World in a Frame:
Understanding Culture Mixing as a New Challenge for Vision-Language Models

Supplementary Material

A. Pipeline Details of CultureMix

A.1. Country Lists and Statistics

Table S1 summarizes, for each country, the number of food dishes and their names. Figure S1 visualizes the distribution of food combinations in MF, showing that the combinations are diverse and well balanced.

A.2. Human Validation and Statistics

We incorporated a human-in-the-loop process throughout all stages of dataset construction. Human annotators validated the generated images, and those that did not meet the quality criteria were regenerated—either using the same model or a more advanced one—followed by another round of human validation. This iterative process was repeated until the final dataset was completed. The criteria used for validation and the statistics of filtered cases are as follows.

Generating Single Food (SF) Images

1. **Background Removal**
   (a) If the image contains *text, hands, people, or tableware*, these elements are removed using either a diffusion model or manual methods.
2. **Human Validation–Comparison with the original image** (manually checked by the author using an in-house platform)
   (a) Criteria (Error Types)
   - Image differs from the original. → *Regenerate*
   - Food item is cropped. → *Regenerate*
   - Tableware appears in the image. → *Regenerate*
   - More than one food item appears. → *Regenerate*
   - Text appears on the food. → *Manually hide the text*
   - Reference food itself is inappropriate (*e.g.*, food placed on a cauldron, making it unsuitable for table placement) → *Replace food image*
   - Other cases where the image appears unnatural → *Replace food image*

   (b) Statistics
   - Regeneration cases: 109 / 295
   - Food image replacement cases: 12 / 295
   - Images with text present: 1 / 295

Generating Multiple Food (MF) Images

1. **Image Generation**
   - Concatenate the two images generated in Step 2 to create a MF image.

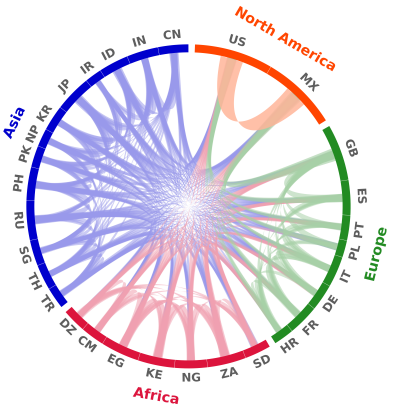


Figure S1. **Visualization of food combinations in MF and MFB.**

2. **Human Validation**
   (a) Criteria (Error Types)
   - Is the food item cropped? → *Regenerate*
   - Are the two food items excessively unbalanced in size? → *Regenerate*

   (b) Statistics
   - Regeneration cases: 74 / 948

Single Food with Background (SFB) and Multi-Food with Background (MFB) images

1. **Background Image Selection**
   - For each category (Street, Landmark), five background images were manually selected from different continents.
2. **Input Image Generation**
   - Concatenate the MF images beneath the selected background image.
3. **Editing**
   - Example prompt (the first two versions were used when initial generations failed):
4. **Human Validation**
   (a) Criteria
   - Background not generated (showing only the food item) → *Regenerate*
   - Food placed unnaturally (*e.g.*, floating in the air or positioned upright as if it would spill). → *Regenerate*
   - Food or background differs from the original. → *Regenerate*

   (b) Statistics. We repeated multiple rounds of refinement. After three attempts, most samples were collected as valid, with only a few remaining

idx	Country	Count	Food1 Names
1	United States Of America (USA)	19	1. Baked beans; 2. Carolina-style pulled pork/barbecue; 3. Chocolate chip cookie; 4. co-shon duh lay; 5. Dutch letter; 6. Eggs Benedict; 7. Hamburger; 8. Jum-buh-lie-ahh; 9. Key Lime Pie; 10. Loco moco; 11. Mochi ice cream; 12. Pecan pie; 13. Pepper jelly; 14. Pie à la Mode; 15. Pumpkin pie; 16. Rainbow cookie; 17. Southern fried chicken; 18. Spaghetti and meatballs; 19. Tootsie Roll
2	China	10	1. Century egg; 2. Cong you bing; 3. Edamame; 4. Lo mai gai; 5. Mantau; 6. Osmanthus cake; 7. Paper wrapped cake; 8. Yangzhou Fried Rice; 9. Yin yang fried rice; 10. Yong tau foo
3	France	10	1. Beef bourguignon; 2. Bouneschlupp; 3. Chouquette; 4. Croissant; 5. Gâteau Basque; 6. Ladyfinger; 7. Nun's puffs; 8. Ratatouille; 9. Sablé; 10. Teurgoule
4	Indonesia	10	1. Bakpia; 2. Kue mangkok; 3. Kue putu; 4. Laksa; 5. Lakso; 6. Nasi Uduk; 7. Sayur asem; 8. Siomay; 9. Tahu campur; 10. Tahu sumedang
5	Kenya	10	1. Bacella alba; 2. Biegnets; 3. Cha-pa-ti or Cha-poh; 4. Githeri; 5. Karaage; 6. Kenyan kachumbari salad; 7. Maize and beans stew; 8. Matoke; 9. Mukimo; 10. Pound maize flour
6	Mexico	10	1. Chilaquiles; 2. Chile relleno; 3. Huarache; 4. Migas; 5. Paste (pasty); 6. Piñata cookie; 7. Puchero; 8. Refried beans; 9. Salsa verde; 10. Sope
7	Philippines	10	1. Aparon; 2. Beef Steak; 3. Laing; 4. Oyster omelette; 5. Puto; 6. Rendang; 7. Silog; 8. Sushi Bake; 9. Taho; 10. Uraro
8	Russia	10	1. Alexandertorte; 2. Bracken fern salad; 3. Mimosa salad; 4. Potato pancake; 5. Pozharsky cutlet; 6. Shchi; 7. Solyanka; 8. Ukha; 9. Vinegret; 10. Zefir
9	South Africa	10	1. African spinach; 2. Bunny chow; 3. Chicken and mushroom pie; 4. Hertzoggie; 5. kota, skhambane; 6. Malva pudding; 7. Melktert; 8. Mopane stew; 9. Potjiekos; 10. tripe
10	Spain	10	1. Andrajos; 2. Arròs negre; 3. Cocido lebaniego; 4. Cocido madrileño; 5. Ensaïmada; 6. Escudella; 7. Hamin; 8. Hornazo; 9. Panellets; 10. Tortillitas de camarones
11	Italy	9	1. Cannoli; 2. Casoncelli; 3. Cavallucci; 4. Cotoletta alla milanese; 5. Florentine biscuit; 6. Gelato; 7. Michetta; 8. Piadina romagnola; 9. Piccata
12	Japan	9	1. Amanattō; 2. Char siu; 3. Dorayaki; 4. Fried ice cream; 5. Melonpan; 6. no; 7. Omurice; 8. Simmered dried strips of daikon radish; 9. Tempura
13	United Kingdom	9	1. Bedfordshire clanger; 2. Blackberry pie; 3. Empire biscuit; 4. Ham and cheese sandwich; 5. Pease pudding; 6. Saveloy; 7. Spotted dick; 8. Stornoway black pudding; 9. Treacle sponge pudding
14	Egypt	8	1. Custard tart; 2. Falafel; 3. Koshary; 4. Molokhia; 5. mombar; 6. Qatayef; 7. Um ali
15	India	8	1. Bhel puri; 2. Curd rice; 3. Neer dosa; 4. Panta bhat; 5. Pulihora; 6. Rice and curry; 7. Thalassery biryani; 8. Unnakai
16	Korea	8	1. Bibim guksu; 2. Bibimbap; 3. Japchae; 4. Jeon; 5. Milmyeon; 6. Sundubu-jjigae; 7. Tteok-bokki; 8. Tteokguk
17	Nigeria	8	1. Bridie; 2. Cooked cassava flakes and Okra soup; 3. E/ku/ru; 4. Ekwang; 5. Lupis; 6. Moin moin; 7. Okra Soup; 8. Vegetable soup with Egusi
18	Portugal	8	1. Aletria; 2. Bacalhau à Gomes de Sá; 3. Cabidela; 4. Cebolada; 5. Mocotó; 6. Pastel de Tentúgal; 7. Polvo à lagareiro; 8. Tripas à moda do Porto
19	Algeria	7	1. Algerian Almond Cookies; 2. Algerian Mekrout; 3. Besbousa; 4. Chermoula; 5. Harira; 6. Tamina; 7. Tikerbabin
20	Germany	7	1. Edi-kang Ikong; 2. Eisbein; 3. Kaiserschmarrn; 4. Maultasche; 5. Poppy seed roll; 6. Toast Hawaii
21	Thailand	7	1. Khanom bueang; 2. Khao kan chin; 3. Khao soi; 4. Kluai khaek; 5. Mi krop; 6. Pad thai; 7. Sakhu sai mu
22	England	6	1. Bacon and egg pie; 2. Bakewell tart; 3. Bath bun; 4. Curry pie; 5. Fish pie; 6. Steak pie
23	Iran	6	1. Baghali polo; 2. Chelow; 3. Kashk bademjan; 4. Mirza ghassemi; 5. Reshteh khoshkar; 6. Sajji
24	Cameroon	5	1. Achu; 2. leaves of gnetum; 3. Plantain Chips; 4. Puff Puff, beans and pape; 5. rice with beans
25	Croatia	5	1. Brudet; 2. Grahova pretepena juha; 3. Međimurska gibanica; 4. soparnik; 5. Zagorski štrukli
26	Nepal	5	1. Chataamari; 2. Chhurpi; 3. Gajar ka halwa; 4. Kwati; 5. Sapu Mhicha
27	Pakistan	5	1. Amba; 2. Bun kebab; 3. Momo; 4. Phirni; 5. Zarda
28	Singapore	5	1. Banmian; 2. Bihun Goreng; 3. Chwee kueh; 4. Noodles with tomato egg sauce; 5. Turnip cake
29	Turkey	5	1. Bamia; 2. Cezerye; 3. Kuru fasulye; 4. Qurabiya; 5. Şöbiyet
30	Poland	4	1. cabbage rolls; 2. Dumplings; 3. Gołąbki; 4. Krumiri
31	Sudan	4	1. Cucumber salad with yogurt; 2. Khachapuri; 3. LoQeymat or zalabia; 4. Vanille cake

