J Samples of Images and Captions Collected

J.1 Samples from Image Crowdsourcing

Varying image and caption quality The crowdsourced images from contributors have varying levels of iamge and caption quality. Within the quality assurance process, we filtered out the samples with low quality image and caption to ensure our SEA-VL data is high quality. We demonstrate the distinction between good quality and low quality images and captions in Figure 4.

Varying image cultural relevance The crowdsourced images from contributors across Southeast Asia (SEA) exhibit varying levels of cultural relevance, as reflected by their average Likert scores (1 - 5). Figure 1 presents samples with high cultural relevance (4.75–5.0), showcasing iconic landmarks, traditional cuisine, and significant cultural elements such as the Petronas Twin Towers, the Merlion statue, and the Kecak Dance of Bali. These images strongly align with regional heritage and traditions, receiving high agreement among annotators regarding their cultural significance. In contrast, Figure 2 illustrates images with medium cultural relevance (2.75–3.25), including elements like local motorcycles, street food, and residential architecture. While these images still contain culturally meaningful aspects, their association with distinct cultural identity may be less pronounced or subject to interpretation. Finally, Figure 3 highlights images with low cultural relevance (1.0–1.25). These images often lack strong ties to local traditions, leading to lower agreement on their relevance to SEA's cultural landscape. The variation in Likert scores across these figures underscores the subjective nature of cultural perception and the importance of carefully curating datasets for cultural representation tasks.

J.2 Samples from Image Crawling

We provide the image samples from our image crawling experiments in Figure 5 and Figure 6 for the CC3M and COYO datasets, respectively.

J.3 Samples from Image Captioning

We provide the image samples from our image captioning experiments in Figure 7-14 for both SEA-VQA and WorldCuisines datasets.

J.4 Samples for Image Generation

We provide the image samples from our image generation experiments in Figure 15, Figure 16, and Figure 17 for cuisine, landmark, and tradition images, respectively.

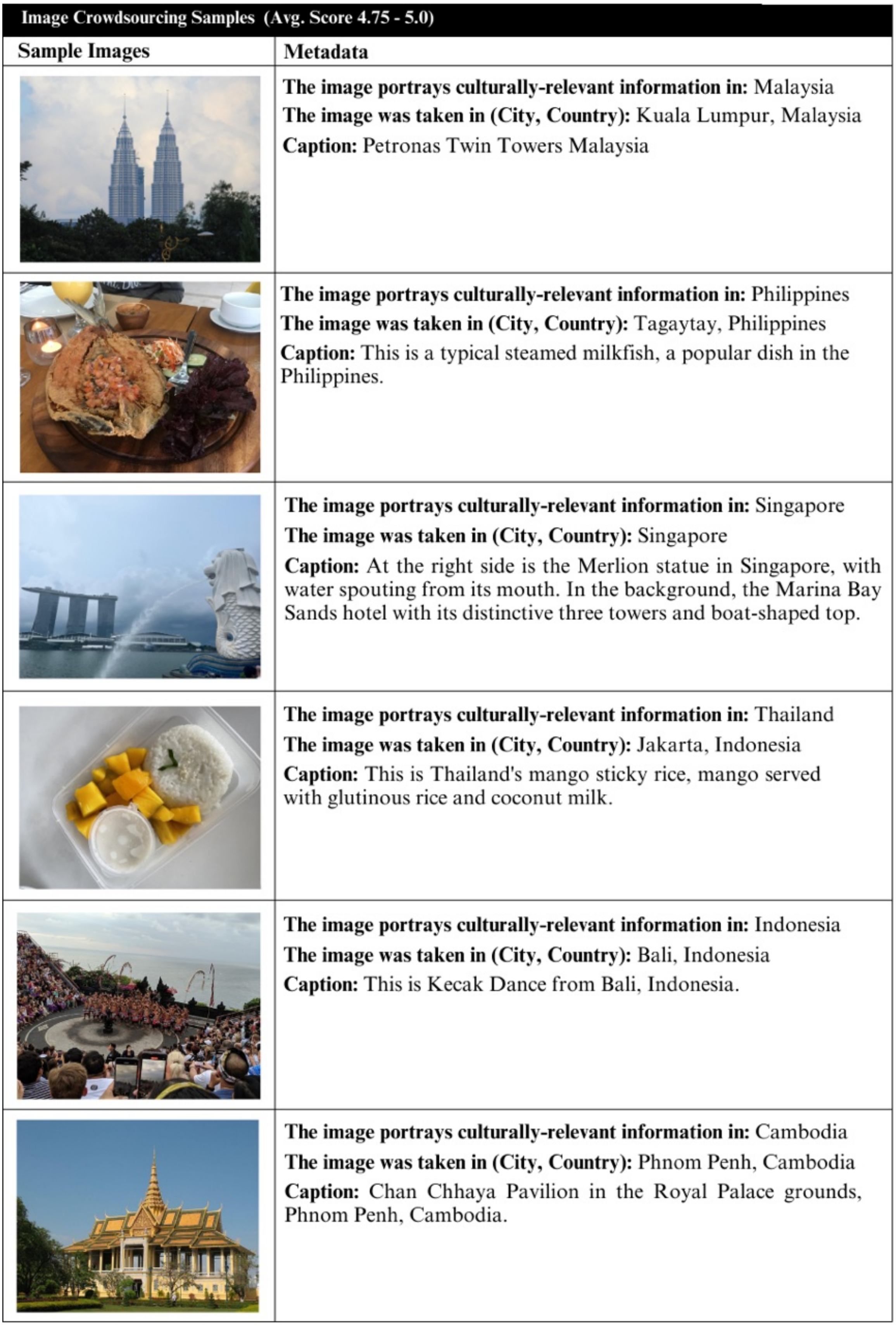


Image Crowdsourcing Samples (Avg. Score 4.75 - 5.0)

Sample Images	Metadata
	The image portrays culturally-relevant information in: Malaysia **The image was taken in (City, Country):** Kuala Lumpur, Malaysia **Caption:** Petronas Twin Towers Malaysia
	The image portrays culturally-relevant information in: Philippines **The image was taken in (City, Country):** Tagaytay, Philippines **Caption:** This is a typical steamed milkfish, a popular dish in the Philippines.
	The image portrays culturally-relevant information in: Singapore **The image was taken in (City, Country):** Singapore **Caption:** At the right side is the Merlion statue in Singapore, with water spouting from its mouth. In the background, the Marina Bay Sands hotel with its distinctive three towers and boat-shaped top.
	The image portrays culturally-relevant information in: Thailand **The image was taken in (City, Country):** Jakarta, Indonesia **Caption:** This is Thailand's mango sticky rice, mango served with glutinous rice and coconut milk.
	The image portrays culturally-relevant information in: Indonesia **The image was taken in (City, Country):** Bali, Indonesia **Caption:** This is Kecak Dance from Bali, Indonesia.
	The image portrays culturally-relevant information in: Cambodia **The image was taken in (City, Country):** Phnom Penh, Cambodia **Caption:** Chan Chhaya Pavilion in the Royal Palace grounds, Phnom Penh, Cambodia.

Figure 1: Samples of crowdsourced images from contributors across SEA with an average Likert score between 4.75 and 5.0 (High).

