Vision Language Models are Confused Tourists

Supplementary Material

A. Country & Image License Details

All data used in this work are sourced from Wikimedia Commons. Each instance whether it represents a flag, an item (e.g., attire, cuisine, or musical instrument), or a landmark, we include its corresponding license information, as shown in Table 3. We strictly include only media that are permitted for research use and redistribution.

B. Evaluation on Prior Benchmarks

We benchmark GPT-5 as the new upper bound model for prior benchmarks namely CVQA [19] and WorldCuisines [26] to assess improvements over previously evaluated models (e.g., GPT-4), which represent the strongest reported results to date.

For CVQA, we evaluated the model in the location-agnostic setting using the prompt format "Question: question Options: options Short Answer:" without including any country information. This setting yielded a score of 74.7%, whereas the best result reported in the paper was 48.7% (InstructBLIP [5]).

For WorldCuisines, we evaluated both Task 1 (cuisine identification) and Task 2 (country prediction) in the multiple-choice (MCQ) setting. We followed the custom prompt provided for each question in the benchmark. GPT5 model achieved 92.7% for task 1, and 78.4% for task 2, compared to the best reported performance of 88.4% and 66.52% (GPT-4o)

C. Image Generation Prompt & Model Settings

The image generation process utilizes the `gemini-2.5-flash-image` model with fixed configuration settings (temperature $= 0.0$, aspect_ratio $= 1:1$, max_output_tokens $= 648$). The methodology aims to test model robustness by generating a photorealistic image where a traditional cultural item ({`item_name`} from {`origin_country`}) is integrated into the environment of an {`adversarial_country`}. The prompt template is constrained to curate a central object placement (occupying $\geq 1/3$ area). The environmental blend is achieved by inserting a {`background_instruction`} corresponding to the test mode (`flag`, `landmark`, or `flag_landmark`) and a supporting instruction based on the object {`category`} (e.g., table for Cuisine).

The complete image generation prompt text is structured as follows, with placeholders filled by the experimental variables:

```
Context: I am providing you a traditional {category}
from {origin_country} called {item_name}. Generate a
photorealistic image of this {category} as if it is
situated in {adversarial_country}, blending it naturally
into the new environment.

Image generation rules:
• The {category} object should be placed in the center
  of the image, taking up at least 1/3 of the total
  image area to ensure it's prominently featured and
  clearly visible.
• {background_instruction}
• Maintain the originality, completeness, and detail of
  the {category} object itself.
• Integrate the {category} and the new scenery by adding
  a relevant supporting object to make the scene natural
  and cohesive ({supporting_bg_instruction}).
• IMPORTANT: Do not include any people, human figures,
  or human silhouettes in the background or anywhere in
  the image. Focus only on the {category} object and
  the environmental scenery.
• The image should be photorealistic, with the object
  and environment being blended seamlessly so it looks
  completely natural and not artificially composed.
• Avoid any obvious compositing artifacts, unnatural
  lighting, or perspective mismatches between the
  {category} and its surroundings.
• Ensure consistent lighting, shadows, and color grading
  across the entire image to enhance realism.
```

D. Manual Evaluation Rubrics

We define the rubrics for evaluating the quality of the model-assisted perturbation below. Worth noting that we use the same number of samples per category, so no category is left behind. Based on that rubric, the generated image perturbation is 4.49, which reflects the high quality of the perturbed outputs.

- **5 (Excellent)**. Perturbation is highly natural and seamless. The added element (flag, landmark, or both) blends perfectly with lighting, perspective, and style. No visible artifacts, mismatched colors, or unnatural edges. The image appears authentic and coherent.
- **4 (Good)**. Perturbation is mostly natural. Minor inconsistencies in lighting, scale, or blending are noticeable on close inspection but do not significantly harm realism. Overall, the image looks believable.
- **3 (Fair)**. Perturbation is moderately convincing. Integration issues such as slight misalignment, contrast mismatch, or unrealistic positioning are visible. The edit is understandable but clearly artificial.
- **2 (Poor)**. Perturbation is visibly artificial. Clear signs of editing, such as wrong lighting, perspective errors, or poor blending. The added element does not integrate well with the original image.
- **1 (Very Poor)**. Perturbation is unnatural or of low quality. Severe mismatches in scale, lighting, or realism. The added element appears pasted-on, distorted, or contextually incoherent. The image looks fake or broken.

Table 2. Geographic Breakdown of Countries in the Dataset by Region and Sub-Region, with Regional Totals.