Table S1. **Unique food counts and names per country included in CultureMix.**

Figure S2. **An example of input images for diffusion model-based image editing.**

invalid. These remaining cases were not verified through the platform but were instead visually inspected and regenerated until satisfactory.
1st attempt: Valid/Total = 0.72
2nd attempt: Valid/Total = 0.70
3rd attempt: Valid/Total = 0.88

A.3. Prompts and Model Configurations for Synthetic Dataset Generation

We provide details on how the image editing models were used to generate `SF`, `MF`, `SFB`, and `MFB` images. The specific model names and configurations are summarized in Table S2, and the corresponding prompts are provided below. Note that we used minor prompt variants (*e.g.*, reordering sentences, substituting synonymous verbs) to regenerate images when the initial output failed to satisfy the required criteria of the human annotators. An example of input Images for the diffusion model is shown in Figure S2.

For transparency, we also provide qualitative examples of image synthesis failures that were filtered out during human validation (Figure S3). These cases illustrate scenarios where the generated images are incomplete or misaligned with the given textual prompt. Based on our results, our synthetic images provide useful insights into the generation pipeline and can help guide future efforts in constructing culture-mixing datasets.

Prompts for `SF` Image Generation

- Leave all the food quality the exact same as the original.
- Modify the background image to pure white.
- Make the image square. Change the food into a top-down view.
- Convert the food to a top-down view.
- Remove any spoons, chopsticks, and human hands.
- Add any missing parts of a plate or a bowl.
- The plate or bowl should be circular or oval.

Prompts for `SFB` Image Generation (FLUX)

- Change the white background underneath a single food item to a table or picnic mat.
- The table or picnic mat should seamlessly be integrated with the background image.
- Rotate the food items along the z-axis so they are viewed from a natural dining perspective — not from the top, but more like how someone sees the plate while sitting at a table, and add realistic shadows to the plates.
- Remove any hands and utensils from the plates.
- Reconstruct the plate if there isn't any or if it's broken.
- Keep the background and the food quality identical to the original.
- The generated image should have only one food item.

Prompts for `SFB` Image Generation (QWEN)

- Leave all the elements the exact same as the original except for the following:
- Add a table or picnic mat underneath the food item.
- The table or picnic mat should seamlessly be integrated with the background image.
- Rotate the food items along the z-axis so they are viewed from a natural dining perspective — not from the top, but more like how someone sees the plate while sitting at a table, and add realistic shadows to the plates.
- Remove any hands and utensils from the plates.
- Reconstruct the plate if there isn't any or if it's broken.
- The generated image should have only one food item.

