A. Constructed Data Sample and Prompting Details

To provide a clear understanding of our data construction process, we present a constructed data sample from the SKV-QA dataset along with the corresponding prompts used for caption generation and GPT-based scoring.

A.1. Constructed Data Sample

```
{
   "path":
       "/home/yl817/LLaVA-NeXT/hf_data/
       nextqa/NExTVideo/2574374895.mp4",
   "question": "what did the baby do
       after throwing the green cup away
       while on the floor near the end A.
       clap proudly B. the lady sitting
       down C. lay on floor D. just
       picked it up E. crawl",
   "debiased_question": "What did the
       baby do after throwing the green
       cup? Choices: Clap, Sit, Lay, Pick
       up, Crawl.",
   "captions": "{
      '[0, 7]': 'The clip begins with a
          child sitting in a blue chair,
          holding a green cup with a
          white G on it. The child is
          wearing a white shirt with
          colorful text and patterned
          shorts. The background shows a
          room with a bed that has a
          colorful blanket and some books
          or papers on it. The child
          appears to be playing with the
          cup, moving it around and
          looking at it. The child then
          leans forward, reaching towards
          the floor, possibly to pick
          something up or to place the
          cup down. The child continues
          to hold the cup while leaning
          forward. The scene remains
          consistent throughout the clip,
          with the child focused on the
          cup and the room's background
          unchanged.',

      '[8, 15]': 'The clip begins with a
          child in a white shirt and
          patterned shorts, crawling on
          an orange play mat. The child
          is reaching towards a blue toy
          chest adorned with colorful
          cartoon characters. The
          background includes a wooden
          cabinet with various items on
          top. As the child continues to
          crawl, they reach for a green
          cylindrical object on the
          floor. The scene progresses
          with the child still on the
          play mat, now closer to the
          green object. The child appears
          to be focused on the object,
          possibly attempting to pick it
          up or interact with it. The
          background remains consistent,
          showing the same wooden cabinet
          and items on top. The clip
          concludes with the child still
          engaged with the green object,
          maintaining their focus on it.',

      '[16, 23]': 'The clip begins with a
          child sitting on an orange
          floor, facing away from the
          camera. The child is wearing a
          white shirt and appears to be
          playing with a green object in
          their hands. In the background,
          there is a blue couch and a
          black object on the floor. As
          the clip progresses, the child
          continues to play with the
          green object, occasionally
          looking around. The child then
          reaches for the black object on
          the floor and picks it up. The
          clip concludes with the child
          holding the black object in
          their hand, while still sitting
          on the orange floor.',

      '[24, 29]': 'The clip begins with a
          child in a white shirt and
          colorful pants crawling on an
          orange floor. The background
          includes a blue piece of
          furniture and a patterned
          blanket. The child appears to
          be moving towards the
          furniture. The scene then
          transitions to the child lying
          on a patterned blanket, still
          wearing the same clothes, and
          seems to be resting or playing
          on the floor. A green cup is
          visible near the child. The
          child then starts to move,
          rolling over and sitting up
          slightly. The background
          remains consistent with the
          blue furniture and patterned
          blanket. The child continues to
          move around on the floor,
          occasionally looking around and
          adjusting their position.'
   }",
   "relevance_score": "{[24,29]: 4}"
}
```

Listing 1. Augmented data sample from SKV-QA

A.2. Caption Generation Prompt

```
System Prompt:\n### Task:\nYou are an
    expert in understanding scene
    transitions based on visual features
    in a video. You are requested to
    create the descriptions for the
```

```
    current clip sent to you, which
    includes multiple sequential frames.
2 ### Guidelines For Clip Description:\n-
    Analyze the narrative progression
    implied by the sequence of frames,
    interpreting the sequence as a whole.
3 - Note that since these frames are
    extracted from a clip, adjacent
    frames may show minimal differences.
4 - When referring to people, use their
    characteristics, such as clothing, to
    distinguish different people.
5 - **IMPORTANT** Please provide as many
    details as possible in your
    description, including colors,
    shapes, and textures.
6 ### Output Format:
7 Your response should look like this: The
    clip begins with..., progresses
    by..., and concludes with...
```

A.3. GPT Scoring Prompt

```
1  You are provided with descriptions of
     segments from a video. Each segment
     is labeled with a starting and ending
     frame index and a description of the
     events in that segment.
2  ### Instructions
3  1. Identify the relevancy between each
     segment and the question, assigning a
     score from 0 to 5 for all segments.
4  - 0 represents no relevancy, and 5
     represents the most relevant.
5  2. Sometimes the question may not be
     explicitly relevant to the
     descriptions. Consider the potential
     connection behind it.
6  3. Only return the starting and ending
     frame index for all segments and
     their corresponding scores.
7  4. Be mindful of the temporal
     relationship between segments and the
     question when scoring.
8  ### Output Format
9  Return the answer as a dictionary-like
     string:
10 Example output:
11 {{[0,6]:3,[7,20]:5}}
```

B. More Qualitative result

In this section, we present additional comparisons among standard token pruning, keyframe selection pipelines, and KVTP on LLaVA-Video-7B. To clearly highlight the differences, we sample eight representative frames from each video. The visualizations demonstrate that KVTP effectively preserves the complete scenes of events, leading to correct video reasoning outcomes while maintaining the same pruning rate.