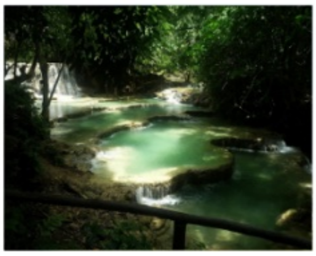
Image Crowdsourcing Samples (Avg. Score 2.75 - 3.25)	
Sample Images	**Metadata**
	The image portrays culturally-relevant information in: Indonesia, Malaysia, Singapore **The image was taken in (City, Country):** Singapore **Caption:** Coconut water, drunk directly from the coconut husk
	The image portrays culturally-relevant information in: Vietnam **The image was taken in (City, Country):** Hanoi, Vietnam **Caption:** This image highlights a classic Honda Super Cub motorcycle, a popular and timeless design on Vietnamese streets, parked against a simple urban backdrop.
	The image portrays culturally-relevant information in: Indonesia **The image was taken in (City, Country):** Belitung, Indonesia **Caption:** One of the houses of residents in Belitung. Built on the edge of the road, then access is a ground/sand road.
	The image portrays culturally-relevant information in: Myanmar **The image was taken in (City, Country):** Mandalay, Myanmar **Caption:** Flat rice noodle salad.
	The image portrays culturally-relevant information in: Cambodia, Thailand **The image was taken in (City, Country):** Siem Reap, Cambodia **Caption:** The Tuktuk is a popular mode of public transportation in Cambodia and nearby countries like Thailand.
	**The image portrays culturally-relevant information in:** Laos **The image was taken in (City, Country):** Luang Prabang, Laos **Caption:** The Kuang Si Waterfall cascades in layers, forming several emerald-green pools.

Figure 2: Samples of crowdsourced images from contributors across SEA with an average Likert score between 2.75 and 3.25 (Medium).

Image Crowdsourcing Samples (Avg. Score 1.0 - 1.25)

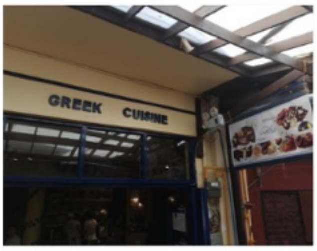
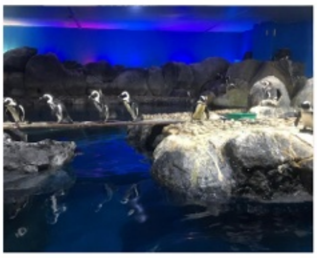
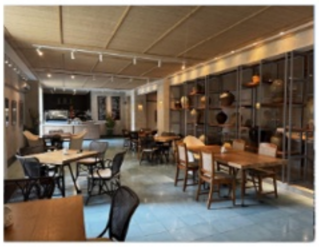
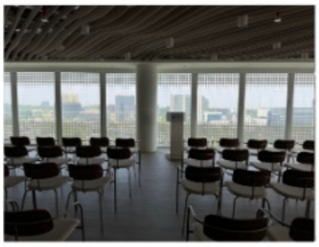
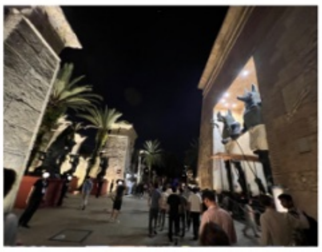
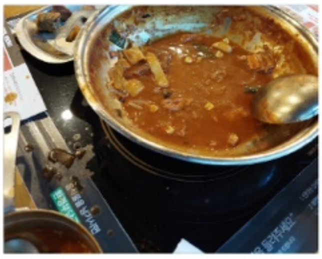
Sample Images	Metadata
	**The image portrays culturally-relevant information in:** Philippines **The image was taken in (City, Country):** Philippines **Caption:** GREEK CUSINE, a western restaurant on Boracay Island, is a restaurant with Greek cuisine, mainly featuring steaks, pork ribs, and other dishes.
	**The image portrays culturally-relevant information in:** Malaysia **The image was taken in (City, Country):** Malaysia **Caption:** The cute little penguins in the Underwater World are lively and active, and they will occasionally play with the tourists.
	**The image portrays culturally-relevant information in:** Indonesia **The image was taken in (City, Country):** Bali, Indonesia **Caption:** A spacious and inviting restaurant features a rustic design with wooden tables and chairs, complemented by a woven ceiling.
	**The image portrays culturally-relevant information in:** Singapore **The image was taken in (City, Country):** Singapore **Caption:** Conference room in one north, Singapore
	**The image portrays culturally-relevant information in:** Singapore **The image was taken in (City, Country):** Singapore **Caption:** A wide-angle shot of the Ancient Egypt area at Universal Studios in Singapore
	**The image portrays culturally-relevant information in:** Singapore **The image was taken in (City, Country):** Singapore **Caption:** Half finished pot of tteobokki (korean rice cakes)

Figure 3: Samples of crowdsourced images from contributors across SEA with an average Likert score between 1.0 and 1.25 (Low).

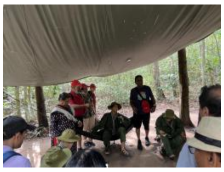
Good-quality Image

The image portrays culturally-relevant information in: Vietnam
The image was taken in (City, Country):" Jakarta, Indonesia
Caption in Native Language: Makanan populer vietnam banh mi dan ga nuong
English Caption: Local vietnam food: bahn mi and ga nuong

Good-quality Caption

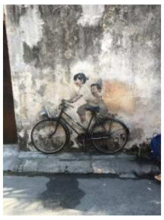
The image portrays culturally-relevant information in: Thailand
The image was taken in (City, Country):" Ayutthaya, Thailand
Caption in Native Language: เศียรพระพุทธรูปถูกรากของต้นไม้ห่อปกคลุม
English Caption: The stone Buddha head is entwined in tree roots.

Bad-quality Image



The image portrays culturally-relevant information in: Vietnam
The image was taken in (City, Country):" Phú Hiệp, Củ Chi, Ho Chi Minh City, Vietnam
Caption in Native Language: The tunnels of Củ Chi are an immense network of connecting tunnels located in the Củ Chi District of Ho Chi Minh City, Vietnam
English Caption: The tunnels of Củ Chi are an immense network of connecting tunnels located in the Củ Chi District of Ho Chi Minh City, Vietnam

Bad-quality Caption



The image portrays culturally-relevant information in: Malaysia
The image was taken in (City, Country):" Malaysia
Caption in Native Language:姐弟共骑是槟城标志性壁画，拍照时模拟骑车姿势，超有趣。也是网红打卡拍照地.
English Caption:姐弟共骑是槟城标志性壁画，拍照时模拟骑车姿势，超有趣。也是网红打卡拍照地.

Figure 4: Samples of crowdsourced images from contributors across SEA being annotated with “Good-quality Image”, “Good-quality Caption”, “Bad-quality Image”, and “Bad-quality Caption”. Validators consider the “bad-quality image“ example as poor due to its PII violation.

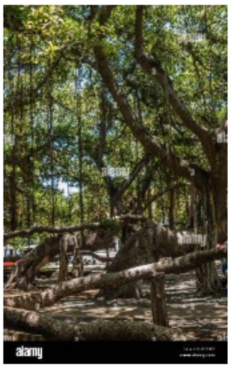
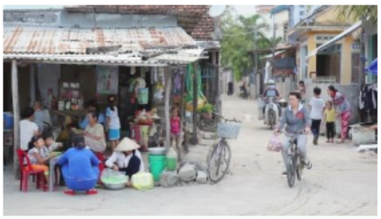
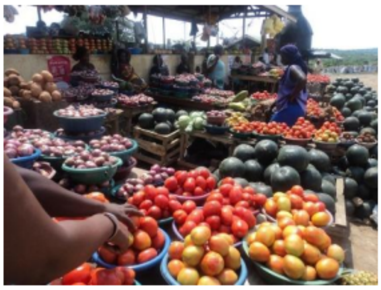
Image Crawling Samples (CC3M)	
SEA culturally-relevant	Not SEA culturally-relevant
	
	

Figure 5: Samples of crawled images from CC3M dataset.

