

Learning to Wander: Improving the Global Image Geolocation Ability of LMMs via Actionable Reasoning

Supplementary Material

S1. Prompt Design Analysis

To ensure fair evaluation and robust automated parsing across diverse Large Multimodal Models (LMMs), we engineered specific prompt structures for both the Baseline and GeoAoT settings. Here we provide the rationale behind key design decisions.

S1.1. Rationale for Structured Output and Constraints

- **Unified JSON Format:** LMMs often exhibit high variance in conversational style. To facilitate automated evaluation without complex regex parsing, we enforce a strict JSON schema. This ensures that critical outputs specifically `final_guess` and `action` are machine readable and consistently formatted across all 20 tested models.
- **Explicit "Punishment" for Refusal:** Safety-aligned models (e.g., GPT-4o, Gemini) often refuse to answer geolocation queries when visual cues are ambiguous, defaulting to responses like "I cannot determine the location." To mitigate this, we include a directive warning of "severe punishment" for refusal. This forces the model to leverage its latent world knowledge and provide its best possible estimate, which is essential for measuring fine-grained geospatial capabilities.
- **Dual Modality for Guessing:** We allow models to predict via either `detailed_description` (textual) or `coordinates` (numerical). This accommodates different model strengths; some LMMs excel at retrieving specific place names (e.g., "Eiffel Tower"), while others are better optimized for spatial coordinate regression.

S1.2. Rationale for GeoAoT Action Space

The GeoAoT prompt introduces specific constraints to enable the **Action of Thought** mechanism within the WanderBench graph structure:

- **Discrete Action Mapping:** The prompt explicitly limits the action space to `rotate_degrees`, `move_to_color`, and `back_to_original`. This maps the continuous visual input to the discrete topological graph of WanderBench, allowing the agent to navigate defined nodes and edges without hallucinating impossible physical actions.
- **State Awareness via Suffix Prompts:** As LMMs are stateless, we inject the current environment state (e.g., "Available move color..") into the user prompt at every turn. This grounds the model's reasoning in the immediate physical reality of the navigation graph, preventing it from attempting to traverse non-existent paths.

S1.3. Baseline Prompt

```
System Prompt:
IMPORTANT: Always respond in this unified JSON
format:
{
  "observations": "What you see in the image in
one or two short sentences",
  "confidence": "low" | "medium" | "high",
  "action": "guess",
  "final_guess": {
    "method": "detailed_description" | "
coordinates",
    "location_description": "[Address] [Street
Name], [City], [State], [Country]",
    "coordinates": {"latitude": 34.0522, "
longitude": -118.2437}
  }
}
```

Since this is a direct prediction (no exploration), always set `action = "guess"` and provide your `final_guess`.

When making `final_guess`:

- method `"detailed_description"`: MUST use EXACT format `"[Address] [Street Name], [City], [State], [Country]"` OR reduced specificity like `"[City], [State], [Country]"` OR `"[State], [Country]"` OR `"[Country]"`. NO descriptive text, NO explanations, ONLY location names in this format. If absolutely no geographical clues available, use EXACTLY `"FAIL TO PREDICT"`. Set `coordinates` to null.
- method `"coordinates"`: Set `coordinates` to `{"latitude": X, "longitude": Y}`, set `location_description` to null.

CRITICAL: `location_description` must be ONLY location names in the specified format. NO descriptive phrases like "suburban area" or "appears to be".

WARNING: Using `"FAIL TO PREDICT"` results in severe punishment. When uncertain, strongly consider using `coordinates` method instead.

User Prompt:

```
"Analyze this Street View panorama and make your
best guess about the location. Provide your
response in the required JSON format."
```

Listing 1. System and User Prompts for the Baseline Model.

