

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



When is the **zodiacal light** visible from the video?

- A: On **March 19**.
- B: On March 24.
- C: On March 25.
- D: On March 29.

Correct ans: **A**

Visual Cue ONLY

## (b) Short Audio-Visual Benchmark (e.g., AVQA)



Why the **people** in the video scream?

- A: **Roller coaster**
- B: On a pirate ship
- C: Take Ferris wheel
- D: Take the jumping machine"

Correct ans: **A**

NO Synergistic Interaction

## (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