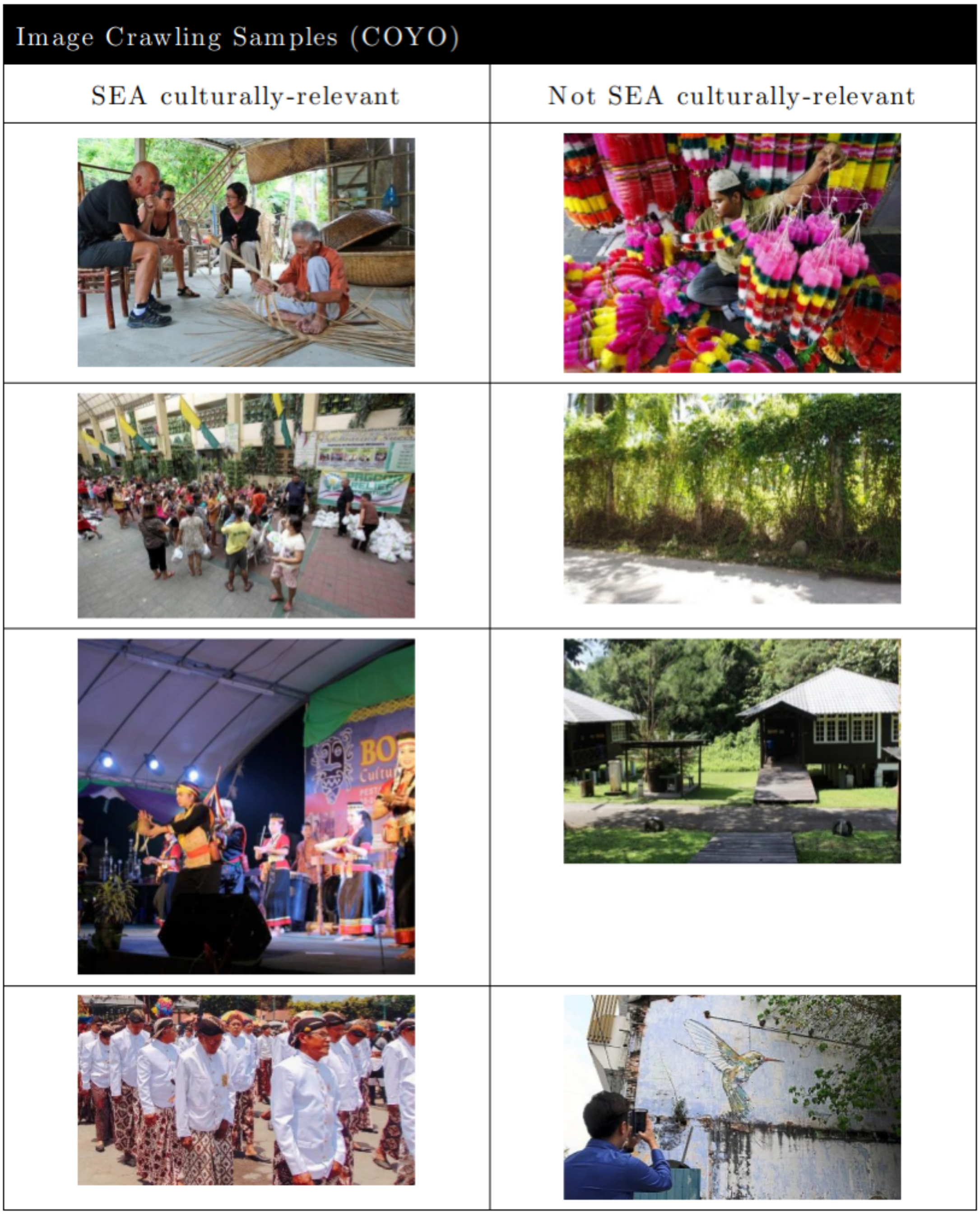
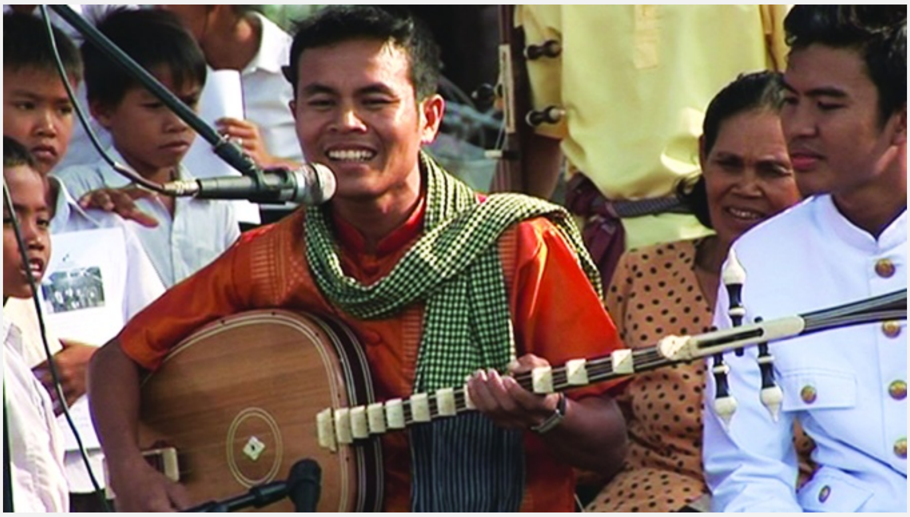


Figure 6: Samples of crawled images from COYO dataset.

Chapei Dang Veng



Human:
The demonstration of Chapei by a student of Cambodian Living Chapei organized by the Cambodian Living Arts at the Svay Antor Primary School, Kampot province.

MAYA (8B) :
A man in an orange shirt and scarf is playing a large guitar-like instrument, surrounded by other men in the scene. The cultural significance of this image lies in the representation of Southeast Asian music, as the man's traditional instrument showcases the region's rich musical heritage. The group of men, dressed in colorful clothing, are standing close together, emphasizing the sense of unity and collaboration often found in Southeast Asian cultures.

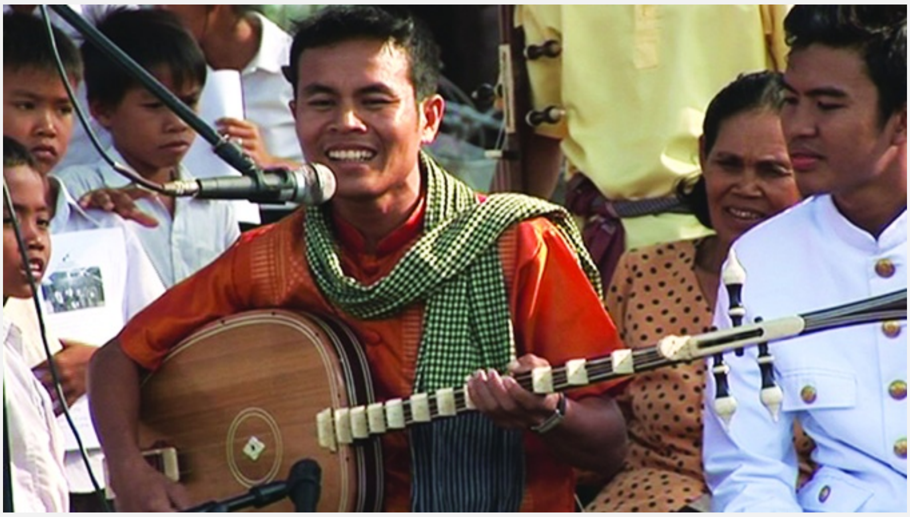
PALI Gemma 2 (10B) :
A medium shot of a man playing a traditional Cambodian guitar. The man is wearing an orange shirt with a green and black checkered scarf wrapped around his neck. He is facing forward and smiling with his eyes closed. His right hand is strumming the guitar, and his left hand is on the neck of the guitar. A microphone is in front of him, and a woman is sitting next to him on the right. She is wearing a brown dress with black polka dots. A man is sitting next to her on the right, and he is wearing a white shirt with gold buttons. A boy is sitting next to the man on the left, and his head is turned to the right. A man is sitting behind the woman on the left, and his head is turned to the right.

Pangea (7B) :
An artist in traditional attire performs with a Southeast Asian stringed instrument, embodying the rich cultural heritage of the region.