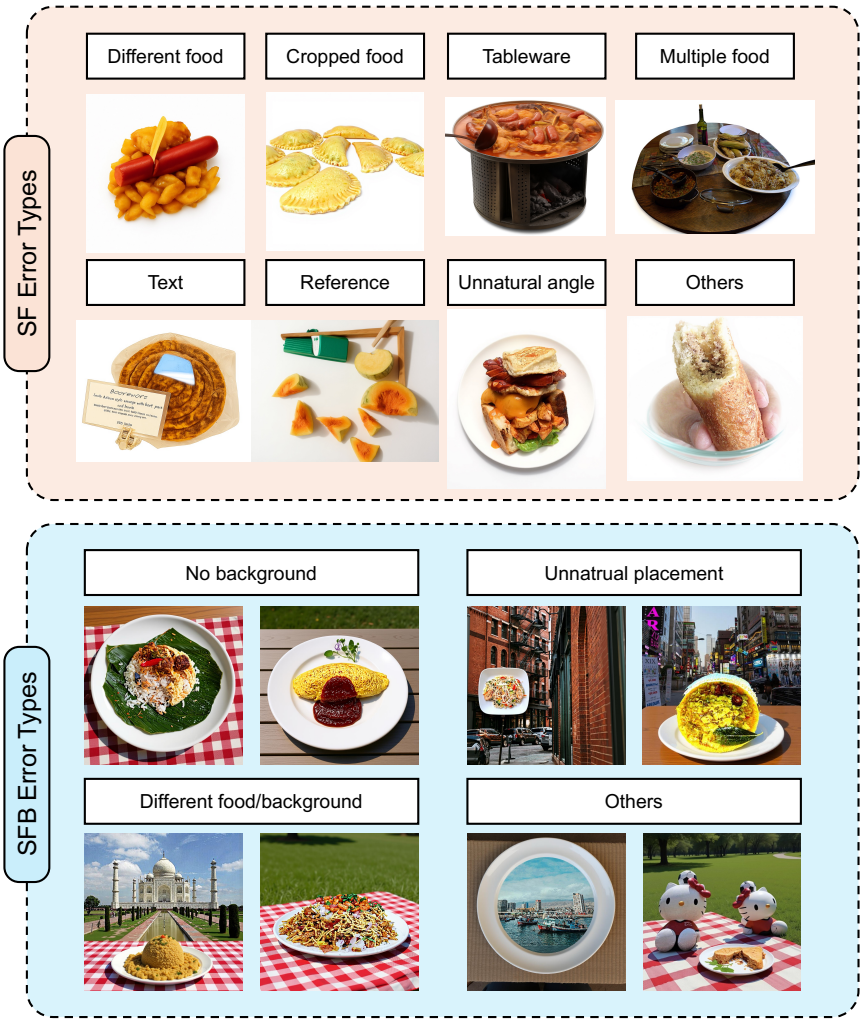


Figure S3. **Sample `SF` and `SFB` error images**. These images did not meet the human evaluation criteria and illustrate common failure patterns observed during the filtering process.

Prompts for `MFB` Image Generation (FLUX)

- Change the white background underneath two food items to a table or picnic mat.
- The table or picnic mat should seamlessly be integrated with the background image.
- Rotate the food items along the z-axis so they are viewed from a natural dining perspective — not from the top, but more like how someone sees the plate while sitting at a table, and add realistic shadows to the plates.
- Remove any hands and utensils from the plates.
- Reconstruct the plate if there isn't any or if it's broken.
- Keep the background and the food quality identical to the original.
- The generated image should have only two food items.

Prompts for `MFB` Image Generation (QWEN)

- Leave all the elements the exact same as the original except for the following:
- Add a table or picnic mat underneath the food item.
- The table or picnic mat should seamlessly be integrated with the background image.
- Rotate the food items along the z-axis so they are viewed from a natural dining perspective — not from the top, but more like how someone sees the plate while sitting at a table, and add realistic shadows to the plates.
- Remove any hands and utensils from the plates.
- Reconstruct the plate if there isn't any or if it's broken.
- The generated image should have only two food items.

Table S2. **Models and Configurations.** Overview of the models and their specific settings used for image synthesis.

Category	Details
Models	`black-forest-labs/FLUX.1-Kontext-dev` `Qwen/Qwen-Image-Edit`

Configs	`guidance_scale = 2.5`
	`num_inference_steps = 50`
	`torch.bfloat16`

A.4. Sample Dataset Images

The complete background image set, consisting of 25 landmark images and 25 street images, is shown in Figure S4 and Figure S5, respectively.

Final Image Generation Samples. Figure S6 presents examples of generated culture-mixing images from our constructed dataset. These samples illustrate the diversity of cross-cultural compositions produced during the construction of our dataset, illustrating variations in food types and cultural geographic backgrounds. These images reflect the range of visual configurations used to evaluate how LVLMs interpret cultural signals when multiple cultural elements coexist within a single image.

Real-World Image Samples. To complement generated images, we also collect a real-world image set featuring naturally occurring food pairings (Figure S7). We categorize half of the collected images into same-country pairs and the other into cross-country pairs. Images were also collected based on a variety of visual complexities, reflecting common sources of variation in real photographs such as mixed lighting, cluttered environments, irregular plating, and inconsistent camera perspectives.

B. Experiments

B.1. Results

Country and Food Name Identification Accuracy. Table S4 reports each model's country and food name identification accuracy across subtasks, providing additional detail that complements the radar chart in Figure 3a in the main text. Figure S8 visualizes how prediction correctness shifts from `SF` to the mixed subtasks (`SFB`, `MF`, `MFB`), showing that culturally mixed contexts often cause models to fail on cases they initially predicted correctly. Also, Figure S9 repeats Figure 3b in the main text with a larger size for closer inspection. Additionally, Table S5 compares the models' country and food name identification accuracy across landmark and street backgrounds. The performance difference between these two background types was minimal for both `SFB` and `MFB`. Figure S10 compares the country, food name identification accuracy, and entropy according to different cultural distances between the target and the distractor, which complements Figure 5 in the main text.

Country and Food Name Identification Entropy. Table S6 reports each model's country and food name identification entropy across subtasks, providing additional detail that complements Figure 6 in the main text.

Real World Dataset. Table S8 reports each model's country and food name identification accuracy of the real-world dataset, providing additional detail that complements the radar chart in Figure 8 in the main text.

B.2. Ablations

Positional Bias. To examine the effect of positional bias with respect to food location on the model performance, we randomly sample 100 multi-food images and compare the predicted labels before and after horizontally flipping each image. As shown in Table S9, the high consistency between the original and flipped predictions indicates that the model's outputs are stable under left–right reversals, suggesting minimal to no positional dependence for food-related attributes.

Size Bias. We also investigate the effect of the food item sizes on the model performance by first comparing the relative sizes of food items appearing on the left and right within multi-food images and then evaluating whether differing size ratios lead to changes in the predicted labels. Table S10 demonstrates that the model exhibits minimal size-related bias when identifying food items. In other words, even when one food item is noticeably larger than the other, the model's identification accuracy remains stable, suggesting that its predictions are largely invariant to object size differences.

B.3. Qualitative Analysis on Food Name Prediction Failure Cases

We sampled 25 instances of incorrect food-name predictions from each subtask (`SF`, `MF`, `SFB`, `MFB`) for Gemini-2.5-pro and InternVL3-8B, resulting in 100 images per model. We then manually checked whether the model's predicted food label visually resembled the ground-truth food (i.e., the model confused it with a similar-looking dish) or whether it was entirely unrelated, using Google Image search results for the predicted food as reference.

Gemini-2.5-pro predicted a visually similar food in 25 out of 100 cases, whereas InternVL3-8B did so in only 2 out of 100. This indicates that both models often predict completely different foods, but between the two, Gemini-2.5-pro tended to make closer guesses, whereas InternVL3-8B's predictions showed little resemblance to the target food. Figure S11 shows examples of similar food prediction for each model.

