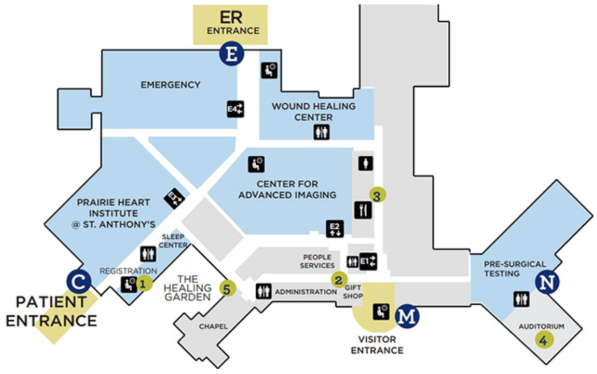
- Question: What haplogroups label appear over Japan and Korea?
  - Correct Answer: **O3e-M134, O2b-M176**
  - LLM Answer: **O1-M119, O2b-M176, O3e-M134**
- Question: Which haplogroup covers the island of Madagascar?
  - Correct Answer: **O1-M119**
  - LLM Answer: **O1-M119**
- Question: What haplogroup appears in high frequency in the Caucasus region?
  - Correct Answer: **G2a-P15**
  - LLM Answer: **R1b**

Figure S3. Sample QA for mixed map type (isopleth + network)



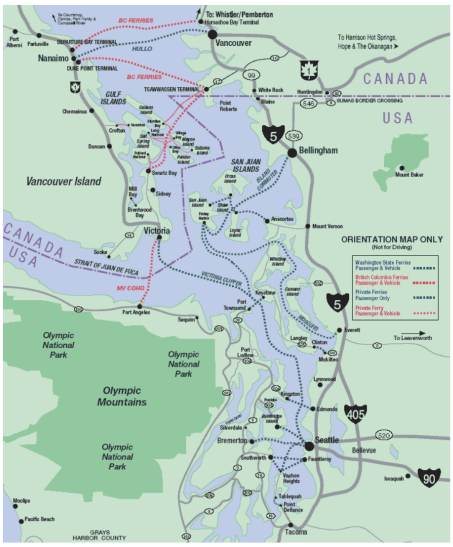
- Question: Which landmark is located directly south of the Panthéon?
  - Correct Answer: **Catacombs**
  - LLM Answer: **Catacombs**
- Question: How many total landmarks are shown?
  - Correct Answer: **14**
  - LLM Answer: **13**
- Question: Name all the museums shown on the map
  - Correct Answer: **Louvre,Orsay**
  - LLM Answer: **Louvre, Orsay**

Figure S4. Sample QA for mixed map type (marker + network)



- Question: How many restrooms are available for women?
  - Correct Answer: **6**
  - LLM Answer: **3**
- Question: Near which entrance is the Gift Shop located?
  - Correct Answer: **Visitor**
  - LLM Answer: **Administration**
- Question: With what letter the patient entrance is depicted?
  - Correct Answer: **C**
  - LLM Answer: **C**

Figure S5. Sample QA for layout map type

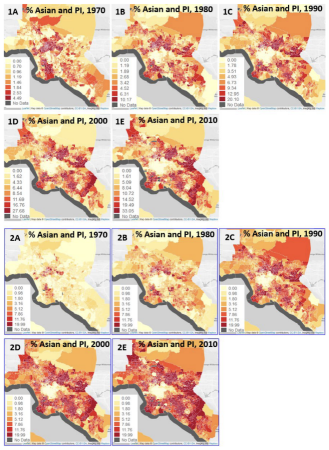


- Question: I'm in the US. I crossed the US - Canada border 1232131231231 times. Which country am I in?
  - Correct Answer: **Canada**
  - LLM Answer: **Canada**
- Question: Name something on the map above that has the word rock in it
  - Correct Answer: **White Rock**
  - LLM Answer: **White Rock**

Figure S6. Sample QA for mixed map type (isopleth + network + marker)

- Question: Does the combined population of Pakistan and Bangladesh exceed 400 million?
  - Correct Answer: **No**
  - LLM Answer: **Yes**
- Question: How many squares represent the population of Oceania?
  - Correct Answer: **82**
  - LLM Answer: **41**
- Question: Order the following Middle Eastern countries from largest to smallest population in 2018 as depicted on the map: Iran, Turkey, Saudi Arabia, Iraq, Egypt.
  - Correct Answer: **Egypt, Iran, Turkey, Saudi Arabia, Iraq**
  - LLM Answer: **Turkey, Iran, Egypt, Saudi Arabia, Iraq**