S1.4. GeoAoT Prompt

```
System Prompt:
IMPORTANT: Always respond in this unified JSON
format:
```

```

{
  "observations": "One or two short sentence
  describe you see in the image",
  "confidence": "low" | "medium" | "high",
  "action": "continue" | "guess",
  "next_move": {
    "type": "rotate_degrees" | "move_to_color" |
    "back_to_original",
    "details": "45" | "red" | null
  },
  "final_guess": {
    "method": "detailed_description" | "
    coordinates",
    "location_description": "[Address] [Street
    Name], [City], [State], [Country]",
    "coordinates": {"latitude": 34.0522, "
    longitude": -118.2437}
  }
}

Actions:
- If action = "continue": Provide next_move (
  rotate_degrees/move_to_color/back_to_original
  )
- If action = "guess": Provide final_guess with
  location

When making final_guess:
- method "detailed_description": MUST use EXACT
  format "[Address] [Street Name], [City], [
  State], [Country]" OR reduced specificity
  like "[City], [State], [Country]" OR "[State
  ], [Country]" OR "[Country]". NO descriptive
  text, NO explanations, ONLY location names in
  this format. If absolutely no geographical
  clues available, use EXACTLY "FAIL TO PREDICT
  ". Set coordinates to null.
- method "coordinates": Set coordinates to {"
  latitude": X, "longitude": Y}, set
  location_description to null.

CRITICAL: location_description must be ONLY
  location names in the specified format. NO
  descriptive phrases like "suburban area" or "
  appears to be".

WARNING: Using "FAIL TO PREDICT" results in
  severe punishment. When uncertain, strongly
  consider using coordinates method instead.

```

Listing 2. System and User Prompts for the GeoAoT.

```

User Prompt (First Step):
"Here's the starting panorama view (total
  continue step expect guess {current_step}/{
  max_step_minus_one}). Analyze the image and
  respond in JSON format."

User Prompt (Follow-up):
"Here's your current view after your last action
  (total continue step except guess {
  current_step}/{max_step_minus_one}). Analyze
  this new view and respond in JSON format."

User Prompt (Available Moves Suffix):
" Available move color you can use is ONLY {
  available_moves}"

```

Listing 3. User Prompts for the GeoAoT During multi-turn interaction.

S1.5. Location Proposing Prompt

```

System Prompt:
You are a GeoGuessr expert creating challenging
  location-based questions.
You must respond with a structured format that
  includes either coordinates or a location
  name with great detail.

IMPORTANT: Always end your response with one of
  these formats:
- COORDINATES: (latitude, longitude)
- LOCATION: specific place name

Be specific about the location and explain why it
  matches the requested difficulty level.

User Prompt (Stage1):
Create a GeoGuessr location for human players
  with approximately {geo-score} geo score in{
  aera}

  REQUIREMENTS:
  - Choose locations that definitely have
  Street View coverage
  - Avoid extremely remote or rural areas
  without Street View
  - Pick populated areas with businesses,
  landmarks, or distinctive features{
  failed_locations_text}

  Consider factors like:
  - Distinctive landmarks or features that
  provide clues
  - Level of urbanization/development (
  prefer moderate to high)
  - Availability of street view coverage (
  ESSENTIAL)
  - Cultural/geographic markers

  Good examples: city centers, suburban
  areas, highway intersections near towns,
  tourist areas.
  Bad examples: remote desert roads,
  unpopulated wilderness areas, extremely rural
  locations.

  Provide either exact coordinates or a
  specific location name that can be geocoded.
  End your response with either:
  COORDINATES: (lat, lon)
  OR
  LOCATION: place name

User Prompt (Stage 2):
I'm showing you a satellite map with colored dots
  representing different Street View locations
  .

  Available colors and their coordinates:
  {chr(10).join(color_descriptions)}

```

```
Please examine the map and select ONE
color that would make for an interesting and
challenging GeoGuessr question.
```

```
Consider factors like:
- Proximity to landmarks or distinctive
features
- Urban vs rural setting
- Availability of visual clues
```

```
IMPORTANT: Respond with exactly this
format:
SELECTED_COLOR: [color_name]
```

```
Then explain your reasoning briefly."
```