Region	Sub-Region	Countries	Region Total
Africa	Sub-Saharan Africa	DR Congo, Ethiopia, Nigeria, South Africa, Tanzania	**10**
	The Middle East & North Africa (MENA)	Egypt, Iran, Iraq, Turkey, Yemen	
Americas	North America	Canada, Cuba, Guatemala, Mexico, United States	**10**
	South America	Argentina, Brazil, Colombia, Peru, Venezuela	
Asia	Central Asia	Kazakhstan, Kyrgyzstan, Tajikistan, Turkmenistan, Uzbekistan	**20**
	East Asia	China, Hong Kong, Japan, South Korea, Taiwan	
	South Asia	Afghanistan, Bangladesh, India, Nepal, Pakistan	
	Southeast Asia	Indonesia, Myanmar, Philippines, Thailand, Vietnam	
Europe	Eastern Europe	Czechia, Poland, Romania, Russia, Ukraine	**10**
	Western Europe	France, Germany, Italy, Spain, United Kingdom	
Oceania	Oceania	Australia, Fiji, New Zealand	**3**
		Grand Total	**53**

E. Image Stacking Perturbation Pseudocode

This section provides the detailed pseudocode of procedure described in Section 3.3.1. The algorithm outlines how adversarial cues—flags and landmarks—are overlaid onto original cultural item images to generate perturbed samples. Each perturbation mode (*flag*, *landmark*, *flag_landmark*) follows a controlled compositing process defined in Algorithm 1.

Algorithm 1: Naive Image Stacking Perturbation

Input: $\mathcal{D}_{pair}, \mathcal{D}_{geo}, m \in \{\text{flag}, \text{landmark}, \text{flag\_landmark}\}, \mathcal{A}$
Output: $\mathcal{I}_S, \mathcal{L}$

foreach $(i, d_i) \in \mathcal{D}_{pair}$ **do**
  foreach $a \in \mathcal{A}$ **do**
    $c_{adv} \leftarrow \psi(d_i, a)$; $\quad \mathbf{g}_{adv} \leftarrow \gamma(\mathcal{D}_{geo}, c_{adv})$;
    $(\mathbf{I}_f, \mathbf{I}_l) \leftarrow \eta(\mathbf{g}_{adv})$;
    $(\mathbf{I}_{ori}, \mathbf{I}_f, \mathbf{I}_l) \leftarrow \nu(\mathbf{I}_{ori}, \mathbf{I}_f, \mathbf{I}_l, m)$;
    if $m \in \{$*flag*, *flag_landmark*$\}$ **then**
      $(w_f, h_f) \leftarrow \nabla(\mathbf{I}_f; \frac{1}{5}\mathbf{I}_{ori})$; place $\mathbf{I}_f$ at $\nearrow$;
    if $m \in \{$*landmark*, *flag_landmark*$\}$ **then**
      $(w_l, h_l) \leftarrow \nabla(\mathbf{I}_l; \frac{1}{5}\mathbf{I}_{ori})$; place $\mathbf{I}_l$ at $\swarrow$;
    $\mathbf{I}_S \leftarrow \Phi(\mathbf{I}_{ori}, \{\mathbf{I}_f, \mathbf{I}_l\})$; $\quad \mathcal{L} \leftarrow \mathcal{L} \cup \{\mathbf{I}_S\}$;
return $\mathcal{I}_S, \mathcal{L}$

F. Complete Evaluation Results

We present the complete results in Table 4. The key observations are consistent with the findings reported in the main paper, demonstrating that the conclusions generalize across all evaluated settings.

- **Proprietary VLMs outperform open-source models**, with the exception of Claude, which performs comparably to Qwen3-VL.
- **Flag perturbation** emerges as the dominant perturbation factor, exhibiting a stronger influence than landmark perturbations.
- **Generative perturbation** consistently outperforms image stacking across all cases, highlighting its effectiveness as a perturbation strategy.
- **Geographic (Geo) and Descriptive (Desc) proxies show similar patterns in all cases**, suggesting that both proxy-based semantic and geographic cultural cues remain relevant.

G. Inference Model Selection & Hyperparmeters

All models are evaluated with a temperature of 0.0 to ensure deterministic outputs. Structured output formatting is used throughout. For GPT and Gemini, structured JSON responses are enabled directly through their respective API features. For other models, we explicitly instruct the model to produce valid JSON outputs via prompt formatting. The open-source model checkpoints used are as follows:

- `OpenGVLab/InternVL3_5-38B`
- `OpenGVLab/InternVL3_5-8B`
- `OpenGVLab/InternVL3_5-4B`
- `Qwen/Qwen3-VL-30B-A3B-Instruct`
- `Qwen/Qwen3-VL-8B-A3B-Instruct`

Table 3. Dataset attribute descriptions and examples. License information is provided as dedicated rows for each URL-bearing field.

