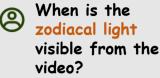
## (a) Long Video Benchmark (e.g., Video MME)







A: On March 19. B: On March 24.

C: On March 25. D: On March 29.

Correct ans: A



**NO Synergistic Interaction** 





A: Roller coaster

B: On a pirate ship
C: Take Ferris wheel
D: Take the jumping
machine"

Correct ans: A

"...Ambient sounds include cheerful male conversations, the gentle clinking of cutlery, and soft background music.

reflecting a casual and inviting atmosphere..."

## Visual Cue ONLY

## -----

## (c) Long Audio-Visual Benchmark (i.e., our proposed AVBench)



What can be inferred about the setting of the scene based on the background sound and visual elements?

A: The environment is a silent, solitary kitchen where someone is preparing food alone.

B: The setting is a formal dining room with a quiet, subdued atmosphere.

C: The scene is set in a busy street market with vendors calling out to passersby. D: The scene takes place in a lively, informal dining environment where people are enjoying a meal together.

Correct ans: D

Synergistic Audio-Visual Integration for Long Videos