User Prompt (Stage 3):

```
Here is the Street View panorama for the location
you selected: {location_info}
```

```
Please examine this panorama and
determine if it would make a good GeoGuessr
question.
```

```
Consider:
- Are there enough visual clues for
players to make educated guesses?
- Is the image quality acceptable?
- Does it match the intended difficulty
level?
- Are there distinctive features or
landmarks visible?
```

```
IMPORTANT: Respond with exactly this
format:
APPROVAL: YES or NO
```

```
Then provide your reasoning. If you say
NO, suggest what type of location would be
better.
```

Listing 4. User Prompts for the Model Location Proposing.

S2. Qualitative Results

In this section, we present qualitative examples to illustrate the performance and reasoning capabilities of LMMs on WanderBench. We examine three key scenarios: (1) the baseline behavior where models rely on static visual cues, (2) the GeoAoT framework where agents actively explore the environment to reduce uncertainty, and (3) the generative capability where models design difficulty-calibrated geolocation tasks.

S2.1. Geolocation Baseline Result

We first examine the **Baseline Direct Answering** setting. In the example below, **LLaVA OneVision 72B** correctly identifies the general linguistic region based on “signs in Spanish,” but fails to extract fine-grained discriminative features (such as the specific phone country code visible in the scene). Consequently, the model hallucinates a location in

Buenos Aires, Argentina (approx. coordinates -34.6, -58.0) rather than the correct location in Valparaíso, Chile. This misclassification leads to a significant **Distance Error of 1235 KM** and a moderate **Geo Score of 50.42**, highlighting the limitations of static inference when distinctive cues are missed or misinterpreted.

Analyze this Street View panorama and make your best guess about the location. Provide your response in the required JSON format.



observations: Busy street with cars, buildings, and a clear sky; signs in Spanish suggesting a Spanish-speaking country.,
confidence: medium,
action: guess,
next_move:
type: guess, details: null
final_guess: coordinates,
location_description: null,
coordinates: **latitude: -34.6037, longitude: -58**

S2.2. Geolocation with GeoAoT Result

Next, we demonstrate the efficacy of the **GeoAoT** framework using the **o3** model. In this challenging Andean environment, the initial view offers limited specificity. Rather than guessing immediately, the agent utilizes the action space to explore. Through a multi-step trajectory—moving towards specific colored nodes and rotating the view—the model actively gathers cumulative evidence. It identifies “Spanish-style architecture,” “clay tile roofs,” and “blue street signs,” allowing it to refine its hypothesis from a generic “hilly urban area” to a precise coordinate in Cusco, Peru, achieving a distance error of only 1.0 km.

Here's the starting panorama view (total continue step except guess 0/5). Analyze the image and respond in JSON format. Available move color you can use is ONLY {red, blue}.



Here's your current view after your last last action (total continue step except guess 2/5). Analyze this new view and respond in JSON format. Available move color you can use is ONLY {red, blue}.



observations: Hilly urban area with red brick buildings and clay tile roofs; paved road with blue and red directional arrows visible.,
confidence: medium,
action: continue,
next_move:
type: move_to_color, details: blue
final_guess: detailed_description,
location_description: null,
coordinates: null



observations: Back at initial spot overlooking valley; blue-tiled guardrail on right, colonial white house with blue street sign on left.,
confidence: medium,
action: continue,
next_move:
type: rotate_degrees, details: 45
final_guess: detailed_description,
location_description: null,
coordinates: null

Here's your current view after your last action (total continue step except guess 1/5). Analyze this new view and respond in JSON format. Available move color you can use is ONLY {red, blue}.