Qwen2-VL (7B) :
A vibrant scene from Southeast Asia captures the essence of cultural heritage. A musician strums a traditional instrument, the rebab, while a group of people, dressed in traditional attire, gather around. The setting is adorned with colorful fabrics and intricate designs, reflecting the rich tapestry of Southeast Asian culture. The atmosphere is one of celebration and community, where music and tradition come together to create a memorable experience.

Figure 7: Example of a **Location-Agnostic** English Caption Generated for SEA-VQA. Correctly identified traditions are highlighted in blue, while incorrectly identified traditions are marked in red.

Chapei Dang Veng



Human:
The demonstration of Chapei by a student of Cambodian Living Chapei organized by the Cambodian Living Arts at the Svay Antor Primary School, Kampot province.

MAYA (8B) :
The image features a young man in a traditional orange and green outfit playing a large stringed instrument, likely a Cambodian lute, surrounded by other men. This scene represents cultural elements from Cambodia, where music plays a significant role in their traditions. The presence of the lute and the men's attire suggest a celebration or performance of traditional Cambodian music, which is an essential part of their culture.

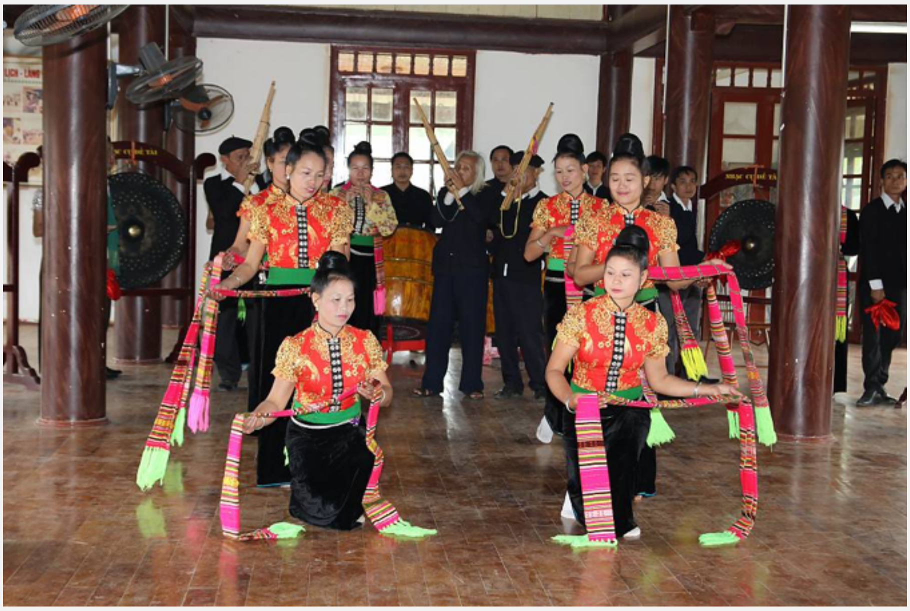
PALI Gemma 2 (10B) :
A medium shot of a man playing a traditional Cambodian guitar. The man is wearing an orange shirt with a green and black checkered scarf wrapped around his neck. He is facing forward, smiling, and looking at the camera. His right hand is strumming the guitar, and his left hand is on the neck of the guitar. A microphone is in front of him, and a woman is sitting next to him on the right. She is wearing a brown dress with black polka dots. A man is sitting next to her on the right, and he is wearing a white shirt with gold buttons. A boy is sitting next to the woman on the left, and his head is turned to the right. A man is sitting behind the woman on the left, and his head is turned to the right. A microphone is in front of him.

Pangea (7B) :
A Cambodian musician captivates the audience with his traditional guitar, adorned in vibrant local attire, embodying the rich cultural tapestry of Southeast Asia.

Qwen2-VL (7B) :
Musicians in traditional attire perform at a cultural festival in Cambodia, showcasing the rich heritage and music of Southeast Asia.

Figure 8: Example of a **Location-Aware** English Caption Generated for SEA-VQA. Correctly identified traditions are highlighted in blue, while incorrectly identified traditions are marked in red.

Art of Xòe dance of the Tai people in Viet Nam



Human:
The performing troupe performs Xòe at the Culture House, Nghĩa An Commune, Nghĩa Lộ Town, Yên Bái Province.

MAYA (8B) :
A group of young women dressed in traditional Southeast Asian clothing perform a dance routine in front of a crowd.

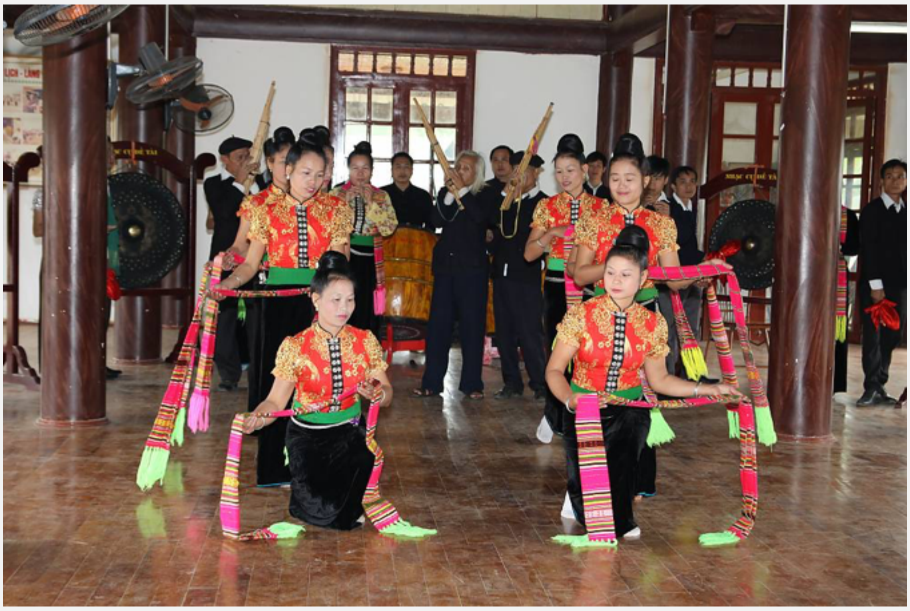
PALI Gemma 2 (10B) :
The image is of a group of women performing a traditional dance in a large room with a brown tile floor. The women are wearing red and black traditional dresses with long, colorful belts that they are holding in their hands. They are kneeling on the floor with their right knee on the ground and their left knee bent. There are men in black uniforms in the background playing drums and holding long wooden sticks. There are two large fans on the left and right side of the room, and there is a window in the back wall with a brown wooden frame.

Pangea (7B) :
A vibrant display of traditional Southeast Asian dance, showcasing the region's rich cultural heritage. The dancers' attire and the use of a large drum are indicative of the area's deep-rooted customs and artistic expressions.

Qwen2-VL (7B) :
Traditional dancers in vibrant costumes perform a cultural dance, accompanied by musicians playing Southeast Asian instruments like the gongs and bamboo flute, in a festive celebration.

Figure 9: Example of a **Location-Agnostic** English Caption Generated for SEA-VQA. Correctly identified traditions are highlighted in blue, while incorrectly identified traditions are marked in red.

Art of Xòe dance of the Tai people in Viet Nam



Human:
The performing troupe performs Xòe at the Culture House, Nghĩa An Commune, Nghĩa Lộ Town, Yên Bái Province.

MAYA (8B) :
A group of young women dressed in traditional Vietnamese costumes are performing a dance routine in front of a crowd.