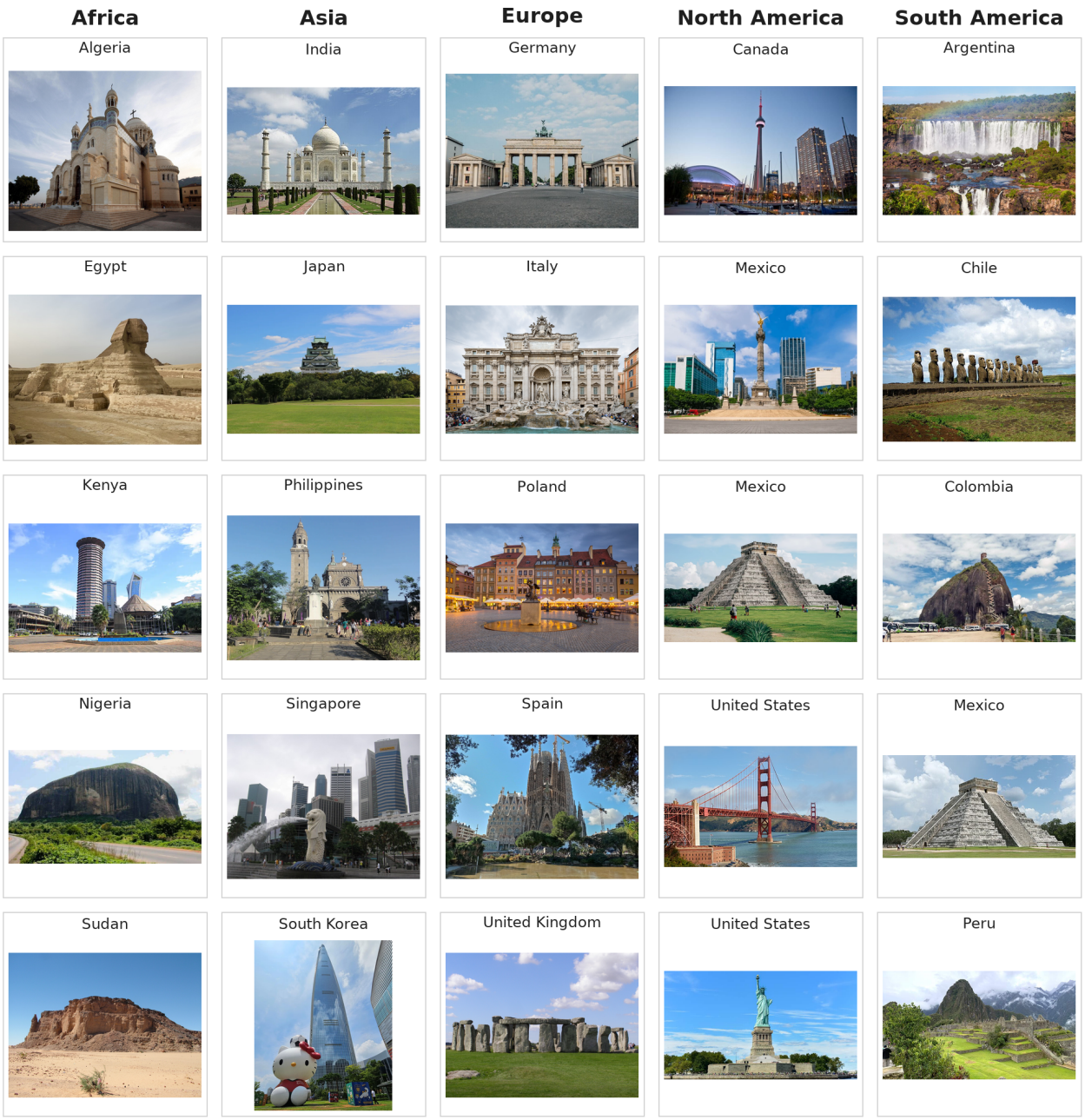


Figure S4. **Background (landmark) image samples.**

C. Mitigation Strategy

To validate whether our mitigation findings also hold for a larger open model, we evaluate InternVL3-38B (Table S7). Due to limited compute, we only test the prompting-based mitigation setting. The results show a similar trend to our earlier findings: simple prompting provides only modest mitigation.

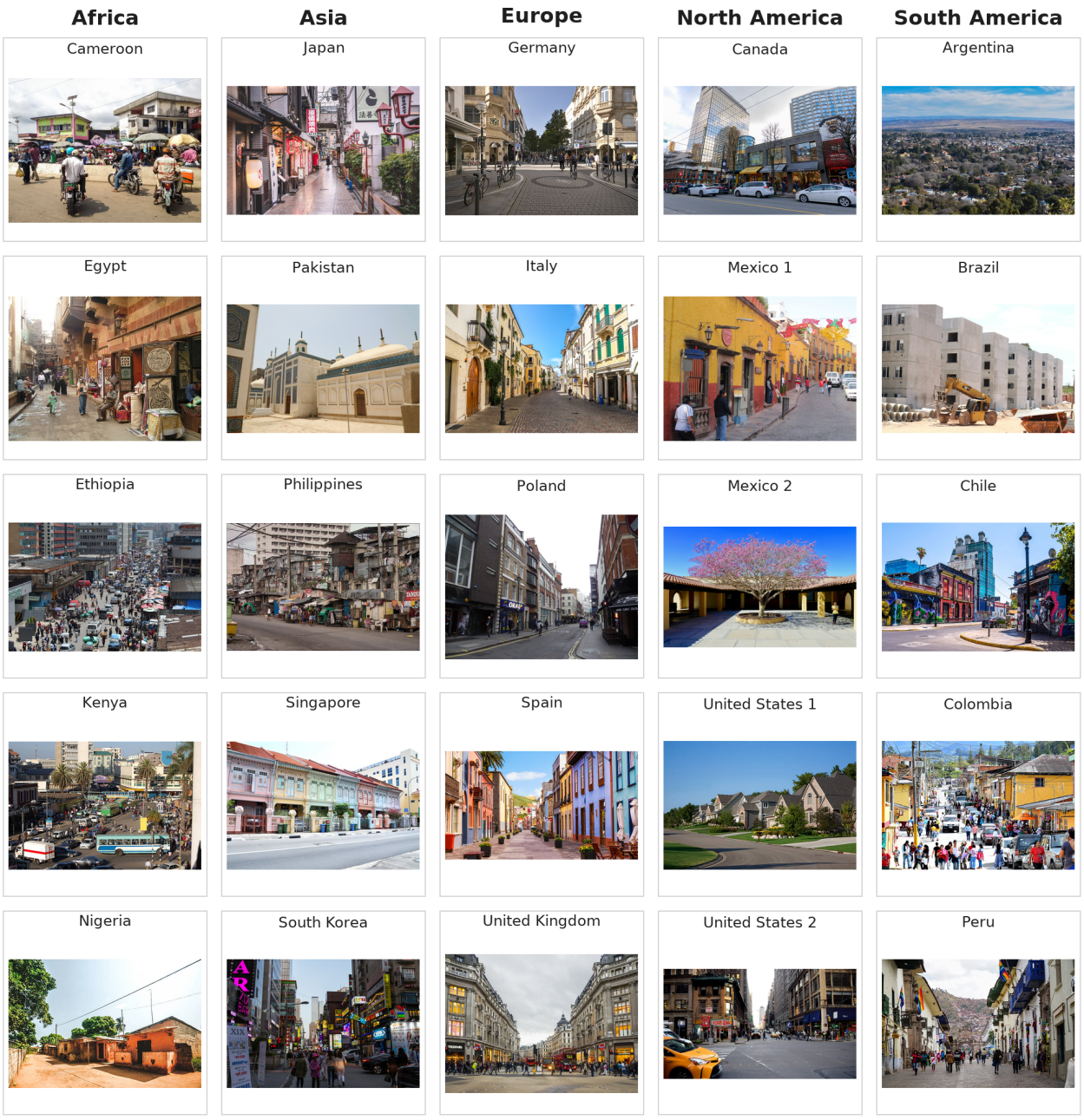


Figure S5. **Background (street) image samples.**

Figure S6. **Image generation examples.** Each row of composite images shows each food from different countries placed in multiple diverse global backgrounds. Results illustrate the cultural combinations represented in our dataset.