Here's your current view after your last action (total continue step except guess 3/5). Analyze this new view and respond in JSON format. Available move color you can use is ONLY {red, blue}.



observations: White adobe house with clay tile roof, mountainous cityscape in background, Spanish-style architecture and a parked white minibus on right.,
confidence: medium,
action: continue,
next_move:
type: move_to_color, details: red
final_guess: detailed_description,
location_description: null,
coordinates: null



observations: View similar to earlier; intersection with white adobe house, blue street sign at corner, hilly cityscape in background.,
confidence: medium,
action: continue,
next_move:
type: move_to_color, details: red
final_guess: detailed_description,
location_description: null,
coordinates: null

Here's your current view after your last action (total continue step except guess 4/5). Analyze this new view and respond in JSON format. Available move color you can use is ONLY {red, blue}.



observations: Hilly Andean cityscape with red-brick buildings and clay-tile roofs, colonial adobe wall on left, overlooking valley under cloudy sky., confidence: high, action: **guess**, next_move: type: back_to_original, details: null final_guess: coordinates, location_description: null, coordinates: null

Please provide your response in the required JSON format.



observations: Hilly Andean cityscape with red-brick buildings and clay-tile roofs, colonial adobe wall on left, overlooking valley under cloudy sky., confidence: high, action: guess, next_move: type: guess, details: null final_guess: coordinates, location_description: null, coordinates: **latitude: -13.5199, longitude: -71.9740**



This location is chosen to be challenging yet solvable for the average player, aiming for an 80 geo score. The viewpoint provides a clear, iconic landmark but from an unexpected perspective, requiring players to use directional and geographical clues to pinpoint their exact location. The player is placed in Malcolm W. Martin Memorial Park in East St. Louis, Illinois. The view looks west across the Mississippi River directly at the St. Louis, Missouri skyline. The Gateway Arch is the unmistakable primary clue, immediately identifying the city. However, the player is not in St. Louis itself, but across the river in a different state. This creates the central challenge. Players must recognize their position relative to the Arch and use the river and visible bridges (like the Eads Bridge to the right) to accurately place their pin in Illinois, rather than simply clicking on the landmark in Missouri. COORDINATES: (38.629532, -90.173553)",



Create a GeoGuessr location for human players with approximately 80 geo score in North American

S2.3. Location Proposing

Finally, we show the **Location Proposing** capability, where the **Gemini 2.5 Pro** model acts as a task creator. Tasked with generating a North American challenge with a target difficulty (approx. 80 Geo Score), the model demonstrates sophisticated meta-reasoning. It selects a satellite location on the Eads Bridge (marked Green) because it offers a distinct view of the Gateway Arch in St. Louis, Missouri, while the physical location is across the state bor-

der in Illinois. This choice creates a "trick" scenario that tests precise spatial reasoning over simple landmark recognition, perfectly matching the requested difficulty profile.

I'm showing you a satellite map with colored dots representing different Street View locations.

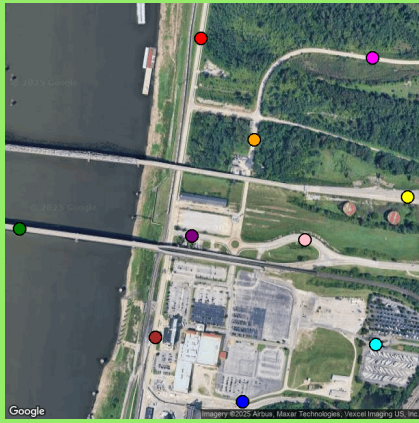
Please examine the map and select ONE color that would make for an interesting and challenging GeoGuessr question.

Consider factors like:

- Proximity to landmarks or distinctive features
- Urban vs rural setting
- Availability of visual clues

IMPORTANT: Respond with exactly this format:

SELECTED_COLOR: [color_name]



Here is the Street View panorama for the location you selected:

Please examine this panorama and determine if it would make a good GeoGuessr question.

Consider:

- Are there enough visual clues for players to make educated guesses?
- Is the image quality acceptable?
- Does it match the intended difficulty level?
- Are there distinctive features or landmarks visible?

IMPORTANT: Respond with exactly this format:

APPROVAL: YES or NO

Then provide your reasoning. If you say NO, suggest what type of location would be better.