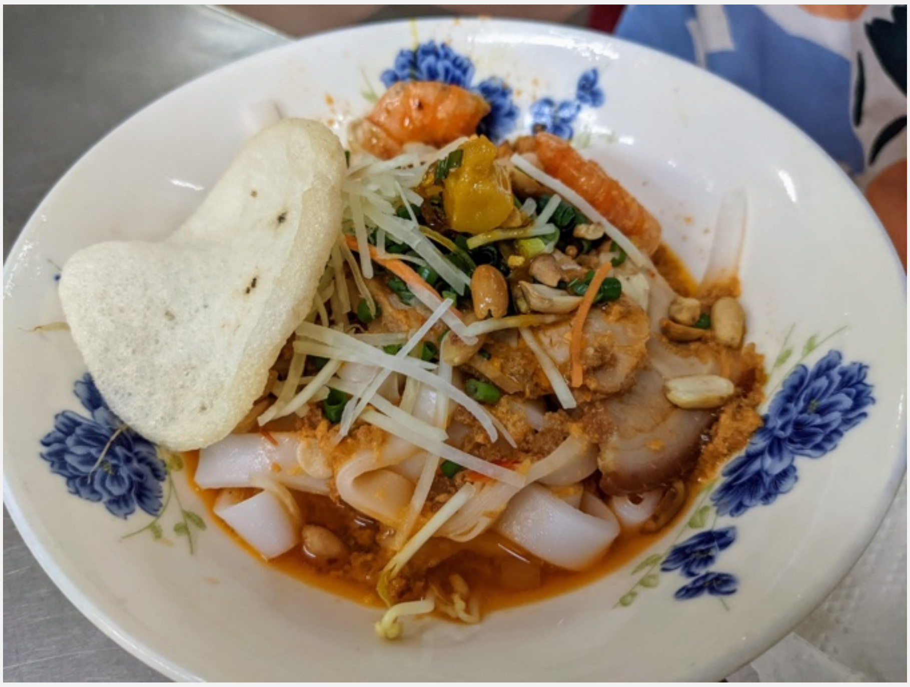
PALI Gemma 2 (10B) :
A group of women are performing a traditional dance in a large room with a wooden floor. The women are wearing traditional Vietnamese clothing, with red tops and black pants. They are holding long, colorful scarves in their hands, which they are waving around. The men in the background are wearing black uniforms and are holding long wooden sticks. There are two large wooden pillars on the left and right side of the room, and a window is in the middle of the wall. The ceiling is white, and there are two fans on the left and right side of the room.

Pangea (7B) :
A vibrant display of Vietnamese cultural heritage, where traditional attire and dance come together in a celebration of the country's rich history. The dancers' movements are fluid and precise, embodying the grace and rhythm of Vietnamese folk dance. The setting, with its wooden architecture and traditional instruments, adds to the authenticity of the scene, offering a glimpse into the enduring traditions of Southeast Asia.

Qwen2-VL (7B) :
Performers in traditional attire engage in a vibrant dance, showcasing the rich cultural heritage of Vietnam, with intricate costumes and rhythmic movements that reflect the country's deep-rooted traditions.

Figure 10: Example of a **Location-Aware** English Caption Generated for SEA-VQA. Correctly identified traditions are highlighted in blue, while incorrectly identified traditions are marked in red.

Mì Quảng



Human:
A picture of Mì Quảng. A dish made out of rice noodles, meat, and herbs, most commonly served with a small amount of broth, which is generally infused with turmeric. Peanut oil can also be added to make the dish more flavorful.

MAYA (8B) :
A white plate filled with a variety of Southeast Asian foods, including tofu, meat, and vegetables, sits on a table.

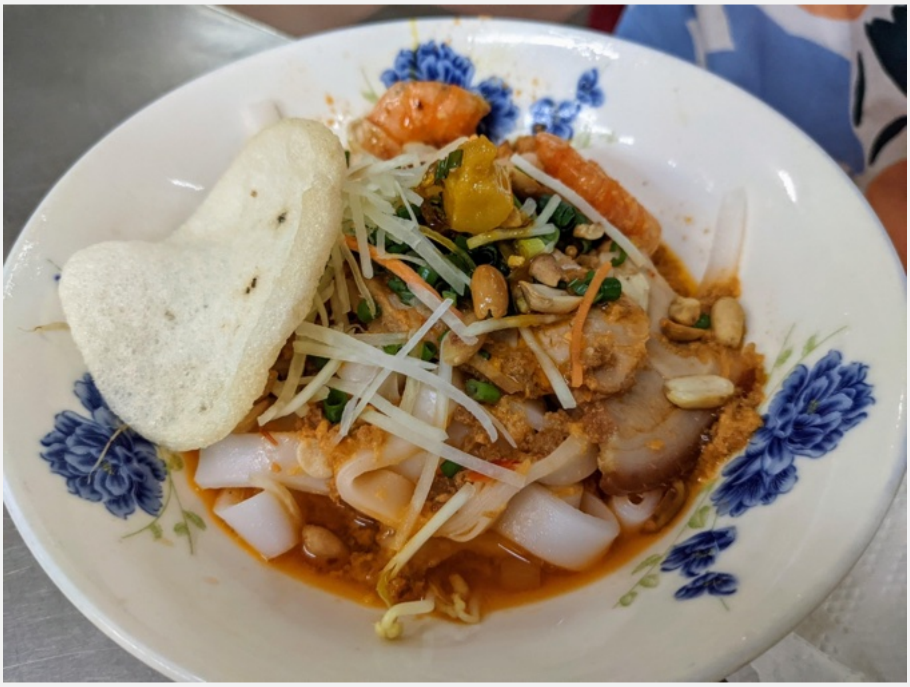
PALI Gemma 2 (10B) :
A close-up view of a bowl of food is seen from an angle above. The bowl is white with blue flowers painted on it. The food is a thick noodle with a red broth and a shrimp on the top right. The shrimp is cooked in a red sauce. The noodles are thick rice noodles. They are white and have a slight sheen to them. The broth is a deep red color and has a slight spice to it. The shrimp is cooked in a red sauce. The shrimp is cooked in a red sauce. The shrimp is cooked in a red sauce.

Pangea (7B) :
A vibrant bowl of traditional Southeast Asian noodle soup, garnished with fresh herbs and a crispy fried shallot, served with a side of crunchy fried tofu. The rich broth and colorful ingredients reflect the region's diverse culinary heritage.

Qwen2-VL (7B) :
A bowl of Vietnamese pho, a traditional noodle soup, is garnished with fresh vegetables, herbs, and a crispy pastry. The dish is served in a beautifully decorated ceramic bowl, reflecting the rich cultural heritage of Southeast Asia.

Figure 11: Example of a **Location-Agnostic** English Caption Generated for WorldCuisines. Correctly identified cuisines are highlighted in blue, while incorrectly identified cuisines are marked in red.

Mì Quảng



Human:
A picture of Mì Quảng. A dish made out of rice noodles, meat, and herbs, most commonly served with a small amount of broth, which is generally infused with turmeric. Peanut oil can also be added to make the dish more flavorful.

MAYA (8B) :
A white plate filled with a variety of Southeast Asian foods, including noodles, meat, and vegetables, sits on a table.