Table S3. **Comparison of cultural benchmark datasets.** This table summarizes existing datasets in terms of cultural content, task type, image modality, geographic and linguistic coverage, and the presence of culture mixing. Our benchmark dataset is the first to explicitly include **cultural mixing**, covering diverse foods and backgrounds across 30 countries in English, and supporting both real and synthetic images.

	Cultural Element	Evaluation Type	Task Type	Image Type	Countries	Languages	Culture Mixing
Bhatia et al. (2024) [?]	Object	Retrieval	Visual grounding	Real	50 countries	English	No
Yin et al. (2021) [?]	Scene	VQA	Commonsense reasoning	Real	4 regions	English	No
Vayani et al. (2025) [?]	Food, Scene, Object	VQA, Captioning	Cultural knowledge	Real	73 countries	100 languages	No
Romero et al. (2024) [?]	Scene, Object	VQA	Cultural knowledge	Real	30 countries	31 languages	No
Nayak et al. (2024) [?]	Object, Scene	VQA	Cultural knowledge	Real	11 countries	English	No
Nikandrou et al. (2025) [?]	Object	VQA	Contextual adaptation	Real / Hybrid	5 Countries	5 languages	No
Zhou et al. (2025) [?]	Food (text only)	Probing	Cultural knowledge	Mostly Text	14 countries	6 Languages	No
Kim et al. (2025) [?]	Ethnicity, Background	VQA	Cultural bias	Hybrid	5 countries	English	Yes
CultureMix	Food, Background	VQA	Cultural knowledge	Both	30 countries	English	**Yes**

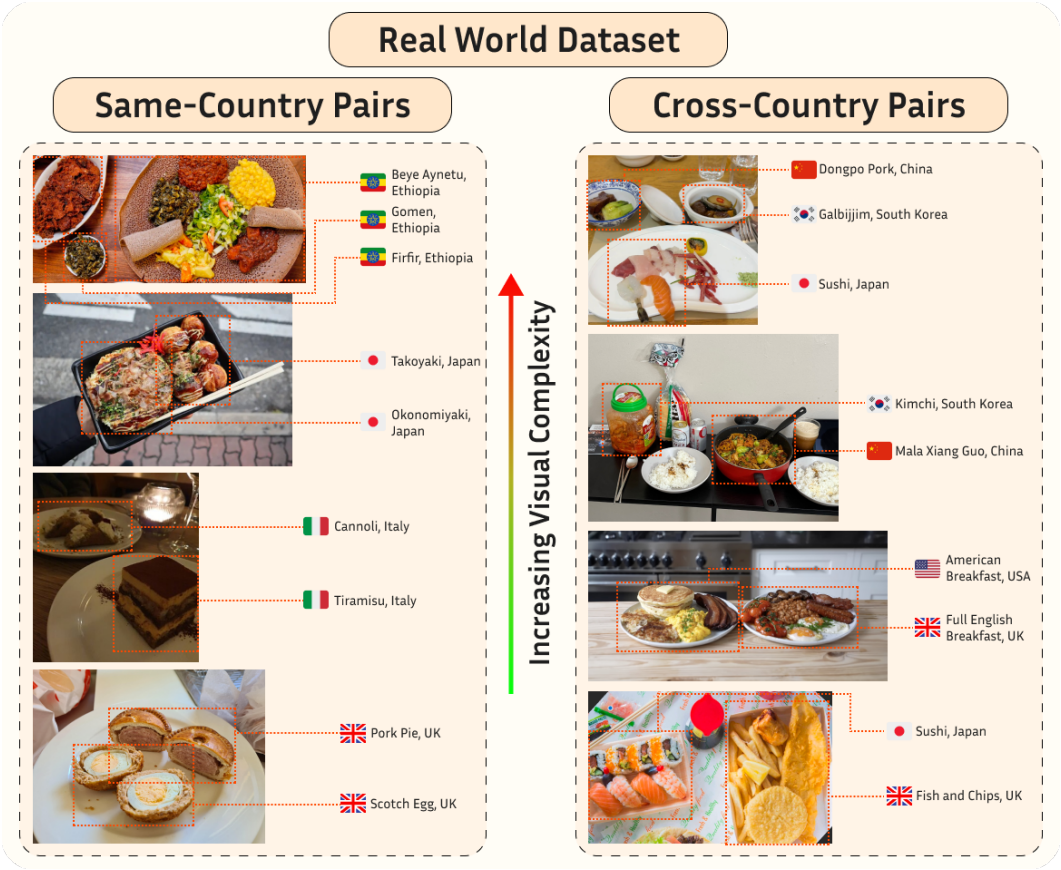


Figure S7. **Real-world dataset examples.** Same-country food pairs (left) and cross-country pairs (right) are shown across a spectrum of visual complexity, reflecting the diversity of appearance, plating, and scene structure encountered in real images.