Attribute	Type	Description	Example
id	int32	Unique identifier for each record.	101
image	image	Main image associated with the item.	(Image file)
item	string	Name or title of the item (e.g., dish or object).	Eiffel Tower
origin_country	string	Country of origin for the item.	France
adversarial_country	string	Country used for adversarial comparison or challenge.	Germany
category	string	Category or type of the item.	Landmark
difficulty	string	Difficulty level of the task or challenge.	Medium
perturb_method	string	Method used to perturb or modify the image/data.	Gaussian Noise
landmark_name	string	Name of the landmark or key feature in the image.	Eiffel Tower
perturb_context	string	Context of the perturbation or transformation applied.	Added clouds and reduced brightness.
pair_method	string	Method used to pair original and perturbed images.	Nearest neighbor similarity
item_url	string	URL linking to more information about the item.	https://en.wikipedia.org/wiki/Eiffel_Tower
item_url_license	string	License for the content at item_url.	CC BY-SA 3.0
flag_url	string	URL to the flag image of the associated country.	https://upload.wikimedia.org/wikipedia/en/c/c3/Flag_of_France.svg
flag_url_license	string	License for the content at flag_url.	Public Domain
landmark_url	string	URL linking to the landmark's reference/source image.	https://commons.wikimedia.org/wiki/File:Eiffel_Tower_Paris.jpg
landmark_url_license	string	License for the content at landmark_url.	CC BY-SA 3.0

- `Qwen/Qwen3-VL-4B-A3B-Instruct`

For the proprietary models, we employed different settings for each variant to assess their impact on performance and robustness. As with the open-weight variants, the *sampling temperature is uniformly set to* 0 across all proprietary models to ensure deterministic output.

- `GPT-5-H+` (GPT-5): In this setting, we utilize *high-level thinking* and *high verbosity*. This configuration serves as the *upper-bound variant* to test the maximum reasoning capacity and its effect on the model's robustness toward the CONFUSEDTOURIST scenario.
- `GPT-5-H` (GPT-5): This configuration uses *high-level thinking* but with a *standard verbosity* setting.
- `GPT-5-L` (GPT-5): This configuration utilizes *low-level thinking* with *standard verbosity*.
- `GPT-5-m` (GPT-5): This setting uses *minimum-level thinking* and *standard verbosity*.
- `GPT-4.1` (GPT-4.1): This model is configured with *standard verbosity* and *standard-level thinking*.
- `G-2.5-Pro` (Gemini 2.5 Pro): For this variant, the model's *dynamic thinking budget is set to* -1 (denoting the highest available thinking budget) with *standard verbosity*.
- `G-2.5-flash` (Gemini 2.5 Flash): This model is configured to use *no internal thinking at all*, with the *dynamic thinking budget set to* 0, and operates with *standard verbosity*.
- `Sonnet-4.5` (Claude 4.5 Sonnet): This variant is tested using its *default configuration settings* as provided by the vendor.

All models are evaluated under identical inference settings, unless otherwise specified.

Table 4. Performance drop comparison across models under different perturbation settings. Columns indicate multilabel exact matching accuracy (%) for naive and AI-based perturbation methods across landmark (L), flag (F), and both (L+F) perturbation context infusion. Difficulty levels correspond to distance for Geo and semantic similarity for Desc as discussed in Sec. 3.2. Detailed inference hyperparameters and settings for each model are outlined in Appendix G.