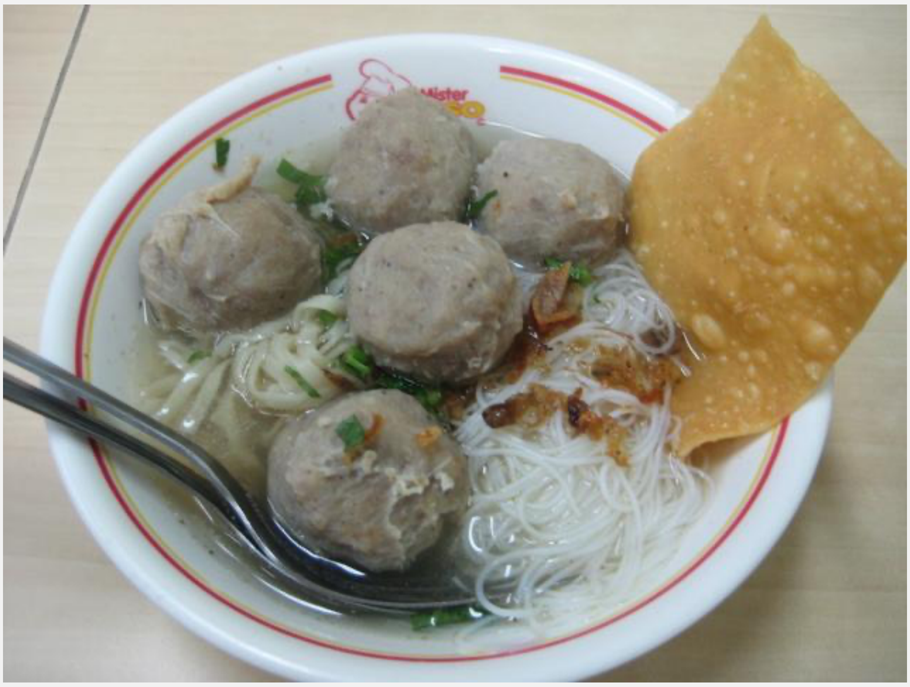
PALI Gemma 2 (10B) :
A close-up view of a white bowl with a blue floral design on the side of it. It has a brown liquid in it with rice noodles and vegetables on top of it. A shrimp is on the right. A piece of a white rice cracker is on the left. The bowl is sitting on a metal surface. A blue and white shirt is visible in the top right corner.

Pangea (7B) :
A vibrant bowl of Vietnamese Pho, a traditional noodle soup, garnished with fresh herbs and peanuts, served with a side of crispy fried shallots. This dish is a testament to the rich flavors and textures that define Vietnamese cuisine.

Qwen2-VL (7B) :
A bowl of Vietnamese pho, a traditional noodle soupis garnished with fresh herbspeanutsand a crispy pastry. The dish is served in a beautifully decorated ceramic bowlreflecting the rich culinary heritage of Vietnam.

Figure 12: Example of a **Location-Aware** English Caption Generated for WorldCuisines. Correctly identified cuisines are highlighted in blue, while incorrectly identified cuisines are marked in red.

Bakso



Human:
A bowl of Bakso. This dish consists of finely ground beef with a small quantity of tapioca flour and salt, and then served in a bowl of beef broth, with yellow noodles, rice vermicelli,broccoli, bean sprout, sprinkled with fried shallots and celery.

MAYA (8B) :
A bowl of noodle soup with meatballs and a tortilla-like cracker on top, served on a white plate.

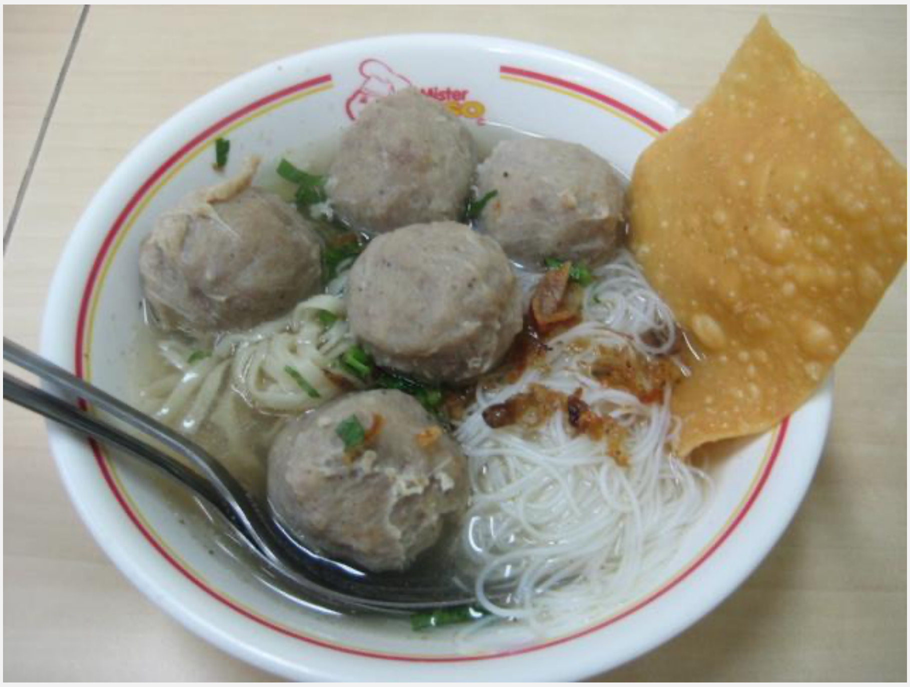
PALI Gemma 2 (10B) :
The image is of a bowl of Bakso, a traditional Indonesian beef ball dish. The bowl is white with a red and yellow line around the edge. The bowl has a red and yellow logo of a chef in the top left corner. The bowl has a large piece of fried tofu on the right side. The tofu is a light brown color and is placed on top of the rice noodles. The rice noodles are placed on the right side of the bowl. The beef balls are placed on top of the rice noodles. The beef balls are a light gray color. There is a metal spoon in the bowl. The spoon is placed on top of the rice noodles. The bowl is placed on a white tile floor.

Pangea (7B) :
A traditional bowl of Indonesian beef noodle soup, known as 'Soto Betawi,' featuring tender meatballs and fresh herbs, served with a side of crispy fried shallots and a crunchy cracker, all in a rich broth that captures the essence of Southeast Asian culinary delights.

Qwen2-VL (7B) :
A bowl of traditional Southeast Asian cuisine, featuring meatballs, noodles, and a crispy pastry, sits on a table, reflecting the rich culinary heritage of the region.

Figure 13: Example of a **Location-Agnostic** English Caption Generated for WorldCuisines. Correctly identified cuisines are highlighted in blue, while incorrectly identified cuisines are marked in red.

Bakso



Human:
A bowl of Bakso. This dish consists of finely ground beef with a small quantity of tapioca flour and salt, and then served in a bowl of beef broth, with yellow noodles, rice vermicelli,broccoli, bean sprout, sprinkled with fried shallots and celery.

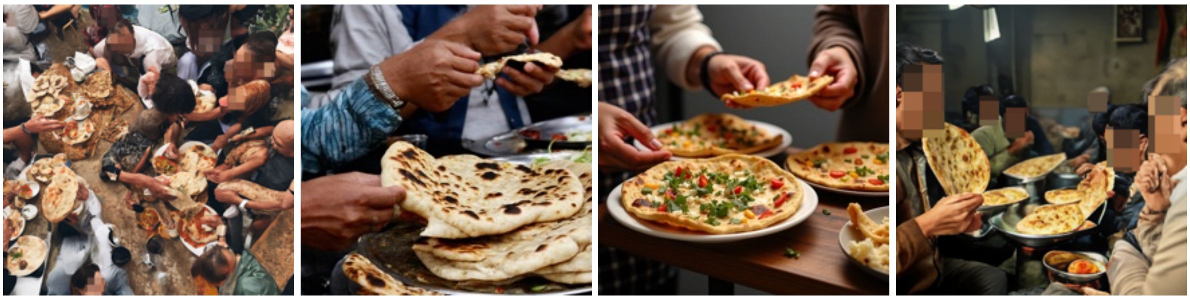
MAYA (8B) :
A bowl of soup with meatballs, noodles, and a tortilla-like piece on a white plate.

PALI Gemma 2 (10B) :
A high-angle view of a bowl of bakso, a traditional Indonesian dish. The bowl is white with a red and yellow line around the edge. There are five meatballs in the broth, with rice noodles on the right side of the bowl. There is a large piece of fried dough on the top right side of the bowl. There is a metal spoon in the broth on the left side of the bowl. There is a red and yellow logo with a person's head on the top left side of the bowl.

Pangea (7B) :
A traditional Indonesian noodle soup, known as 'Mie Ayam' or 'Ayam Taliwang,' served with meatballs and a side of crispy fried shallots. This dish is a delightful representation of Southeast Asian culinary traditions, featuring fresh herbs and a rich broth that captures the essence of Indonesian flavors.