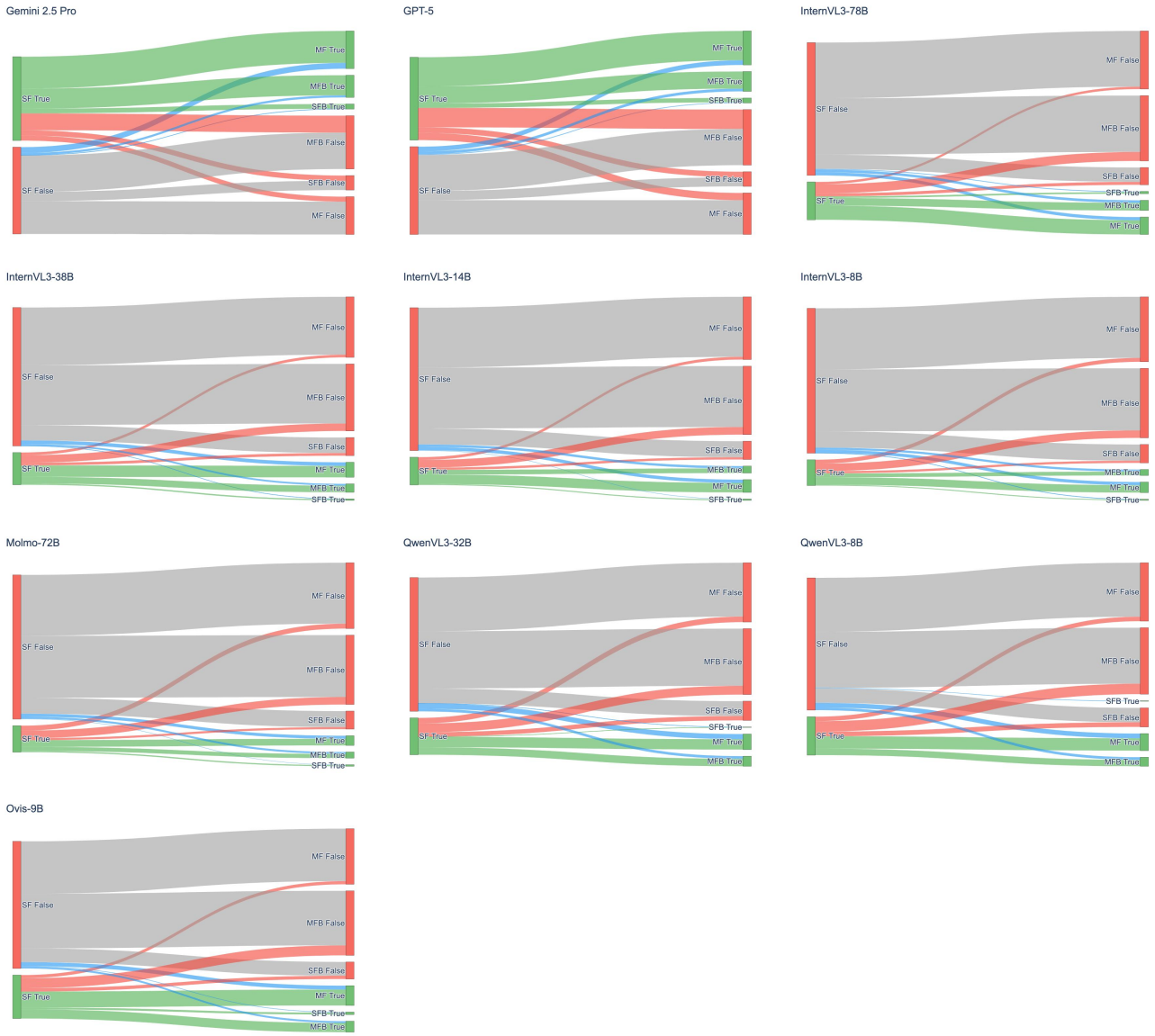


Figure S8. **Sankey diagram of country prediction for each LVLM.** We visualize how prediction correctness shifts from SF to the mixed subtasks (SFB, MF, MFB). Green indicates True → True, Red indicates True → False, Blue indicates False → True, and Gray indicates False → False. Although a small portion of predictions fall into Blue (False → True), a much larger portion appears in Red (True → False), showing that culturally mixed contexts confuse the models and often cause them to fail on cases they initially predicted correctly. While closed-source models perform better on country identification in SF, they also exhibit substantial True → False shifts in the mixed subtasks, resulting in reduced accuracy in culturally mixed settings.

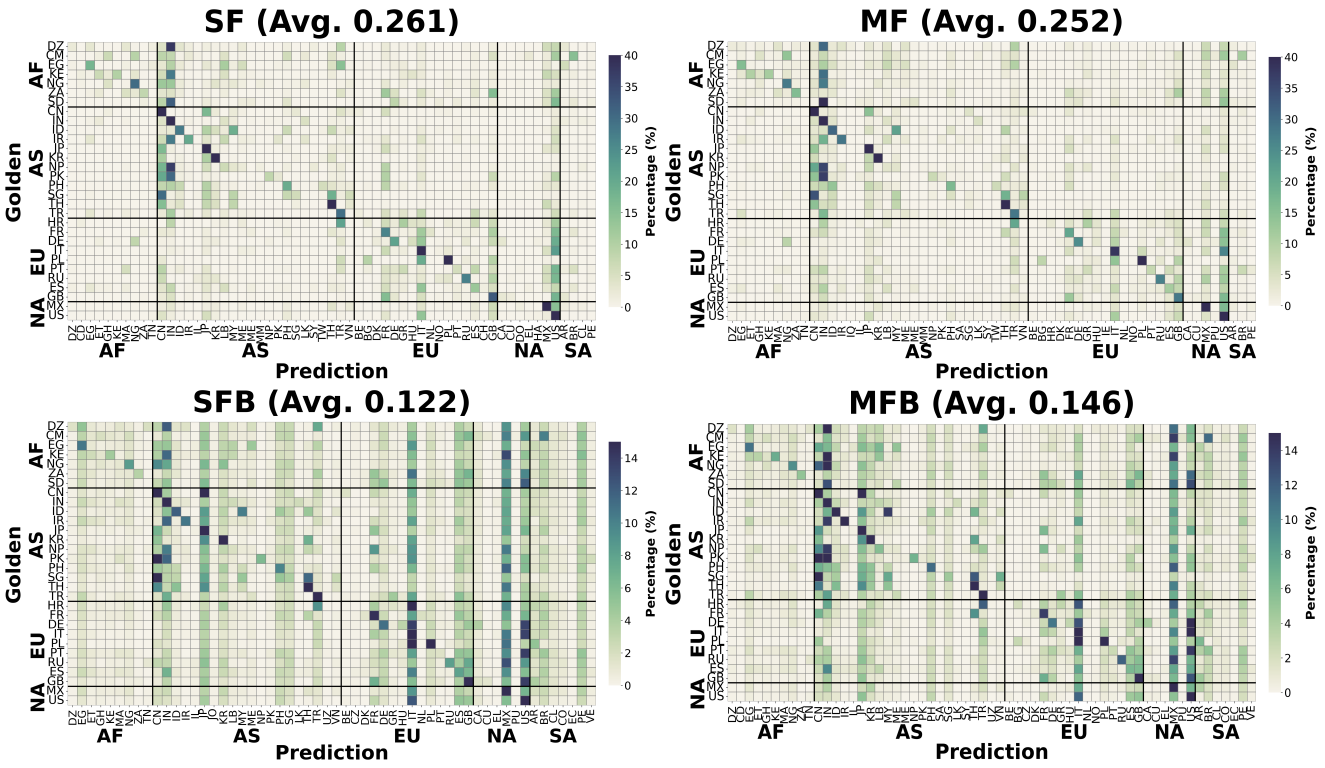
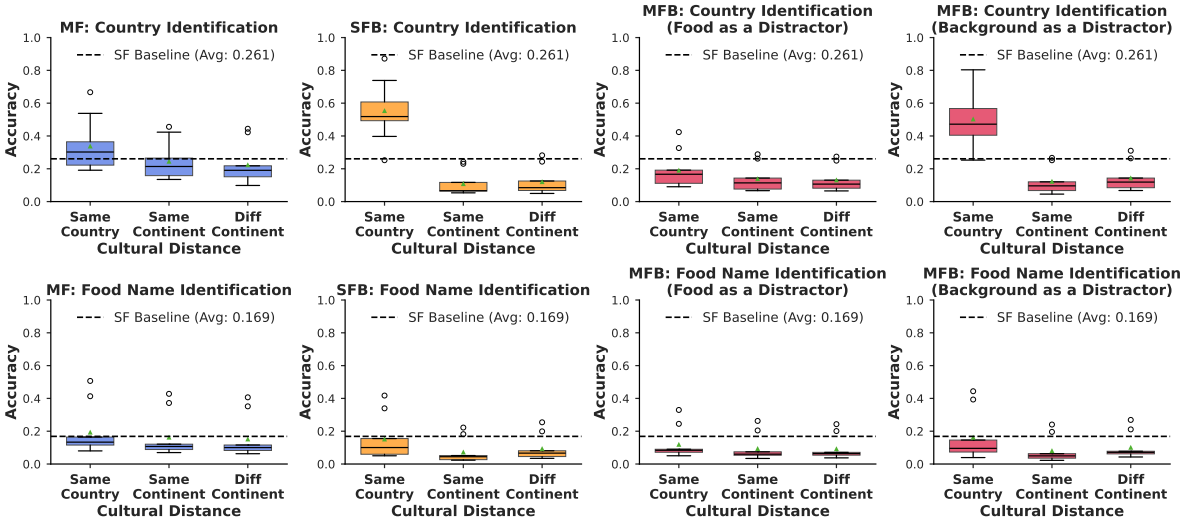
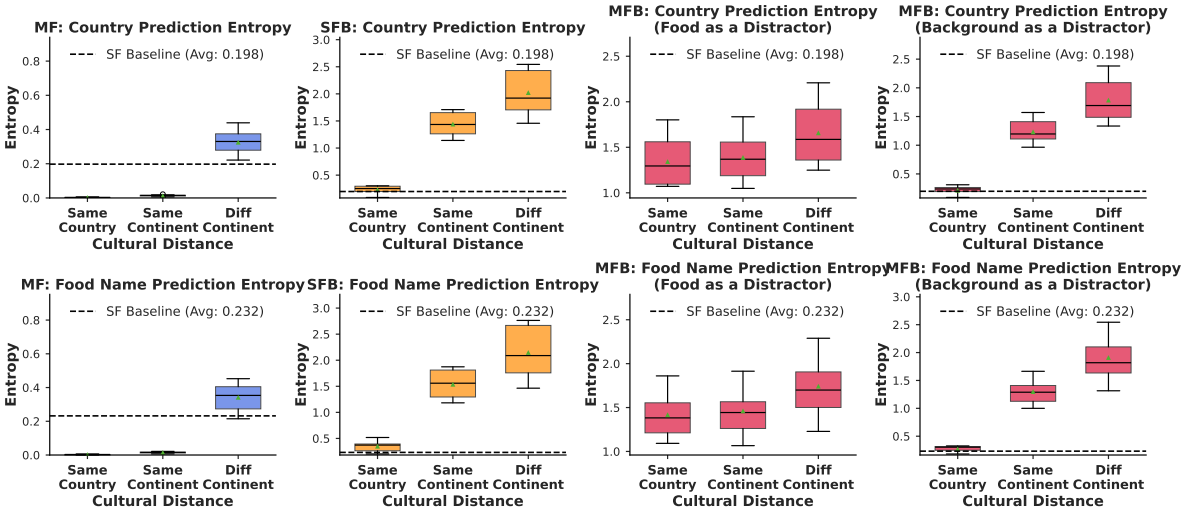