SELECTED_COLOR: GREEN

This location is on the Eads Bridge, offering a spectacular and unmistakable view of the St. Louis skyline, including the Gateway Arch. The challenge isn't identifying the city, but rather pinpointing the exact location on a specific, famous bridge. Players must use the river, the structure of the bridge itself, and their position relative to the skyline to get a perfect score, making it an interesting test of precision rather than just general location identification.”,



APPROVAL: YES

This panorama is excellent for a GeoGuessr question. It has an unmistakable primary landmark, the Gateway Arch, which immediately identifies the city as St. Louis. The image quality is clear and unobstructed. The challenge, which matches the intended difficulty, comes from the player's specific position: they are on the historic Eads Bridge, looking west from Illinois into Missouri. Players will easily identify the city, but getting a high score requires them to pinpoint their location on the correct bridge over the Mississippi River, making it a great test of precision.

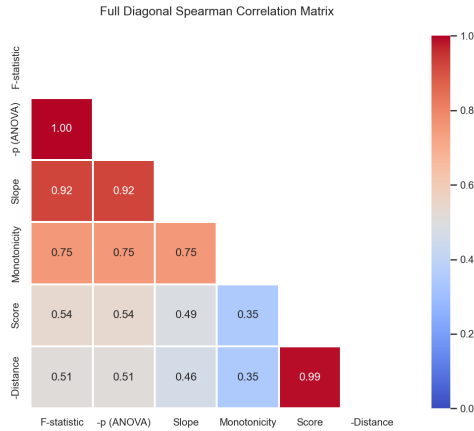


Figure S1. Full diagonal Spearman correlation matrix. The heatmap reveals not only the link between proposing metrics and geolocation accuracy (bottom rows) but also the strong internal consistency between statistical separation (ANOVA) and trajectory ordering (Slope/Monotonicity).

S3. Correlation Analysis: Proposing vs. Geolocation

To understand the relationship between a model’s ability to act as a “proposer” (generating locations with calibrated difficulty) and its effectiveness as a “solver” (geolocating images), we performed a Spearman rank correlation analysis across the 11 models common to both tasks.

We hypothesize that models with a refined meta-understanding of geospatial difficulty—evidenced by their ability to statistically separate “Easy” from “Hard” locations—possess the underlying world knowledge required for accurate geolocation. Figure S1 visualizes the full pairwise correlations between proposing quality metrics (ANOVA F -statistic, Significance $-p$, OLS Slope, Monotonicity) and downstream geolocation performance (Average Score, Negated Distance).

S3.1. Quantitative Analysis

The correlation matrix (Figure S1) reveals two distinct patterns regarding model capability:

1. Internal Consistency of Proposing Metrics: We observe extremely strong correlations between the variance-based metrics (ANOVA F -statistic) and the trajectory-based metrics. Specifically, the F -statistic shows a correlation of $\rho = 0.92$ with the OLS Slope and $\rho = 0.75$ with Monotonicity. This implies that models achieving high statistical separation between difficulty levels are not generating random noise; rather, they are consistently ordering locations in the correct difficulty direction (Easy \rightarrow Hard). A model that successfully “separates” distributions (high F) almost invariably does so by following a positive difficulty gradient

(high Slope).

2. Predictive Power for Geolocation: As hypothesized, the ability to calibrate difficulty correlates positively with downstream geolocation performance. The ANOVA F -statistic shows the strongest relationship with Geolocation Score ($\rho = 0.54$) and Distance ($\rho = 0.51$). While the Monotonicity indicator is internally consistent with the other proposing metrics, its direct correlation to geolocation score is weaker ($\rho = 0.35$). This suggests that the *magnitude* of the difficulty separation (captured by the F -statistic and Slope) is a more reliable predictor of geospatial reasoning mastery than the binary presence of monotonic ordering alone.

In conclusion, the models that best solve the geolocation task are the same models that possess a nuanced, statistically significant understanding of what makes a location “Hard” or “Easy.”