		Without Perturb.	Naive						AI					
			Geo			Desc			Geo			Desc		
			L	F	L+F	L	F	L+F	L	F	L+F	L	F	L+F
Difficulty			Easy/Hard	Easy/Hard	Easy/Hard	Easy/Hard	Easy/Hard	Easy/Hard	Easy/Hard	Easy/Hard	Easy/Hard	Easy/Hard	Easy/Hard	Easy/Hard
Proprietary														
GPT-5 (High$^+$)	Country	66.7	65.9 / 61.6	62.6 / 49.3	61.6 / 48.8	64.7 / 59.2	57.4 / 54.5	57.9 / 56.4	49.8 / 49.8	46.9 / 41.2	46.0 / 42.7	47.7 / 41.7	43.4 / 34.6	44.7 / 37.0
	Item	57.6	54.5 / 53.1	54.0 / 47.9	51.2 / 49.3	54.9 / 49.3	51.9 / 50.2	51.9 / 46.9	42.2 / 49.8	40.8 / 39.8	42.7 / 40.3	45.5 / 37.4	40.4 / 34.1	36.2 / 33.2
GPT-5 (High)	Country	66.7	66.4 / 63.0	64.5 / 49.8	56.9 / 50.2	62.1 / 59.2	60.4 / 52.6	59.6 / 51.7	53.1 / 49.8	46.9 / 37.9	46.0 / 41.7	48.1 / 40.3	43.4 / 34.1	44.7 / 33.6
	Item	56.4	50.7 / 49.8	50.7 / 44.5	46.4 / 44.1	51.5 / 44.5	50.2 / 39.8	48.1 / 42.2	40.3 / 45.0	39.3 / 37.4	39.3 / 37.9	42.1 / 34.1	40.0 / 30.3	37.4 / 31.8
GPT-5 (Low)	Country	63.8	60.7 / 57.3	57.8 / 44.5	53.1 / 47.9	63.8 / 56.4	56.2 / 49.3	56.2 / 49.8	46.4 / 47.4	42.7 / 37.9	40.3 / 37.9	47.2 / 36.0	39.6 / 32.2	42.6 / 29.9
	Item	50.2	47.9 / 44.1	46.4 / 42.2	43.1 / 41.2	45.5 / 42.7	44.7 / 35.5	43.4 / 38.9	36.5 / 40.8	36.5 / 35.1	36.5 / 35.5	36.6 / 30.3	33.2 / 28.9	37.4 / 25.1
GPT-5 (Minimal)	Country	63.4	63.0 / 57.8	51.7 / 35.1	47.4 / 34.1	61.3 / 58.8	52.3 / 40.8	50.6 / 40.3	46.0 / 44.5	37.9 / 27.0	38.4 / 28.0	43.4 / 34.1	36.2 / 24.2	32.3 / 25.1
	Item	46.5	42.7 / 43.6	37.9 / 30.8	34.1 / 29.4	41.7 / 41.2	40.9 / 28.9	37.0 / 30.8	30.3 / 37.0	26.1 / 25.6	28.4 / 25.6	31.1 / 24.6	27.2 / 18.0	25.5 / 21.3
GPT-4.1	Country	65.0	66.8 / 57.3	55.0 / 31.3	50.2 / 29.9	60.9 / 55.9	55.3 / 45.0	51.5 / 42.2	45.0 / 37.0	31.3 / 19.4	28.4 / 21.3	43.0 / 31.3	33.2 / 22.7	29.8 / 20.9
	Item	57.2	57.8 / 54.0	48.8 / 40.8	46.0 / 39.8	54.0 / 50.2	51.1 / 40.3	48.1 / 37.4	40.8 / 38.9	28.9 / 29.9	27.5 / 28.4	37.9 / 28.0	31.1 / 21.3	28.9 / 18.5
Gemini-2.5-Pro	Country	65.4	65.4 / 55.0	63.5 / 34.1	60.2 / 32.7	62.1 / 60.2	59.6 / 48.8	57.9 / 47.4	56.4 / 46.0	43.6 / 24.2	46.4 / 29.9	51.9 / 44.5	43.4 / 32.7	46.4 / 32.7
	Item	65.8	62.6 / 62.6	60.2 / 47.4	58.3 / 46.9	63.4 / 55.9	59.6 / 49.8	56.6 / 46.0	54.0 / 54.0	46.0 / 41.7	46.0 / 46.0	51.1 / 45.0	44.7 / 34.1	46.8 / 35.5
Gemini-2.5-Flash	Country	66.7	64.0 / 52.1	55.9 / 30.3	55.0 / 27.6	62.8 / 58.8	53.4 / 43.3	51.5 / 45.0	51.2 / 44.3	41.7 / 20.9	40.3 / 24.6	48.1 / 40.5	46.0 / 30.3	41.0 / 32.2