Figure S9. **Country identification target-prediction heatmaps for each subtask.** For every golden country, the plots show the distribution of predicted countries, illustrating both correct predictions and systematic confusions across models. The figure is repeated here in a larger size to facilitate closer examination.



(a) Cultural distance Vs. Accuracy



(b) Cultural distance Vs. Entropy

Figure S10. **Effect of cultural distance between target and distractor.** Models perform best when the target and distractor originate from the same country. Accuracy declines and entropy increases as cultural distance increases, indicating room for improving culture-mixing understanding in LVLMs.

Table S4. **Identification accuracy performance.** Most LVLMs show relatively higher accuracy in single food settings (SFB and SF) compared to that of multiple food settings (MFB and MF) for both country and food name identification. **Bold** and underline indicate the highest and lowest accuracy among subtasks.

(a) Country Identification

Model	**SF**	**MF**	**SFB**	**MFB**
Data Size	988	948	12,350	9,485
Gemini-2.5-Pro	0.457	**0.499**	0.286	0.313
GPT-5	**0.487**	0.450	0.250	0.271
InternVL3-78B	**0.234**	0.231	0.110	0.125
InternVL3-38B	**0.205**	0.199	0.088	0.112
InternVL3-14B	0.161	**0.170**	0.075	0.087
InternVL3-8B	**0.152**	0.140	0.065	0.071
Molmo-72B	**0.139**	0.129	0.080	0.081
QwenVL3-32B	**0.242**	0.213	0.077	0.124
QwenVL3-8B	**0.252**	0.227	0.060	0.124
Ovis-9B	**0.285**	0.263	0.133	0.148
Avg.	**0.261**	0.252	0.122	0.146

(b) Name Identification

Model	**SF**	**MF**	**SFB**	**MFB**
Data Size	988	948	12,350	9,485
Gemini-2.5-Pro	0.399	**0.435**	0.252	0.268
GPT-5	**0.379**	0.371	0.199	0.212
InternVL3-78B	**0.128**	0.113	0.065	0.069
InternVL3-38B	**0.100**	0.093	0.056	0.062
InternVL3-14B	**0.076**	0.071	0.043	0.045
InternVL3-8B	0.070	**0.072**	0.035	0.040
Molmo-72B	**0.090**	**0.093**	0.067	0.058
QwenVL3-32B	**0.164**	0.079	0.045	0.084
QwenVL3-8B	**0.145**	0.129	0.034	0.071
Ovis-9B	**0.152**	0.112	0.077	0.075
Avg.	**0.170**	0.157	0.087	0.098

Table S5. **Country and food name identification accuracy by background (`SFB` and `MFB`).** The models perform similarly for identifying the country and food with the landmark and street background.