C. Ablation study on context fusion head

To verify the effectiveness of context fusion head, we ablated the context fusion head and observed a clear performance drop, confirming its effectiveness. This ablation is conducted on the KVTP+Prumerge pipeline.

	VideoMME	EgoSchema	NeXT-QA
w/ head	63.3	54.7	76.7
w/o head	61.7	53.2	75.6

Table 6. Ablation study of the context fusion head on three SKV-QA subsets.

D. Results on the full datasets

To validate the generalizability of our method, we evaluated our method on the full versions of the three main datasets, along with two additional long-form datasets (*LongVideoBench* and *MVLU*), showing consistent improvements and confirming generalizability.

	1	2	3	4	5	FLOPs
LLaVA-Video-7B	63.3	57.3	83.2	70.8	58.2	100%
PruMerge	59.6	54.1	78.2	64.1	52.8	36%
KeyVideoLLM	52.1	51.9	73.5	61.9	50.0	36%
FastV	<u>61.8</u>	<u>56.7</u>	**81.8**	<u>66.1</u>	<u>56.7</u>	64%
KVTP+Prumerge (ours)	**62.7**	**57.2**	<u>80.5</u>	**66.2**	**57.3**	36%

Table 7. Evaluation results on five datasets: 1 = VideoMME, 2 = EgoSchema, 3 = NeXT-QA, 4 = MVLU, 5 = LongVideoBench.

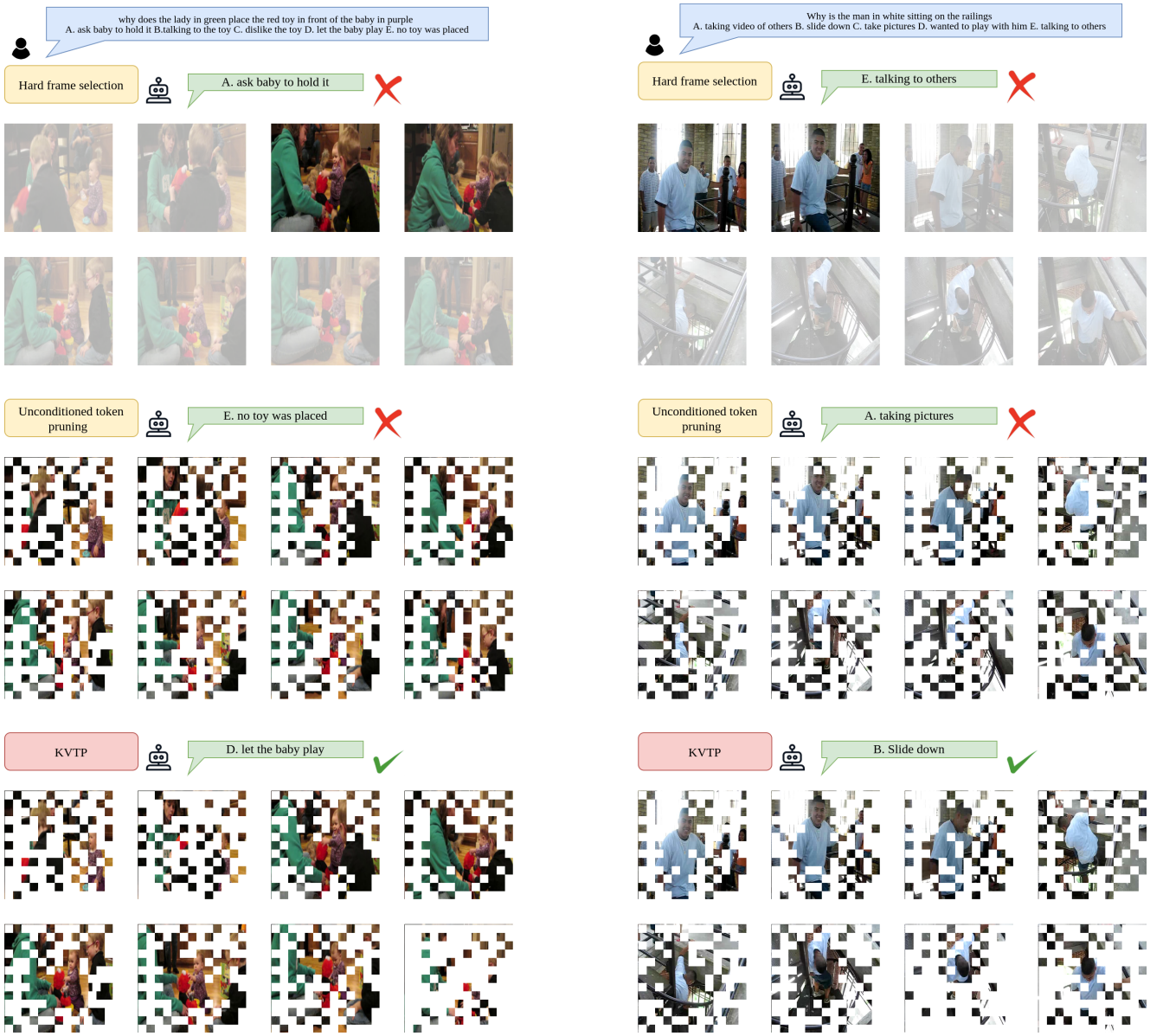

Figure 8. Qualitative results on LLaVA-Video-7B.