	Item	64.2	59.7 / 57.3	58.8 / 47.4	56.4 / 48.1	59.0 / 54.0	54.7 / 46.2	51.9 / 44.5	48.3 / 50.0	44.1 / 35.1	42.7 / 40.3	49.4 / 38.1	45.5 / 34.1	42.7 / 33.6
Claude-4.5-Sonnet	Country	53.8	52.6 / 52.1	37.1 / 31.4	32.4 / 29.4	49.6 / 51.9	41.6 / 30.2	40.5 / 28.9	31.9 / 32.7	26.1 / 24.2	22.3 / 23.2	29.4 / 21.4	22.1 / 17.5	24.9 / 14.8
	Item	49.6	48.3 / 49.3	33.3 / 34.3	31.0 / 32.2	45.7 / 45.7	38.2 / 28.8	37.9 / 27.0	29.5 / 35.5	25.1 / 28.0	20.9 / 24.6	28.1 / 21.0	21.2 / 18.5	23.6 / 15.3
Open-Source														
Qwen3-VL-30B	Country	52.3	52.6 / 47.4	34.6 / 19.4	24.2 / 16.6	49.8 / 47.4	37.4 / 28.0	31.5 / 18.5	38.4 / 31.3	26.5 / 14.7	22.3 / 15.2	34.0 / 27.5	27.7 / 17.1	22.1 / 11.4
	Item	42.0	42.2 / 38.4	33.2 / 26.1	28.4 / 27.5	40.0 / 37.4	32.3 / 24.6	30.2 / 22.7	28.9 / 32.2	23.7 / 23.2	20.4 / 20.4	27.2 / 24.2	25.1 / 18.0	16.2 / 12.8
Qwen3-VL-4B	Country	49.8	49.3 / 46.9	25.6 / 18.5	24.6 / 15.2	46.0 / 45.5	32.8 / 23.7	27.2 / 21.3	37.0 / 37.4	18.5 / 13.7	23.7 / 17.5	32.8 / 29.9	18.3 / 16.6	22.6 / 14.7
	Item	32.9	30.3 / 33.2	24.2 / 24.6	23.7 / 23.7	31.1 / 27.5	26.0 / 20.4	22.6 / 19.9	27.0 / 26.1	20.9 / 18.0	21.3 / 19.9	23.8 / 21.3	16.6 / 16.6	16.6 / 13.3
Qwen3-VL-8B	Country	44.9	41.7 / 38.4	10.4 / 8.1	9.5 / 8.1	39.6 / 38.4	14.9 / 10.0	14.5 / 9.0	30.8 / 26.1	9.0 / 5.7	12.3 / 8.1	23.0 / 21.3	10.6 / 8.1	12.3 / 6.6
	Item	32.9	29.4 / 28.9	20.9 / 19.9	21.3 / 21.8	29.4 / 28.4	20.0 / 18.0	22.1 / 16.1	21.8 / 22.3	11.8 / 13.3	13.7 / 11.4	16.6 / 14.2	12.8 / 10.0	11.1 / 6.6
InternVL3.5-38B	Country	25.5	26.5 / 23.7	11.4 / 7.6	10.9 / 6.2	21.7 / 22.3	12.8 / 9.0	11.5 / 7.1	17.1 / 13.7	8.1 / 7.1	9.5 / 6.2	14.9 / 13.7	8.5 / 5.7	7.2 / 6.6
	Item	19.8	21.3 / 20.9	13.3 / 12.3	11.4 / 11.8	14.0 / 18.0	11.9 / 10.0	9.8 / 8.1	14.7 / 10.9	8.1 / 9.0	8.5 / 9.0	11.1 / 10.9	7.2 / 6.2	4.3 / 5.7
InternVL3.5-4B	Country	20.2	17.5 / 19.0	9.5 / 6.6	5.7 / 5.7	17.4 / 17.1	10.6 / 8.5	8.5 / 6.2	10.9 / 14.2	5.2 / 5.2	4.7 / 3.8	9.4 / 10.4	6.8 / 5.2	4.7 / 3.8
	Item	11.9	10.4 / 11.4	5.7 / 6.2	3.8 / 7.1	11.9 / 9.0	6.4 / 4.7	5.5 / 3.8	7.1 / 7.6	4.7 / 7.6	5.2 / 4.7	3.8 / 5.2	6.4 / 4.3	3.4 / 2.8
InternVL3.5-8B	Country	21.0	23.2 / 21.8	8.5 / 7.6	7.1 / 5.7	18.7 / 19.4	6.4 / 10.0	4.3 / 8.5	11.8 / 12.3	5.7 / 5.7	4.3 / 7.1	10.2 / 8.1	6.4 / 4.7	4.3 / 5.7
	Item	12.3	14.7 / 14.2	5.2 / 6.2	6.2 / 5.7	12.3 / 11.8	5.5 / 6.6	4.7 / 6.6	5.7 / 9.0	3.3 / 4.7	3.8 / 6.2	4.7 / 5.7	4.3 / 3.3	3.8 / 4.3

