Question:
What elements in Image1 and Image2 suggest these places have a rich historical background?

Ground Truth:
Elements in Image1 and Image2 that suggest these places have a rich historical background can be found by examining their prominent architectural features and settings. In Image1, the stone construction, tall, fortified towers, and extensive defensive walls all suggest a medieval origin. Such structures were typically built to last, using materials that could withstand the ravages of time and conflict. The architectural design, aimed at defense, points to historical conflicts and the need for such fortifications. The mere fact that this structure remains intact today indicates its historical prominence and the significant events it likely witnessed. In Image2, the combination of old stone buildings, roof tiles, and a mixture of architectural styles similarly indicates a long and varied history. The juxtaposition of simple residential buildings with intricate church spires and towers suggests continuous habitation and use over many centuries. The presence of large, ornate churches or cathedrals indicates historical wealth and cultural significance, as such buildings were often funded by the prosperous portions of medieval society or influential religious orders.

Question:
Compare the cooking techniques shown in Image1 and Image2.

Ground Truth:
The cooking techniques displayed in Image1 and Image2 reveal distinct culinary methods reflecting different traditions and end-results. In Image1, the main technique is deep-frying. This process involves immersing the doughnuts in hot oil, which cooks them rapidly at high temperatures. Deep frying is a technique known for producing a crunchy texture and rich flavor due to the Maillard reaction, where proteins and sugars brown upon exposure to heat. It is a quick method but requires precision to ensure the product is neither undercooked nor burnt. In contrast, Image2 presents a classical Italian cooking method where spaghetti is being tossed in tomato sauce. This involves boiling the spaghetti until al dente, then gently incorporating it into the sauce in a pan. This technique emphasizes absorption and coating, where the pasta not only holds onto the sauce but absorbs some of the flavors, providing a harmonious blend. It is a more controlled cooking method compared to deep frying, focusing on subtlety and meld of flavors over high-temperature cooking.

Figure 1. **Left: Visual Perception Reasoning.** The most basic visual reasoning. **Right: Visual Relationship Reasoning.** Understand and articulate the connections and relationships between different images.

Question:
How might the emotional expressions or symbolic meanings differ between the structures shown in Image1 and Image2?

Ground Truth:
The emotional and symbolic expressions conveyed by the structures in Image1 and Image2 reflect their distinct purposes and cultural contexts. The ruins in Image1 likely evoke a sense of historical grandeur and lost glory. They symbolize the might and reach of the Roman Empire, embodying the practicality and efficiency of Roman engineering. There is a sense of nostalgia and reflection on the advancements in construction, urban planning, and societal governance that these ruins represent. In contrast, the Gothic cathedral in Image2 carries a more transcendent and spiritual symbolism. Its towering spires and intricate designs are meant to inspire awe and reverence. The cathedral's architecture directs one's gaze heavenward, symbolizing a connection between earth and the divine. The use of light, seen through large stained-glass windows, creates a mystical atmosphere, enhancing the spiritual experience. Symbolically, these buildings represent the medieval church's authority, the ubiquitous presence of Christianity, and the aspiration to reach towards heaven through human craftsmanship.

Question:
What kinds of wildlife might thrive in the environments depicted in Image1 and Image2?

Ground Truth:
The types of wildlife thriving in the environments shown in Image1 and Image2 would be markedly different due to their distinct ecosystems. In the lush, green environment of Image1, the biodiversity is likely to be dense and varied, hosting species adapted to forest and hillside habitats. Mammals such as deer, monkeys, and a variety of rodents could be common, along with numerous bird species ranging from small songbirds to larger raptors. The undergrowth and forest canopy would support insects, reptiles, and amphibians, contributing to a rich and interconnected web of life. Conversely, Image2 would support a different array of wildlife, more suited to coastal and marine environments. Marine life, including fish, crustaceans, and coral species, would dominate. Seabirds such as gulls, pelicans, and terns would be common, capitalizing on the abundant food resources provided by the sea. The coastal flora, consisting of salt-tolerant plants and shrubs, would provide habitat for various small mammals and insects that thrive in drier, windier conditions.

Figure 2. **Left: High-Level Semantic Reasoning.** Grasping the deeper meanings, symbols, or abstract concepts across multiple images. **Right: Cross-modal reasoning.** Extracting visual information from an image and combining it with textual cues to infer knowledge or information that goes beyond the content of the image.

Table 1. We compare MMDU with several existing datasets.

Datasets	Average Turns	Average Images	Max Turns	Max Images
MMvet	1	1	1	1
ConvBench	3	1	3	1
Qbench2	1	2	1	2
SEEDBench2	1	2.7	1	/
BLINK	1	/	1	4
Spot-the-Diff	4	2	4	2
Dreamsim	2	3	3	3
Coinstruct	7.5	2.7	/	4
MMDU (Ours)	**15**	**3.8**	**27**	**20**

Table 2. Results on MMDU with different maximum number of tokens.

Models	Max tokens	C	R	VP	LC	AA	IRU	Overall Score
LLaVa-1.5	2k	19.0	19.0	21.8	29.3	22.5	19.6	20.9
	4k	25.4	25.6	31.1	40.8	32.9	29.5	30.0
LLaVa-1.5+mmdu-45k	2k	20.0	20.1	22.1	29.4	23.5	21.6	22.3
	4k	31.5	32.3	34.9	45.0	36.3	33.8	34.9
	8k	34.2	34.3	36.1	48.2	39.6	34.7	**37.1**

Table 3. We test LLaVa-1.5 on several multi-image benchmarks.

Models	MMMU (multi-pics)	BLINK	Qbench2	Mantis (sequence)	Mantis (merge)	MMDU
LLaVa-1.5	27.7	37.1	46.0	37.8	41.9	32.2
LLaVa-1.5+mmdu-45k	**29.8**	**40.1**	**48.5**	**44.7**	**44.7**	**37.2**
Δ	+2.1	+3.0	+2.5	+6.9	+2.8	+5.0

Table 4. Results on MMDU with different SFT strategies.

	C	R	VP	LC	AA	IRU	Overall
Continue training	34.3	34.5	36.7	47.2	38.5	35.5	37.2
Add to the existing pool	34.3	36.3	37.1	47.3	38.9	35.7	37.3