Supplementary Material

Where Do Vision-Language Models Fail? World Scale Analysis for Image Geolocalization

This supplementary provides additional implementation details, qualitative analysis and additional results. We show qualitative comparison of semantically similar images identified as confusing samples (Sec. 1), biome label distributions and classification prompts (Secs. 2–3), Top-5 biome accuracy (Sec. 4), preliminary LoRA fine-tuning (Sec. 5), and full Urban/Rural and GER tables with per-cell standard deviations (Sec. 6).

1. Qualitative Analysis

Table 7 presents qualitative examples of neighboring-country confusion produced by **Qwen3-VL-4B-Instruct** under a blind two-turn prompting setup. Three image pairs are shown, each drawn from geographically close countries: (1) Pakistan vs. India, (2) Bangladesh vs. India, and (3) USA vs. Canada. For each pair, the model is given the same prompt asking it to predict the top-5 most likely countries for each image. The results are shown with color-coded annotations, where green indicates a correct top-1 prediction, red indicates an incorrect top-1, and yellow denotes cases where the ground truth appears within the top-5 but not at the top. The first pair, the model correctly identifies one country; in the second pair, it fails to correctly identify either; and in the third pair, it achieves one correct prediction and one partial success within the top-5. Overall, the examples demonstrate the model's difficulty in distinguishing visually similar neighboring regions.

Table 7. Neighbouring country confusion analysis using **Qwen3-VL-4B-Instruct** in a blind two-turn setting (Turn 1: predict; Turn 2: justify). Green = correct Top-1, Red = wrong Top-1, Yellow = GT in Top-5 but not Top-1.

	Pair 1: India / Pakistan		**Pair 2: India / Bangladesh**		**Pair 3: USA / Canada**	
	Pakistan	**India**	**India**	**Bangladesh**	**Canada**	**USA**
Images						
Prompt	*"I am showing you two images from two different locations in the world. For each image, predict the top 5 countries it is most likely from. Respond only in JSON format."*		*"I am showing you two images from two different locations in the world. For each image, predict the top 5 countries it is most likely from. Respond only in JSON format."*		*"I am showing you two images from two different locations in the world. For each image, predict the top 5 countries it is most likely from. Respond only in JSON format."*	
Predictions & Justification	**Top-1: India ✗** 1. India 2. Pakistan 3. United States 4. Australia 5. Canada *Stone arch + arid landscape → defaulted to India.*	**Top-1: India ✓** 1. India 2. Pakistan 3. United States 4. Australia 5. Canada *Ancient temple, stone fortification → India ranked #1.*	**Top-1: USA ✗** 1. United States 2. Canada 3. Mexico 4. Brazil 5. Argentina *Street sign + American-style lamppost misled model.*	**Top-1: India ✗** 1. India 2. Indonesia 3. Thailand 4. Philippines 5. Vietnam *Dense tropical forest → South/SE Asia.*	**Top-1: USA ✗** 1. United States 2. Canada 3. Australia 4. New Zealand 5. Mexico *Suburban homes + fire hydrant → defaulted to USA.*	**Top-1: USA ✓** 1. United States 2. Canada 3. Australia 4. New Zealand 5. Mexico *Fire hydrant + utility poles → USA ranked #1.*

2. Biome Label Distribution

Table 8 summarizes the distribution of consensus biome labels across the three datasets, considering only the samples where all three annotators-CLIP, SigLIP, and Qwen3-VL-4B-are in agreement. It reports the frequency of each biome category under this strict consensus criterion, providing an overview of the recognized environmental classes across datasets.

Table 8. Distribution of consensus biome labels where all three annotators (CLIP, Qwen, SigLIP) agree. CityGuessr-68k: 10,264 of 68,269 (15.0%); GeoGuessr-50k: 13,400 of 49,997 (26.8%); OSV5M: 18,001 of 50,000 (36.0%).

	GeoGuessr-50k		**CityGuessr-68k**		**OSV5M**	
Biome	**Samples**	**%**	**Samples**	**%**	**Samples**	**%**
Temperate	8,446	63.0	1,609	15.7	5,910	32.8
Mediterranean	1,514	11.3	4,710	45.9	96	0.5
Arid	1,646	12.3	3,115	30.3	8,304	46.1
Boreal	1,607	12.0	99	1.0	569	3.2
Tundra	109	0.8	683	6.7	785	4.4
Tropical	78	0.6	48	0.5	2,337	13.0

3. Biome Classification Prompts

Table 9 lists the text prompts used for zero-shot biome classification. For CLIP and SigLIP, each prompt is used as a text embedding and the highest cosine similarity biome is assigned. For Qwen3-VL-4B, the six category names and descriptions are provided directly and the model is asked to select one.

Table 9. Prompts used for zero-shot biome classification. CLIP and SigLIP use these as text embeddings for cosine similarity; Qwen3-VL-4B is prompted to select among these descriptions.