Table 5. Distraction rate (%) across models under different perturbation settings. Columns indicate the percentage of wrong predictions that were distracted by the adversarial country across landmark (L), flag (F), and both (L+F) perturbation context infusion.

Model	Naive						AI					
	Geo			Desc			Geo			Desc		
	L	F	L+F	L	F	L+F	L	F	L+F	L	F	L+F
Difficulty	Easy/Hard	Easy/Hard	Easy/Hard	Easy/Hard	Easy/Hard	Easy/Hard	Easy/Hard	Easy/Hard	Easy/Hard	Easy/Hard	Easy/Hard	Easy/Hard
Proprietary												
GPT-5 (High)	8.5 / 37.2	33.3 / 75.5	29.7 / 78.1	10.1 / 36.0	23.7 / 60.0	25.3 / 60.8	16.2 / 48.1	41.1 / 72.5	35.1 / 75.6	21.3 / 43.7	36.1 / 71.9	33.1 / 65.7
GPT-5 (Low)	8.4 / 26.7	34.8 / 66.7	30.3 / 67.3	11.8 / 32.6	20.4 / 58.9	22.3 / 62.3	22.1 / 46.8	41.3 / 69.5	31.7 / 70.2	21.8 / 41.5	40.1 / 72.0	38.5 / 65.5
GPT-5 (Minimal)	10.3 / 24.7	57.8 / 80.3	54.1 / 82.7	9.9 / 33.3	43.8 / 75.2	50.0 / 73.8	36.8 / 58.1	64.1 / 82.5	63.8 / 86.8	36.8 / 56.1	62.7 / 80.6	50.3 / 77.8
GPT-4.1	15.7 / 45.6	60.0 / 91.7	61.0 / 93.2	17.4 / 41.9	47.6 / 80.2	49.1 / 82.8	35.3 / 75.9	77.2 / 94.7	76.2 / 97.0	46.3 / 69.7	73.9 / 89.6	66.7 / 89.2
Gemini-2.5-Pro	18.4 / 55.4	57.0 / 81.6	65.3 / 82.2	20.7 / 49.4	52.8 / 74.8	54.4 / 78.4	37.9 / 69.2	68.3 / 85.6	69.0 / 84.3	34.7 / 60.0	66.1 / 78.2	52.2 / 81.8
Claude-4.5-Sonnet	9.0 / 35.6	67.4 / 85.4	65.5 / 86.6	10.3 / 34.0	55.1 / 75.5	55.1 / 80.0	39.9 / 52.8	65.4 / 85.0	57.9 / 79.0	40.4 / 56.4	68.9 / 79.9	59.4 / 77.0
Open-Source												
Qwen3-VL-30B	12.0 / 46.8	85.5 / 95.9	90.0 / 98.9	18.6 / 33.3	76.2 / 90.1	88.8 / 95.3	50.0 / 73.1	85.8 / 97.2	89.6 / 97.8	45.2 / 63.4	79.4 / 90.9	79.2 / 93.0
Qwen3-VL-4B	14.0 / 38.4	90.4 / 91.9	90.6 / 91.6	15.0 / 35.7	86.7 / 88.8	86.5 / 89.8	42.9 / 54.5	90.7 / 94.0	75.8 / 92.0	45.6 / 52.7	89.6 / 89.8	84.1 / 85.6
Qwen3-VL-8B	26.8 / 50.8	94.2 / 94.3	94.2 / 96.4	26.8 / 43.1	94.5 / 94.7	96.0 / 95.3	58.2 / 67.9	95.3 / 97.0	93.5 / 97.4	63.5 / 62.0	96.2 / 94.8	94.7 / 93.4
InternVL3.5-38B	12.9 / 37.9	47.6 / 63.1	59.6 / 74.7	23.4 / 28.0	72.2 / 58.9	79.8 / 66.8	36.6 / 46.7	56.2 / 70.4	76.4 / 72.7	41.5 / 51.1	80.0 / 62.8	82.6 / 72.6
InternVL3.5-4B	9.8 / 31.0	42.4 / 59.4	46.7 / 61.3	27.8 / 24.0	75.2 / 56.5	77.7 / 59.6	27.1 / 45.9	43.0 / 56.5	50.7 / 67.0	47.9 / 42.3	76.7 / 59.0	82.6 / 67.5
InternVL3.5-8B	6.8 / 26.1	50.3 / 56.4	55.6 / 63.8	19.4 / 24.1	80.5 / 62.1	82.7 / 66.3	25.8 / 42.2	44.7 / 64.8	62.9 / 76.0	44.1 / 44.3	75.9 / 59.7	82.7 / 75.4

H. Prompt Ablation Attempt

To mitigate hallucination and misgrounding issues observed during evaluation, we conducted a *prompt ablation study* aimed at reducing over-attention to perturbational cues. The core idea is to remove prompt tokens that may induce attention overload toward contextual or background signals, which can cause vision-language models (VLMs) to rely excessively on spurious visual elements (e.g., flags or landmarks). Instead, we introduced a more explicit focusing instruction directing the model to concentrate on the target object itself: `"Identify the traditional name and origin of the category in the image. Please identify based on the object itself, not from surrounding cues."`

This ablation was applied across all three cultural categories (attire, cuisine, and music). Figures 7–12 illustrate examples before and after prompt refinement. Empirically, the refined prompts often improved grounding, with attention maps showing stronger alignment to the main object

rather than the added adversarial features. However, this improvement was **inconsistent**: while several cases showed clearer localization and accurate predictions, others exhibited persistent confusion, suggesting that textual prompt control alone may not fully resolve multimodal bias.

Overall, the ablation results highlight that language-side intervention can partially reorient model attention toward the intended visual concept, but its stability varies across categories and perturbation contexts.

I. Perturbed Image Visual Examples

Figure 13 The results illustrate the effects of different perturbations. Notably, the generative perturbation preserves the original image content, ensuring that the main object remains the controlled variable, while only the background is altered.

J. Cultural Pairing Algorithm

Algorithm 2: Cultural Pairing Dataset Construction (Short)

Input: P_{cty} (countries), $\mathcal{C} = \{\text{Attire}, \text{Music}, \text{Cuisine}\}$, Top-5 items per (country, category), country centroids $(\text{lat}_c, \text{lon}_c)$.

Output: Culture pool P_c, Landmark pool P_l, Triplets $\mathcal{T} = \{(p_i, p_j, l)\}$.