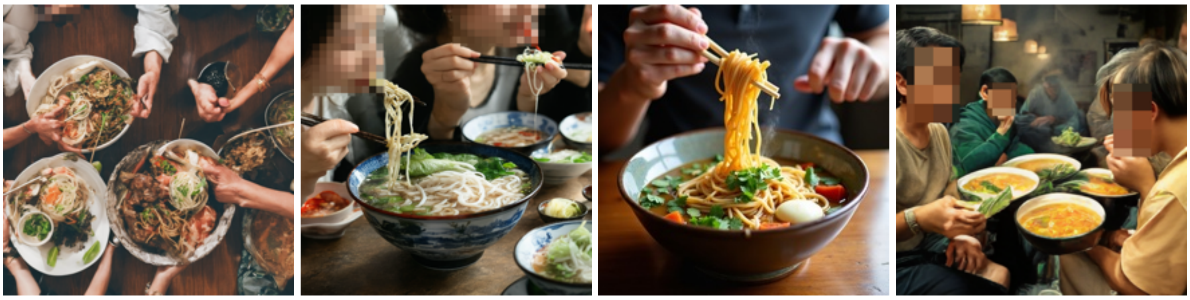
Qwen2-VL (7B) :
A bowl of Indonesian meatball soupknown as "bakso," is served with thin rice noodles and a crispy fried pastry. The dish is a popular street food in Indonesiaoften enjoyed with a side of sambal for added flavor.

Figure 14: Example of a **Location-Aware** English Caption Generated for WorldCuisines. Correctly identified cuisines are highlighted in blue, while incorrectly identified cuisines are marked in red.



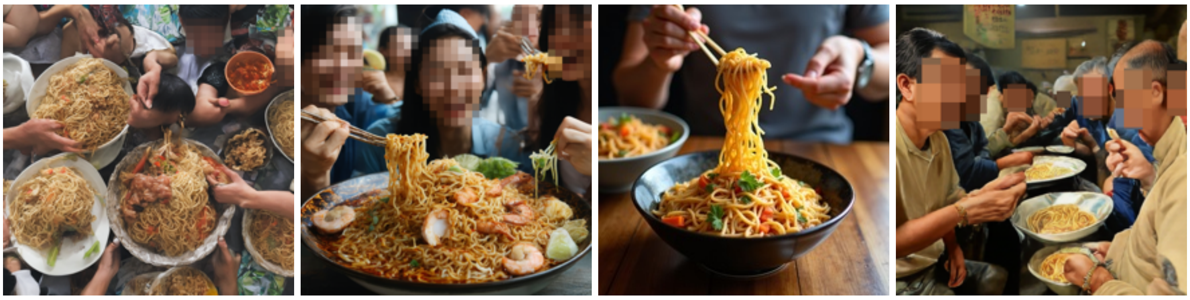
Stable Diffusion 2 | Stable Diffusion 3.5 | FLUX.1-dev | Janus Pro (7B)

An image of people eating Naan.



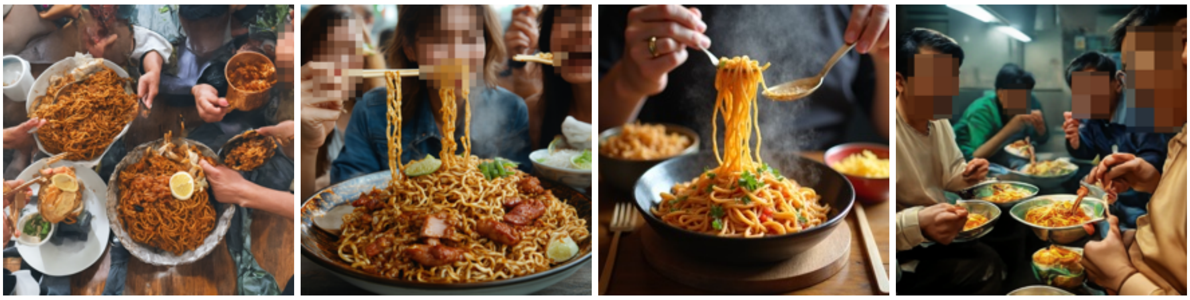
Stable Diffusion 2 | Stable Diffusion 3.5 | FLUX.1-dev | Janus Pro (7B)

An image of people eating Pho.



Stable Diffusion 2 | Stable Diffusion 3.5 | FLUX.1-dev | Janus Pro (7B)

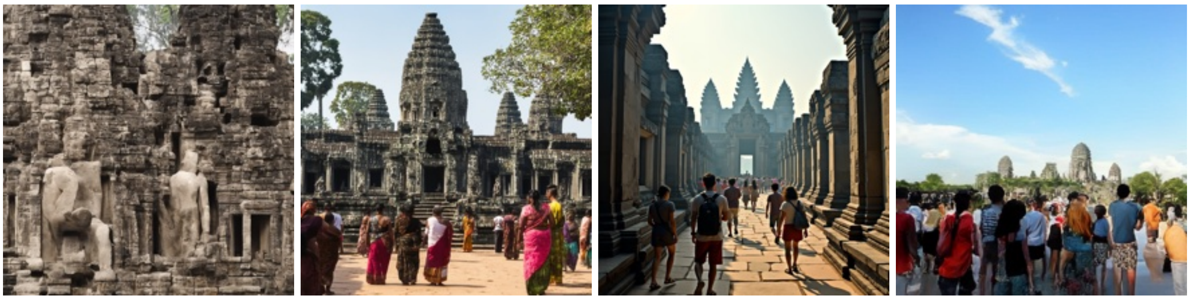
An image of people eating Pancit.



Stable Diffusion 2 | Stable Diffusion 3.5 | FLUX.1-dev | Janus Pro (7B)

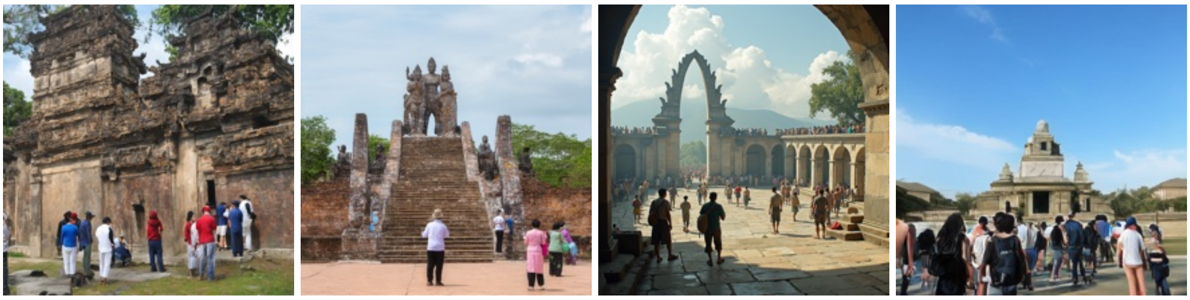
An image of people eating Mee goreng.

Figure 15: Qualitative comparison of generated images. Stable Diffusion 3.5 performs better for culinary images.



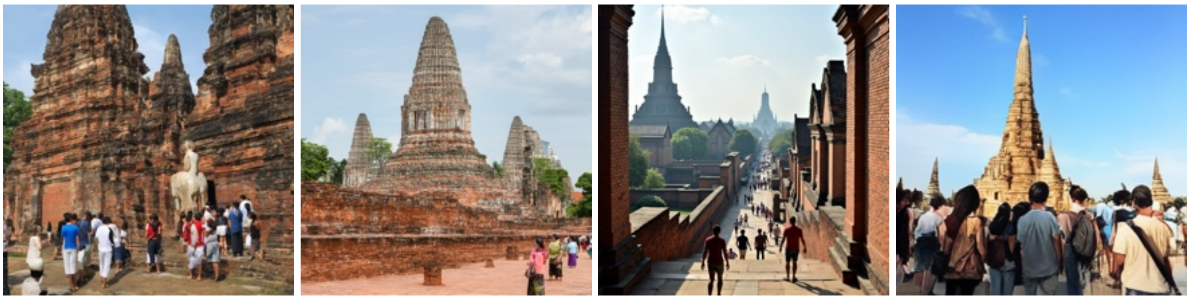
Stable Diffusion 2 | Stable Diffusion 3.5 | FLUX.1-dev | Janus Pro (7B)

An image of people at Angkor.