Biome	**Prompt**
Tropical	“a tropical rainforest or jungle scene”
Arid	“a dry desert or arid landscape”
Temperate	“a temperate forest or grassland scene”
Mediterranean	“a Mediterranean coastal or dry summer landscape”
Tundra	“a cold tundra, snow, or polar landscape”
Boreal	“a boreal forest or taiga with conifer trees”

4. Top-5 Biome Accuracy

Table 10 presents the Top-5 accuracy for nine Vision-Language Models (VLMs) across three datasets and six environmental biomes: Arid, Boreal, Mediterranean, Temperate, Tropical, and Tundra. The data highlights a dominant performance by the Qwen3-VL family, particularly the 2B and 4B variants, which achieve the highest scores in most categories. Conversely, the InternVL2.5 series shows accuracy steadily improving as model size increases. Finally, the results underscore that tropical regions are the most challenging to geolocalize across all models, whereas boreal regions often yield the highest accuracy due to more distinctive geographic cues.

Table 10. Top-5 accuracy (%) across biome categories using consensus labels (all three annotators agree). GeoGuessr-50k: n=13,400 (26.8%); CityGuessr: n=10,264 (15.0%); OSV5M: n=18,001 (36.0%). Best per-biome results are **bold**. [*L = LLaVA]

	GeoGuessr-50k (n=13,400)						**CityGuessr** (n=10,264)						**OSV5M** (n=18,001)					
Model	Arid	Bor.	Med.	Tmp.	Tro.	Tun.	Arid	Bor.	Med.	Tmp.	Tro.	Tun.	Arid	Bor.	Med.	Tmp.	Tro.	Tun.
InternVL2.5-1B	69.9	28.7	69.0	66.2	30.8	39.4	64.7	52.5	83.0	60.4	54.2	60.5	45.5	85.8	62.5	58.7	26.4	52.1
InternVL2.5-4B	67.6	71.4	76.4	73.0	38.5	45.0	74.1	63.6	83.7	62.6	52.1	65.3	45.4	90.7	61.5	58.5	26.4	47.6
InternVL2.5-8B	76.0	**93.3**	85.6	83.0	51.3	60.6	**82.2**	65.7	84.1	66.4	58.3	70.7	49.5	96.5	76.0	61.4	25.7	61.3
*L-Mistral-7B	48.7	45.6	60.2	62.8	37.2	44.0	63.8	15.2	55.3	47.0	31.2	40.3	33.5	88.2	57.3	43.2	16.3	54.3
*L-Vicuna-7B	57.4	69.4	46.0	61.3	11.5	39.4	55.2	46.5	57.6	52.5	37.5	40.4	34.6	87.3	41.7	48.0	9.8	30.2
*L3-LLaVA-8B	61.8	87.7	69.2	69.5	43.6	48.6	55.3	57.6	65.2	59.1	37.5	54.9	42.1	96.0	54.2	53.2	21.1	52.6
Qwen3-VL-2B	**81.4**	87.2	88.1	85.0	67.9	71.6	78.2	64.6	87.5	71.1	**60.4**	71.7	60.3	98.1	**80.2**	**72.0**	**36.5**	71.8
Qwen3-VL-4B	80.1	93.2	**89.8**	**87.1**	**70.5**	**76.1**	80.8	65.7	**89.2**	**74.8**	**60.4**	78.8	**61.6**	**98.4**	78.1	**72.0**	33.8	70.6
Qwen3-VL-8B	66.9	72.1	83.8	77.8	64.1	**76.1**	81.5	**66.7**	88.9	74.6	54.2	**83.6**	57.7	97.0	**80.2**	69.8	32.0	**75.5**

5. Preliminary LoRA Fine-Tuning

To assess parameter-efficient fine-tuning, we fine-tune InternVL2.5-1B using LoRA (r=8, α=32, all-linear modules) on a top-20 country subset of GeoGuessr-50k (12,116 train / 3,030 test images), training for 2 epochs at 5×10^{-5}. Top-1 accuracy improves from 40.40% to **79.44%** (+39.04 pp). The gain is larger for rural scenes (+45.96 pp) than urban (+25.70 pp), suggesting the zero-shot model's rural weakness is a data-coverage issue addressable by fine-tuning. LoRA also improves error quality: GER-Weak rises from 17.55% to 39.17%, indicating the fine-tuned model's remaining errors are over twice as likely to be visually justified. Full exploration of larger ranks and scales is left to future work.

6. Full Urban/Rural and GER Tables

Table 11 extends Table 4 (main paper) with per-cell standard deviations. Urban accuracy is highly stable across labellers (std $<$ 1 pp for most cells), while rural variance reaches 5–6 pp, confirming that annotator disagreement concentrates on the rural boundary. Top-5 gaps are consistently narrower than Top-1, e.g., Qwen3-VL-4B's urban–rural difference on GeoGuessr-50k compresses from $\sim$15 pp (Top-1) to $\sim$8 pp (Top-5).

