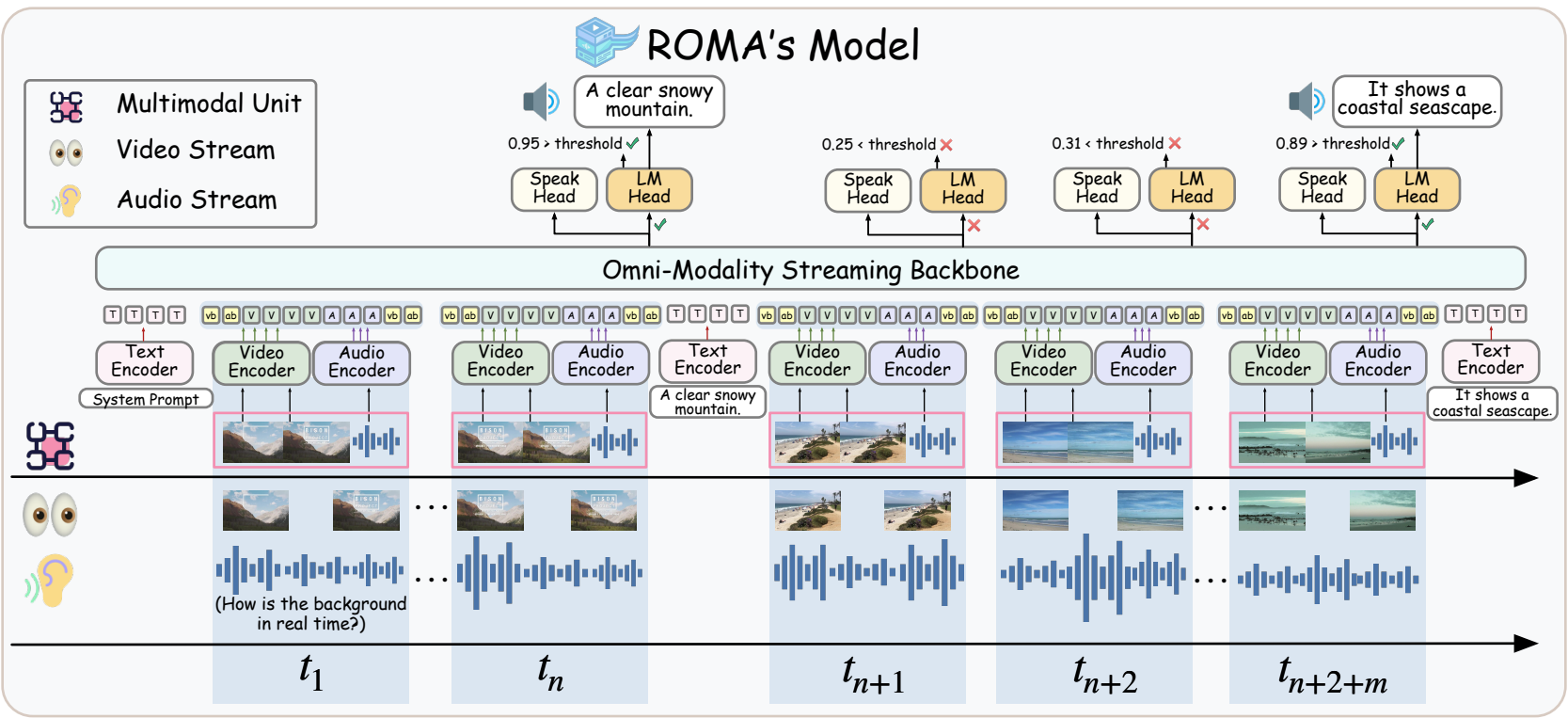
ROMA's Model
Multimodal Unit
Video Stream
Audio Stream
A clear snowy mountain.
0.95 > threshold
Speak Head
LM Head
0.25 < threshold
0.31 < threshold
It shows a coastal seascape.
0.89 > threshold
Omni-Modality Streaming Backbone
Text Encoder
System Prompt
Video Encoder
Audio Encoder
A clear snowy mountain.
It shows a coastal seascape.
(How is the background in real time?)
t_1
t_n
t_{n+1}
t_{n+2}
t_{n+2+m}