

Video 1



Video 2



Caption The camera remains still, a boy with short black hair and wearing a white shirt is playing the violin, the background is a music room, soft afternoon lighting

Input Query

Your task is to compare two videos generated based on the same caption and provide an overall judgment along with a judgment for each dimension.

Video 1: First 4 input frames; Video2: Next 4 input frames.

The caption is: <caption>

Evaluation Dimensions:

1. Text Alignment : Alignment between video content and prompt
2. Video Quality : The visual aesthetics of the video
3. Movement Quality : Level of motion coherence

Frames and Analysis Rules: 8 sampled frames are provided, evenly downsampled from 128 frames; Insufficient frames? Request more

Format Requirement:

1. Snapshot: summarize any information that might be useful for your final judgment within <snapshot></snapshot> tags.
2. Think: Place all reasoning content within <think></think> tags.
3. Answer: output the answer within <Answer></Answer> tags. If final answer is uncertain, output the recommended answer and confidence level within <Recommend Answer></Recommend Answer> tags.

1 represents Video 1, 2 represents Video 2, and 0 represents Tie. The confidence levels range from high to low as 1, 2, and 3.



VideoSearch Reasoner With Initial Input

<snapshot>

Four input frames from Video 1 and Video 2 show a boy playing the violin in a music room. The lighting is soft, and the boy is wearing a white shirt

</snapshot>

<think>

..... it is difficult to draw conclusions at this stage. To further analyze, I will select additional frames to zoom in on the boy's movements and the overall visual quality, particularly focusing on the violin-playing dynamics between the first and second input frames.

</think>

<recommend answer>

TA = 2, MQ = 1, VQ = 0, OA = 0, CF = 2

</recommend answer>

<tool_call> { "name": "select_frames",
"arguments": { "target_frames": [12, 16, 20, 24,
60, 64, 68, 72] } } </tool_call>

Final Round?

Tool Call?

Video 1
RetrievedVideo 2
Retrieved

VideoSearch Reasoner With Retrieved Frames



<snapshot>

..... In Video 1, the boy's violin-playing actions appear blurry and the finger movements seem less natural. In Video 2, more dynamics are displayed, especially in the violin-playing posture, which appears smoother and more natural.....

</snapshot>

<think>

The selected frames confirm that Video 2 has a higher visual quality due to the clearer posture. The motion quality in Video 2 is good, and its alignment with the caption is consistent.....

</think>

<answer>

TA = 2, MQ = 0, VQ = 2, OA = 2, CF = 2

</ answer>

Final Round?

Tool Call?