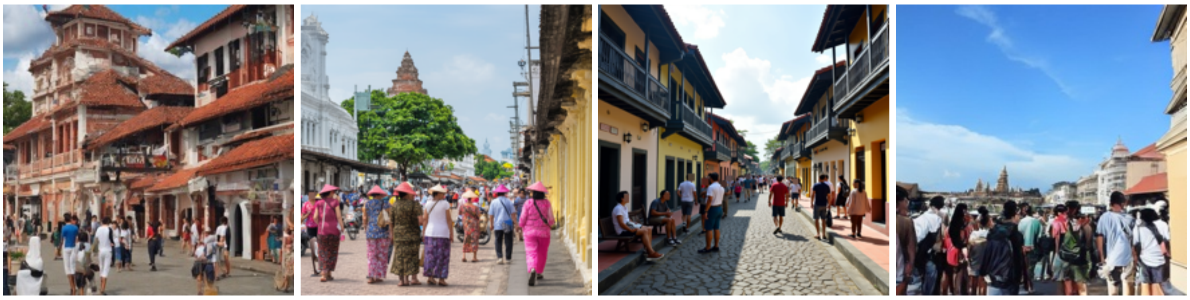
Stable Diffusion 2 | Stable Diffusion 3.5 | FLUX.1-dev | Janus Pro (7B)

An image of people at Complex of Hué Monuments.



Stable Diffusion 2 | Stable Diffusion 3.5 | FLUX.1-dev | Janus Pro (7B)

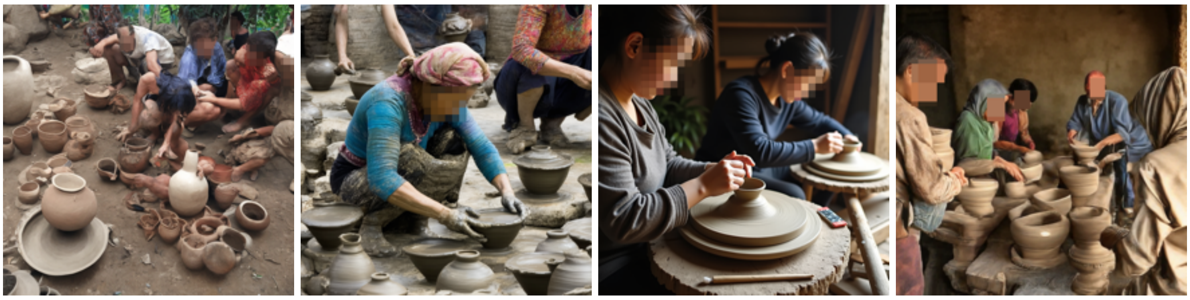
An image of people at Historic City of Ayutthaya.



Stable Diffusion 2 | Stable Diffusion 3.5 | FLUX.1-dev | Janus Pro (7B)

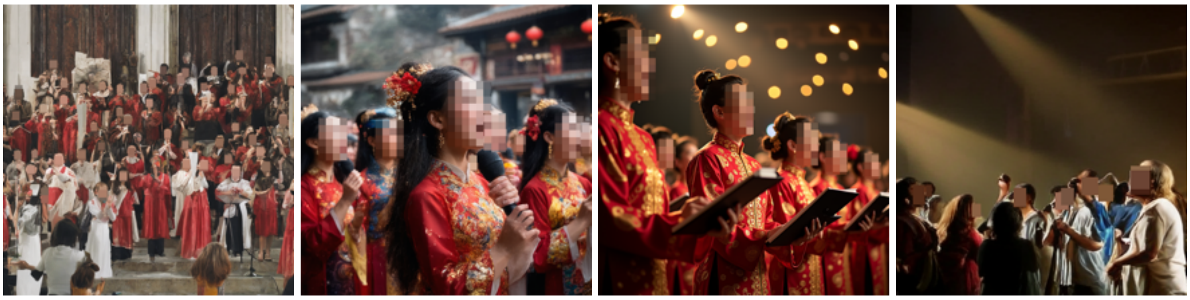
An image of people at Melaka and George Town, Historic Cities of the Straits of Malacca.

Figure 16: Qualitative comparison of generated images. None of the models perform well for landmark images.



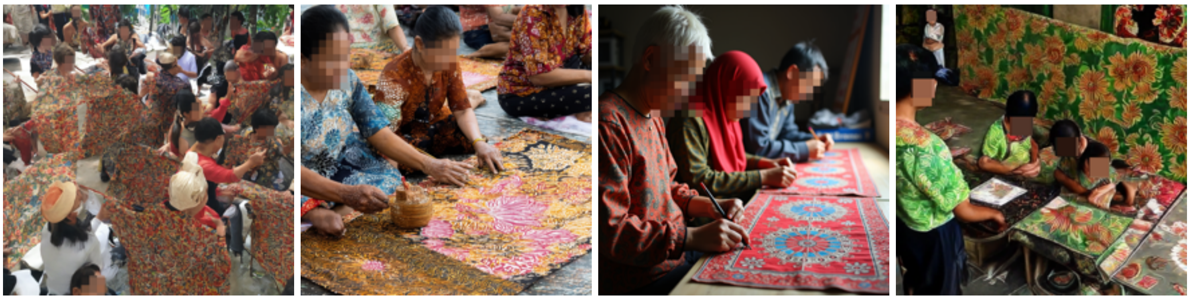
Stable Diffusion 2 Stable Diffusion 3.5 FLUX.1-dev Janus Pro (7B)

An image of people doing Art of pottery-making of Chăm people.



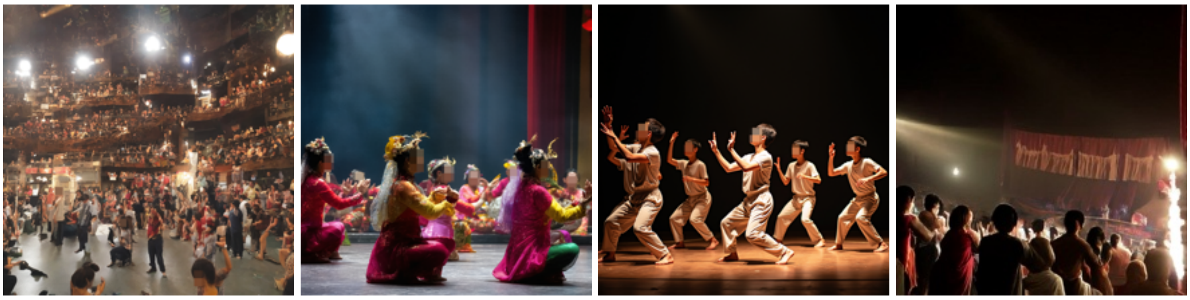
Stable Diffusion 2 Stable Diffusion 3.5 FLUX.1-dev Janus Pro (7B)

An image of people doing Ca trù singing.



Stable Diffusion 2 Stable Diffusion 3.5 FLUX.1-dev Janus Pro (7B)

An image of people doing Indonesian Batik.



Stable Diffusion 2 Stable Diffusion 3.5 FLUX.1-dev Janus Pro (7B)

An image of people doing Mak Yong theatre.

Figure 17: Qualitative comparison of generated images. Stable Diffusion 3.5 and FLUX.1-Dev perform acceptably for a portion of tradition images.