$P_c \leftarrow \varnothing$, $P_l \leftarrow \varnothing$, $\mathcal{T} \leftarrow \varnothing$.
foreach $c \in P_{cty}$ **do**
  foreach $\kappa \in \mathcal{C}$ **do**
    add top-5 (c, κ) as $p = (\text{Item}, c, \text{Desc}, I, \text{lat}_c, \text{lon}_c)$ to P_c
  add $l_{c,k} = (\text{Landmark}, c, \text{Desc}, I, \text{lat}, \text{lon})$ for $k = 1..3$ to P_l
foreach $p \in P_c$ **do**
  $z_p \leftarrow E_{\text{mE5}}(\text{Desc}(p))$
foreach $\kappa \in \mathcal{C}$ **do**
  $S \leftarrow \{p \in P_c : \text{cat}(p) = \kappa\}$
  foreach $p_i \in S$ **do**
    $j^+ \leftarrow \arg\max_{j \neq i,\, p_j \in S} \frac{z_{p_i} \cdot z_{p_j}}{\|z_{p_i}\| \|z_{p_j}\|}$,
    $j^- \leftarrow \arg\min_{j \neq i,\, p_j \in S} \frac{z_{p_i} \cdot z_{p_j}}{\|z_{p_i}\| \|z_{p_j}\|}$
    $g^+ \leftarrow \arg\min_{j \neq i,\, p_j \in S} D_{\text{hav}}(p_i, p_j)$,
    $g^- \leftarrow \arg\max_{j \neq i,\, p_j \in S} D_{\text{hav}}(p_i, p_j)$
    sample $l^{(1)} \in P_l$ from countries of $\{p_i, j^+\}$; similarly $l^{(2)}, l^{(3)}, l^{(4)}$
    $\mathcal{T} \leftarrow \mathcal{T} \cup \{ (p_i, j^+, l^{(1)}), (p_i, j^-, l^{(2)}), (p_i, g^+, l^{(3)}), (p_i, g^-, l^{(4)}) \}$
return $P_c, P_l, \mathcal{T}$

Figure 7. Original attire prompt, which has right country (**wrong**).

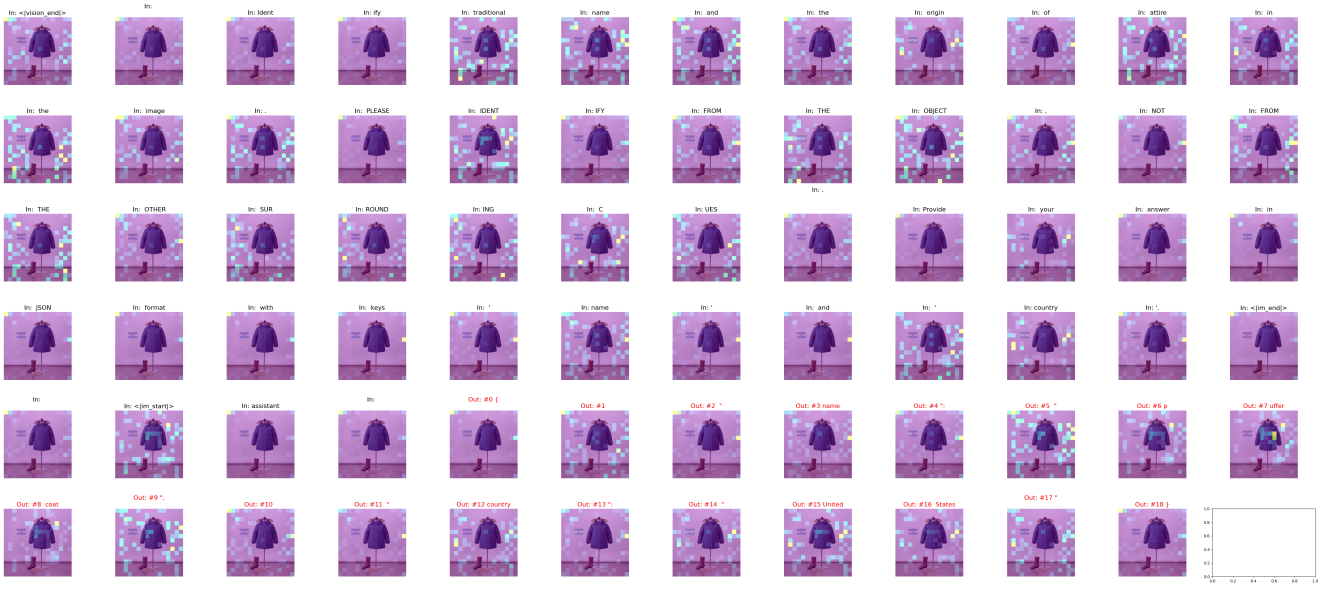

Figure 8. Refined attire prompt, this is instability case where now, instead of just country, both name and country is wrong. (**wrong**).

