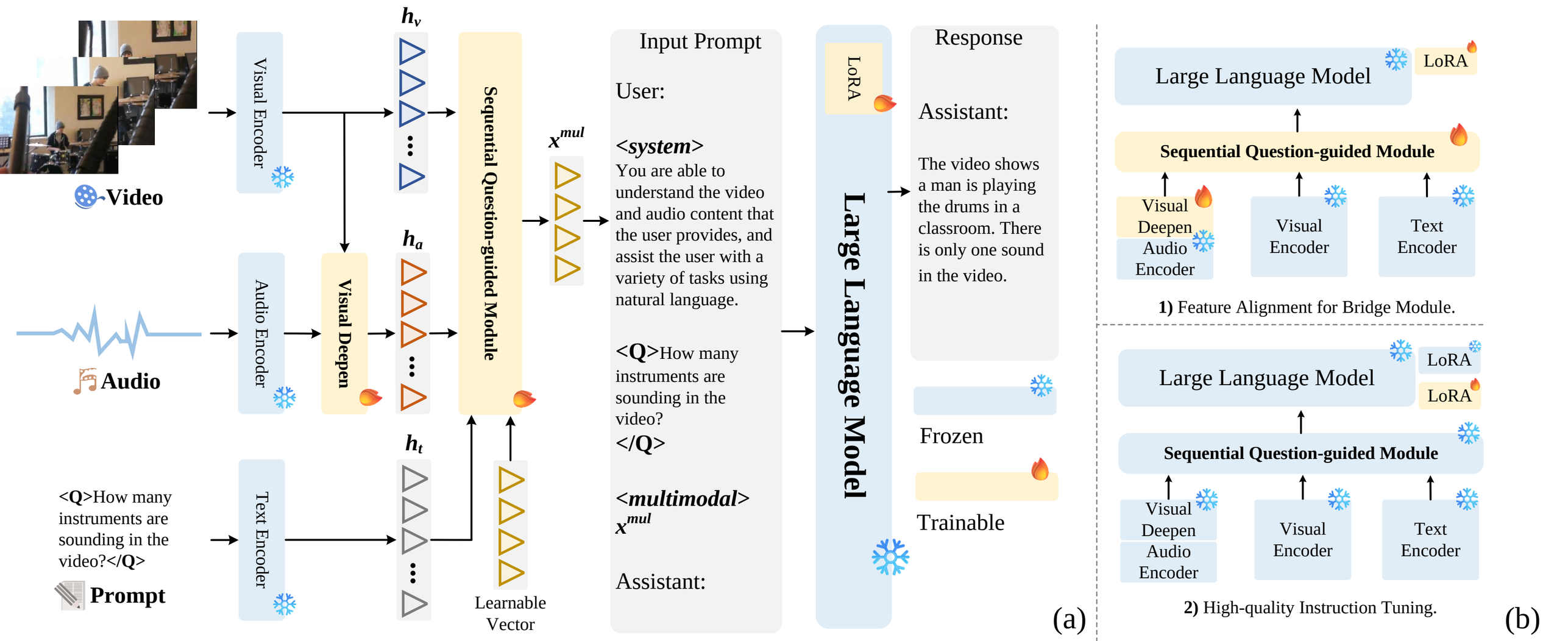
Video
Audio
<Q>How many instruments are sounding in the video?</Q>
Prompt
Visual Encoder
Audio Encoder
Visual Deepen
Text Encoder
h_v
h_a
h_t
Sequential Question-guided Module
x^{mul}
Learnable Vector
Input Prompt
User:
<system>
You are able to understand the video and audio content that the user provides, and assist the user with a variety of tasks using natural language.
<Q>How many instruments are sounding in the video?</Q>
<multimodal>
x^{mul}
Assistant:
LoRA
Large Language Model
Response
Assistant:
The video shows a man is playing the drums in a classroom. There is only one sound in the video.
Frozen
Trainable
(a)
Large Language Model
LoRA
Sequential Question-guided Module
Visual Deepen
Audio Encoder
Visual Encoder
Text Encoder
1) Feature Alignment for Bridge Module.
Large Language Model
LoRA
LoRA
Sequential Question-guided Module
Visual Deepen
Audio Encoder
Visual Encoder
Text Encoder
2) High-quality Instruction Tuning.
(b)