(a) `SFB` Accuracy by Background Type

Model	**Country**		**Food Name**	
	Landmark	**Street**	**Landmark**	**Street**
Gemini-2.5-Pro	0.285	0.288	0.255	0.250
GPT-5	0.249	0.252	0.200	0.199
InternVL3-78B	0.119	0.101	0.068	0.063
InternVL3-38B	0.086	0.089	0.057	0.055
InternVL3-14B	0.078	0.073	0.044	0.041
InternVL3-8B	0.067	0.063	0.037	0.033
Molmo-72B	0.080	0.081	0.067	0.068
QwenVL3-32B	0.033	0.122	0.003	0.086
QwenVL3-8B	0.028	0.092	0.000	0.067
Ovis-9B	0.139	0.126	0.080	0.075
Avg.	0.116	0.129	0.081	0.094

(b) `MFB` Accuracy by Background Type

Model	**Country**		**Food Name**	
	Landmark	**Street**	**Landmark**	**Street**
Gemini-2.5-Pro	0.311	0.314	0.264	0.272
GPT-5	0.265	0.277	0.204	0.220
InternVL3-78B	0.134	0.117	0.071	0.068
InternVL3-38B	0.116	0.107	0.064	0.060
InternVL3-14B	0.095	0.080	0.045	0.044
InternVL3-8B	0.076	0.067	0.039	0.040
Molmo-72B	0.082	0.080	0.057	0.059
QwenVL3-32B	0.133	0.115	0.088	0.079
QwenVL3-8B	0.129	0.118	0.072	0.070
Ovis-9B	0.156	0.139	0.078	0.072
Avg.	0.150	0.141	0.098	0.098

Table S6. **Predicted label entropy by model and subtask.** The LVLMs show relatively high prediction uncertainty in single food settings (SFB and SF) compared to that of multiple food settings (MFB and MF) for both country and food name identification. **Bold** and underline indicate the highest and lowest accuracy among subtasks, respectively.

(a) Country Identification Entropy

Model	**SF**	**MF**	**SFB**	**MFB**	**Object**
Gemini-2.5-Pro	0.1674	0.4466	**1.9906**	1.7028	0.3127
GPT-5	0.1406	0.3774	**1.6294**	1.4498	0.4073
InternVL3-78B	0.1545	0.5471	**2.0483**	1.7942	0.4895
InternVL3-38B	0.1358	0.5294	**1.5570**	1.4269	0.5136
InternVL3-14B	0.2109	0.6882	**2.5861**	2.3096	0.7084
InternVL3-8B	0.2279	0.7483	**2.6767**	2.5273	0.7784
Molmo-72B	0.3794	0.8183	1.7660	**1.9765**	1.0114
QwenVL3-32B	0.2124	**1.0243**	2.7046	1.7793	0.6309
QwenVL3-8B	0.1894	0.6087	**2.5548**	1.5627	0.6015
Ovis-9B	0.1628	0.7224	**2.4157**	2.2161	0.4315
Avg.	0.1981	0.6511	**2.1929**	1.8745	0.5831

(b) Food Name Identification Entropy

Model	**SF**	**MF**	**SFB**	**MFB**	**Object**
Gemini-2.5-Pro	0.7720	0.9071	**2.2122**	2.0513	0.9076
GPT-5	0.8014	0.8144	**1.9083**	1.7651	0.8574
InternVL3-78B	0.7264	0.9045	**2.2624**	1.9843	0.8797
InternVL3-38B	0.7515	0.8252	**1.5961**	1.4558	0.8136
InternVL3-14B	0.7840	0.9458	**2.8754**	2.5452	0.9444
InternVL3-8B	0.8079	0.9485	**2.9280**	2.7138	0.9707
Molmo-72B	0.9270	1.0470	1.7244	**2.1658**	1.1527
QwenVL3-32B	0.7933	**1.1436**	3.1793	2.0625	0.9798
QwenVL3-8B	0.7914	0.9590	**2.7944**	1.6854	0.9032
Ovis-9B	0.7685	1.0390	**2.3858**	2.2773	0.9155
Avg.	0.7923	0.9534	**2.3866**	2.0707	0.9325

Table S7. Effect of direct prompting on InternVL3-38B.

OpenGVLab	**Entropy (↓)**			**Accuracy (↑, %)**			
InternVL3-38B	MF	SFB	MFB	SF	MF	SFB	MFB
Base	0.80	1.56	1.13	0.080	0.074	0.041	0.052
$Prompt_{Direct}$	1.03	**1.52**	**0.93**	**0.12**	0.040	**0.042**	0.017

Table S8. **Real world dataset identification accuracy.** LVLMs generally show better performance in identifying multiple foods from the same culture, even in real-world settings.

(a) Country Identification Accuracy

Model	Single	Multi (Same)	Multi (Diff)
Gemini-2.5-Pro	0.868	**0.922**	0.856
GPT-5	0.877	**0.904**	0.750
InternVL3-78B	0.721	**0.809**	0.654
InternVL3-38B	0.667	**0.791**	0.683
InternVL3-14B	0.562	**0.609**	0.558
InternVL3-8B	0.566	**0.678**	0.663
Molmo-72B	0.539	**0.670**	0.500
QwenVL3-32B	**0.658**	0.600	0.356
QwenVL3-8B	0.731	**0.748**	0.375
Ovis-9B	0.749	**0.896**	0.692
Avg.	0.694	**0.763**	0.609

(b) Name Identification Accuracy

Model	Single	Multi (Same)	Multi (Diff)
Gemini-2.5-Pro	0.699	**0.765**	0.692
GPT-5	**0.635**	0.617	0.519
InternVL3-78B	0.511	**0.583**	0.471
InternVL3-38B	0.443	**0.513**	0.471
InternVL3-14B	0.365	**0.400**	0.375
InternVL3-8B	0.384	0.391	**0.413**
Molmo-72B	0.311	**0.365**	0.298
QwenVL3-32B	**0.489**	0.478	0.327
QwenVL3-8B	**0.530**	0.461	0.337
Ovis-9B	0.511	**0.548**	0.500
Avg.	0.488	**0.512**	0.440

Table S9. **The effect of food item Loc on identification accuracy.** We observe almost no shift when comparing predictions on the original multi-food images and their horizontally flipped counterparts, indicating minimal positional bias in both country and name identification.

(a) Country Identification Accuracy

Model *Data Size*	MF 350	Loc Shift 100	Diff (Loc − MF)
InternVL3-8B	0.111	0.090	−0.021
Ovis-9B	0.237	0.250	+0.013
QwenVL3-8B	0.203	0.180	−0.023

(b) Name Identification Accuracy

Model *Data Size*	MF 350	Loc Shift 100	Diff (Loc − MF)
InternVL3-8B	0.049	0.020	−0.029
Ovis-9B	0.086	0.100	+0.014
QwenVL3-8B	0.131	0.120	−0.011

Table S10. **The effect of size on identification accuracy.** We observe almost no change in model performance between the original single-food images and their resized variants, indicating that identification accuracy is largely invariant to object size.

(a) Country Identification

Model *Data Size*	SF 247	Size Shift 741	Diff
InternVL3-8B	0.158	0.139	−0.019
Ovis-9B	0.271	0.290	+0.019
QwenVL3-8B	0.239	0.224	−0.015

(b) Food Name Identification

Model *Data Size*	SF 247	Size Shift 741	Diff
InternVL3-8B	0.061	0.066	+0.005
Ovis-9B	0.150	0.152	+0.003
QwenVL3-8B	0.142	0.143	+0.001

Figure S11. **Similar food prediction examples.** Gemini predicted *Vanille cake* as *Gugelhupf*, and InternVL3-8B predicted *Carolina-style pulled pork* as *Pernil*. In both cases, the predicted dishes are visually similar to the ground-truth foods yet are distinct items.