Table 11. **Urban/Rural** Top-1 and Top-5 accuracy (%), reported as mean $\pm$ std across three independent labellers (CLIP, SigLIP, Qwen3-VL-4B). The maximum standard deviation across all cells is 6.66 pp. [*L = LLaVA]

	GeoGuessr-50k				**CityGuessr**				**OSV5M**			
	Top-1		Top-5		Top-1		Top-5		Top-1		Top-5	
Model	Urb	Rur	Urb	Rur	Urb	Rur	Urb	Rur	Urb	Rur	Urb	Rur
InternVL2.5-1B	48.81±.43	27.73±1.21	76.14±.55	59.76±1.39	46.93±.64	30.06±3.48	68.07±.49	46.44±4.25	33.57±5.96	30.15±.98	54.69±6.55	51.22±1.28
InternVL2.5-4B	69.29±.85	46.71±1.57	84.23±.54	67.54±1.25	57.60±.71	34.89±4.46	71.16±.53	50.20±4.19	36.38±6.57	31.66±1.20	54.35±6.39	50.25±1.16
InternVL2.5-8B	74.47±.34	58.35±1.34	88.29±.23	78.67±.87	59.61±.58	38.16±3.72	72.74±.49	52.48±4.21	38.24±6.08	36.26±1.23	56.67±5.39	54.77±1.06
*L-Mistral-7B	46.53±1.81	34.36±1.64	59.72±1.41	49.43±1.26	31.95±.61	19.16±4.55	48.75±.58	28.31±4.62	19.60±4.89	17.38±1.08	33.93±5.06	36.02±1.19
*L-Vicuna-7B	42.76±2.11	30.73±1.82	58.99±1.40	50.03±1.27	38.35±.69	23.75±4.85	53.16±.66	32.95±5.12	22.26±5.00	18.82±1.09	38.15±5.79	36.26±1.26
*L3-LLaVA-8B	59.85±.55	42.86±1.41	74.89±.14	64.65±1.05	40.73±.53	26.13±4.45	58.75±.42	40.25±4.23	28.53±5.17	28.00±1.16	46.25±5.73	46.98±1.29
Qwen3-VL-2B	82.56±.43	67.78±1.14	92.04±.22	82.98±.78	67.25±.69	43.27±4.77	79.28±.46	61.07±2.51	50.96±6.23	46.07±1.14	68.68±4.55	64.01±.77
Qwen3-VL-4B	83.83±.49	68.49±1.16	93.38±.18	85.03±.75	68.84±.78	41.54±5.21	80.10±.45	61.80±3.08	51.26±6.30	44.80±1.06	68.31±4.25	64.54±.79
Qwen3-VL-8B	79.82±.79	58.56±1.58	89.50±.64	73.70±1.20	68.31±.86	41.80±4.75	80.29±.54	63.87±2.65	48.23±6.66	41.05±1.10	65.77±5.27	62.63±.98

Table 12 supplements Table 4 (main paper) with per-cell standard deviations across the CLIP and SigLIP neighbour sets. The low maximum standard deviations (1.44 pp for GER-W, 0.85 pp for GER-S) confirm that GER rankings are robust to the choice of embedding model. Weaker models show higher relative variance (e.g., LLaVA-Mistral-7B: 13.25 ± 1.44 GER-W on GeoGuessr-50k), while the Qwen3-VL family is near-identical across both backbones (e.g., 17.18 ± 0.04 on OSV5M).

Table 12. **Geographic Error Reasonableness (GER)** reported as mean ± std across CLIP and SigLIP nearest neighbours ($K=5$); maximum SD across all cells is 1.44 pp (GER-W: 1.44, GER-S: 0.85). GER-W: predicted country in $\geq$1 of 5 neighbours' ground-truth countries; GER-S: in $\geq$2. Computed over incorrect Top-1 predictions only (%). [*L = LLaVA]

	GeoGuessr-50k		**CityGuessr**		**OSV5M**	
Model	GER-W	GER-S	GER-W	GER-S	GER-W	GER-S
InternVL2.5-1B	15.98±.53	7.18±.28	5.32±1.15	1.98±.43	11.52±.42	4.70±.15
InternVL2.5-4B	21.34±1.27	11.43±.65	7.34±1.38	2.88±.50	11.54±.55	5.00±.25
InternVL2.5-8B	33.72±.63	18.65±.30	7.34±1.29	2.94±.48	14.08±.32	6.22±.10
*L-Mistral-7B	13.25±1.44	6.97±.75	3.16±.76	1.17±.29	5.69±.39	2.42±.19
*L-Vicuna-7B	10.67±1.25	5.70±.59	3.10±.67	1.18±.31	5.53±.39	2.45±.20
*L3-LLaVA-8B	21.41±1.44	10.80±.85	3.95±.90	1.44±.32	10.02±.59	4.14±.28
Qwen3-VL-2B	40.48±.49	23.19±.47	7.84±.94	3.06±.29	17.18±.04	7.56±.03
Qwen3-VL-4B	**44.37±.45**	**25.92±.76**	**8.30±1.04**	**3.34±.43**	**17.14±.14**	**7.77±.02**
Qwen3-VL-8B	30.22±.58	16.96±.44	7.79±.95	3.11±.40	15.35±.19	6.73±.13