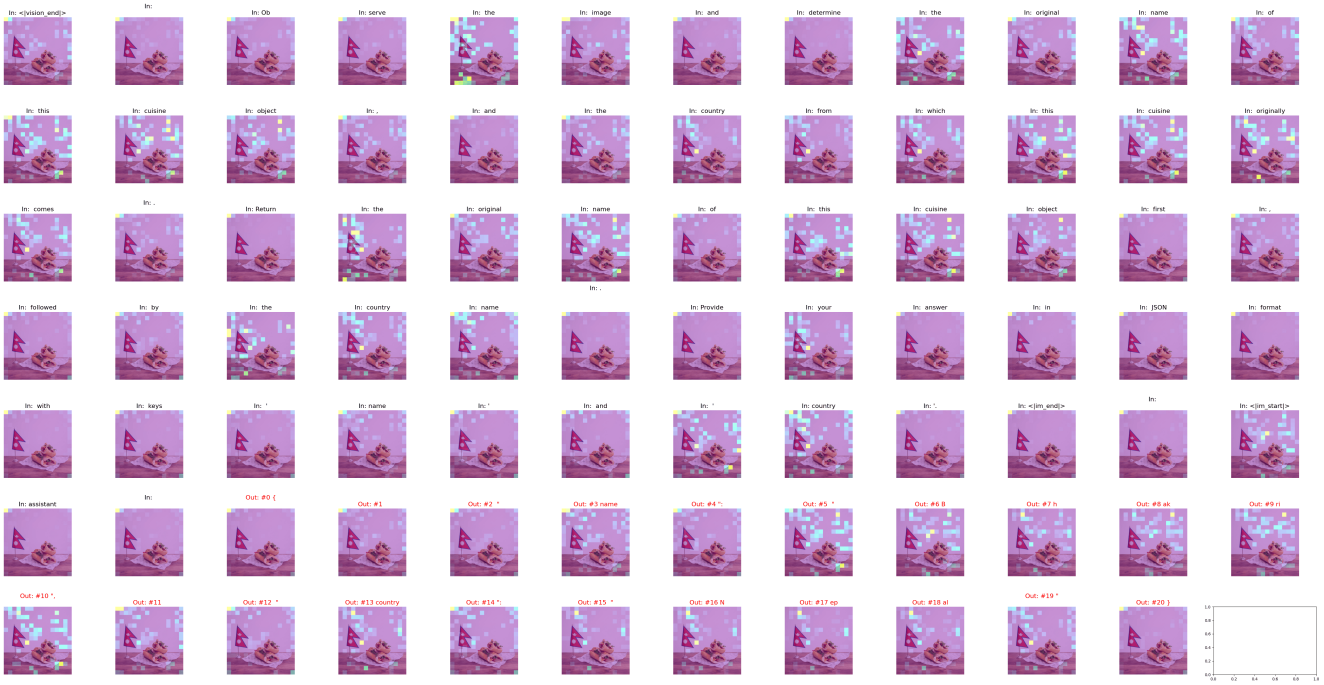

Figure 9. Original cuisine prompt (**wrong**).

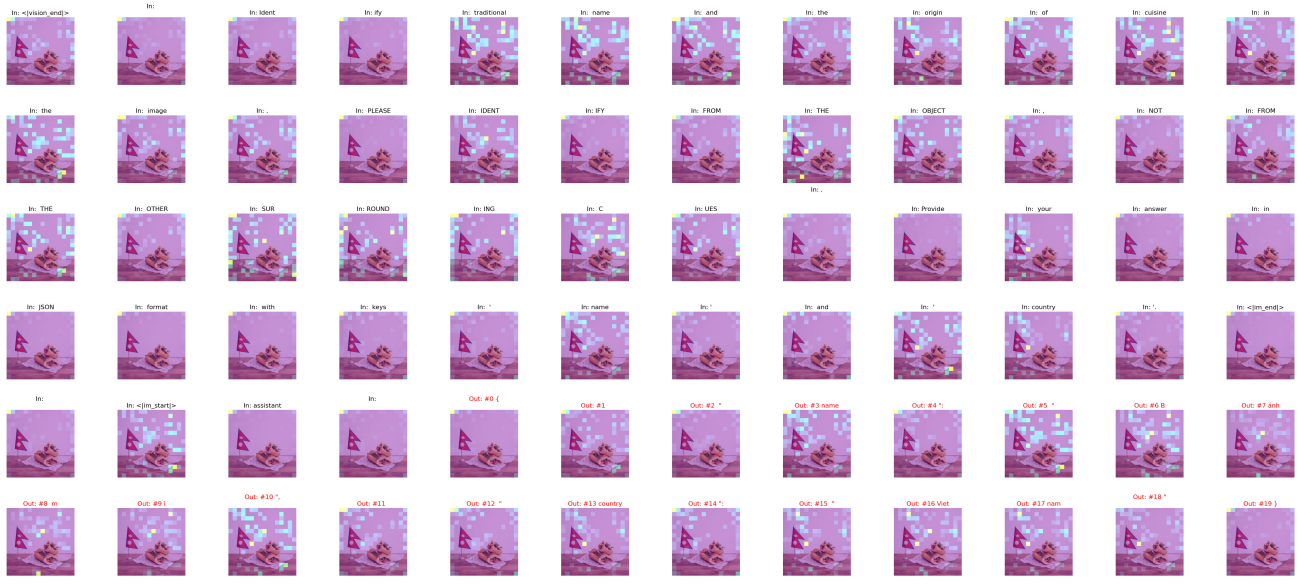

Figure 10. Refined cuisine prompt (**correct**).

Figure 11. Original music prompt (**wrong**).

Figure 12. Refined music prompt (**correct**).

Figure 13. Perturbed Image Visual Examples