Figure S7. Sample QA for cartogram map type



- Question: What are the lowest and highest density values in Panel 2B's legend?
  - Correct Answer: **0.0, 19.99**
  - LLM Answer: **0.98, 11.76**
- Question: What is the difference between the highest density value in 1A and 1B?
  - Correct Answer: **5.68**
  - LLM Answer: **1.89**
- Question: Which panel has more ""No Data"" entries: 2D or 2E?
  - Correct Answer: **2E**
  - LLM Answer: **2D**

Figure S8. Sample QA for multiple map type

S1.6. Custom Prompt for Question Answering

This subsection contains the full text of the custom prompt used to guide our AI agent's analysis of map data, as detailed in the main paper. The prompt provides a structured set of instructions and examples to ensure a consistent and logical approach to interpreting spatial information.

Custom Prompt

You are an AI Agent with specialised knowledge in reading and understanding map data. Analyze the following map and using information from the steps and examples given below, answer the question.

Steps to follow:

1. Identify Map-Related Elements in the Question
2. Locate the Identified Elements on the Map
3. Apply Logical Reasoning
4. Formulate a Concise Answer

Based on your reasoning, arrive at a clear and accurate answer. Return only a word or phrase, as required—no explanation is needed. If adequate data is not present, give answer as "no data". If you have all the data and there is no answer to the question, give answer as "none". If it is a counting problem, give answer 0. If you have all the data and it is not possible to answer the question, give answer "not possible".

Assuming we are talking about a map with election results for USA. This map contains the voter breakdown across the United States, including the number of votes cast and the winning party in each state. Some examples of questions and their answers are as follows:

Question: Count the number of states on the west coast where Democrats won.
Answer: 3
Question: Based on the information given in the map, who won the election, Democrats or Republicans?
Answer: Democrats
Question: Based on the information given in the map, if both Democrats and Republicans win 25 states each, do we have more blue states or red states?
Answer: neither
Question: List the top 4 states in terms of seats where the republicans won
Answer: Texas, Georgia, Missouri, Tennessee
Question: Rank these states in ascending order of seats - kansas, south carolina, nebraska, oklahoma, colorado, wisconsin
Answer: nebraska, kansas, oklahoma, south carolina, colorado, wisconsin
Question: Based on reasoning, Answer the following: Montana : Wyoming :: North Dakota : ?
Answer: South Dakota

Now, Answer the Question below based on the information, instruction and examples above:

S1.7. Annotator Instructions

The following instructions were provided to human annotators for creating the question-answer pairs in the MAPVERSE benchmark. The goal was to produce questions that require genuine geospatial reasoning and are difficult for large language models to answer correctly.

Instructions

Task Objective: Your task is to create map-based questions along with answers based on the information provided in a map, ensuring that a human can easily answer them.

- Provide objective questions grounded on the given map.
- Provide the correct answer to the question (single word or a few words).
- Include the answer provided by the LLM.

Answers must be based on:

- The information presented in the map.
- General common understanding of how maps are read (spatial common sense).

Instructions for Annotation:

- Use the "Add More Questions" button to get input spaces for question, correct answer, and answer by LLM.
- Once done creating all the questions, press the submit button. Your annotation would be saved by the portal.

Preferred

✔ Analyze the map carefully to understand the information provided.
✔ Use only the information in the map to answer the question.
✔ Keep your questions simple and straightforward.
✔ Keep your answers concise (within a few words).
✔ Use common sense to provide additional context where necessary.
✔ Use proper grammar, spelling, and punctuation.
✔ Create at least 10 questions per map.

Avoid

✘ Using outside knowledge not present in the map.
✘ Creating questions that are too long or difficult to understand.
✘ Creating ambiguous questions.
✘ Providing answers that are overly verbose or unclear.
✘ Providing subjective or ambiguous answers.
✘ Using abbreviations or acronyms not defined in the map.
✘ Creating more than three questions of the same type on a single map.

Important Note Avoid using information that you may know if you believe it is not generally known.
The following examples clarify the instructions. Please review them carefully.

[Link to Sample Annotation]