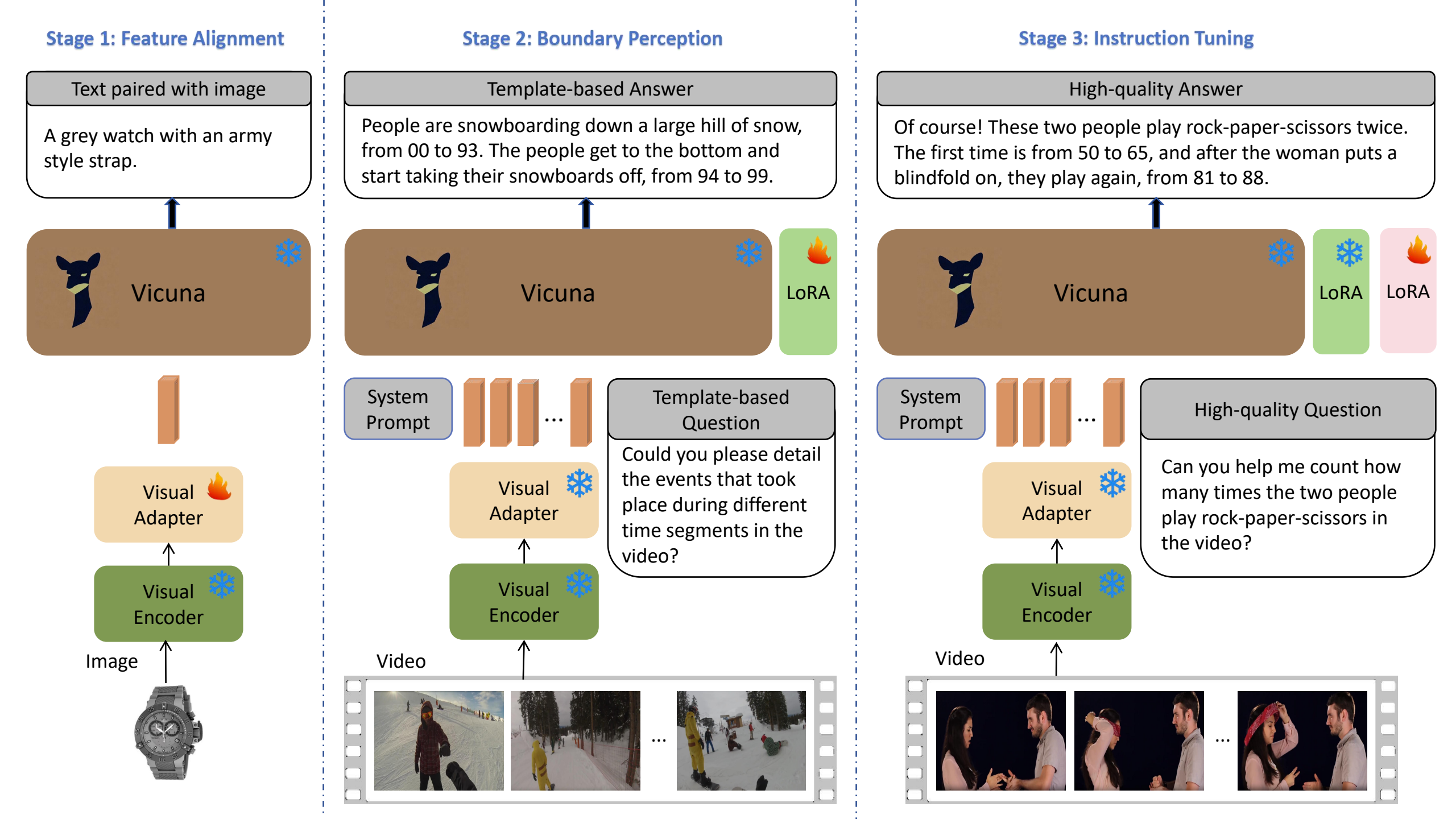
Stage 1: Feature Alignment
Text paired with image
A grey watch with an army style strap.
Vicuna
Visual Adapter
Visual Encoder
Image
Stage 2: Boundary Perception
Template-based Answer
People are snowboarding down a large hill of snow, from 00 to 93. The people get to the bottom and start taking their snowboards off, from 94 to 99.
Vicuna
LoRA
System Prompt
Template-based Question
Could you please detail the events that took place during different time segments in the video?
Visual Adapter
Visual Encoder
Video
Stage 3: Instruction Tuning
High-quality Answer
Of course! These two people play rock-paper-scissors twice. The first time is from 50 to 65, and after the woman puts a blindfold on, they play again, from 81 to 88.
Vicuna
LoRA
LoRA
System Prompt
High-quality Question
Can you help me count how many times the two people play rock-paper-scissors in the video?
Visual Adapter
Visual Encoder
Video