1. Single-turn dialogue (ST)

Please give me a description of the frames …

Final video caption is …
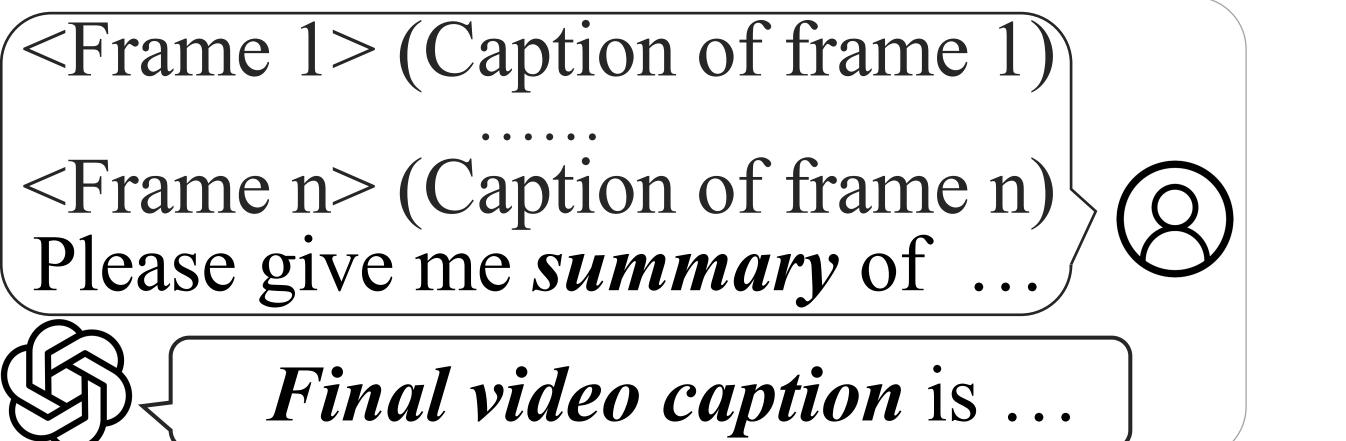
2. Multi-turn dialogue within the same context (MTSC)

Please describe this frame …

This frame shows …

………

Please describe this frame …

This frame shows …

<Frame 1> (Caption of frame 1)
……
<Frame n> (Caption of frame n)
Please give me ***summary*** of …

Final video caption is …

3. Multi-turn dialogue across different contexts (MTDC)

Please describe this frame …

This frame shows …

Caption of the ↓ previous frame

<The previous frame shows …>
Please describe this frame …

This frame shows …

Caption of the ↓ previous frame

………

Caption of the ↓ previous frame

<The previous frame shows …>
Please describe this frame …

This frame shows …

<Frame 1> (Caption of frame 1)
<Frame 2> (Caption of frame 2)
……
<Frame n> (Caption of frame n)
Please give me ***summary*** of …

Final video caption is …