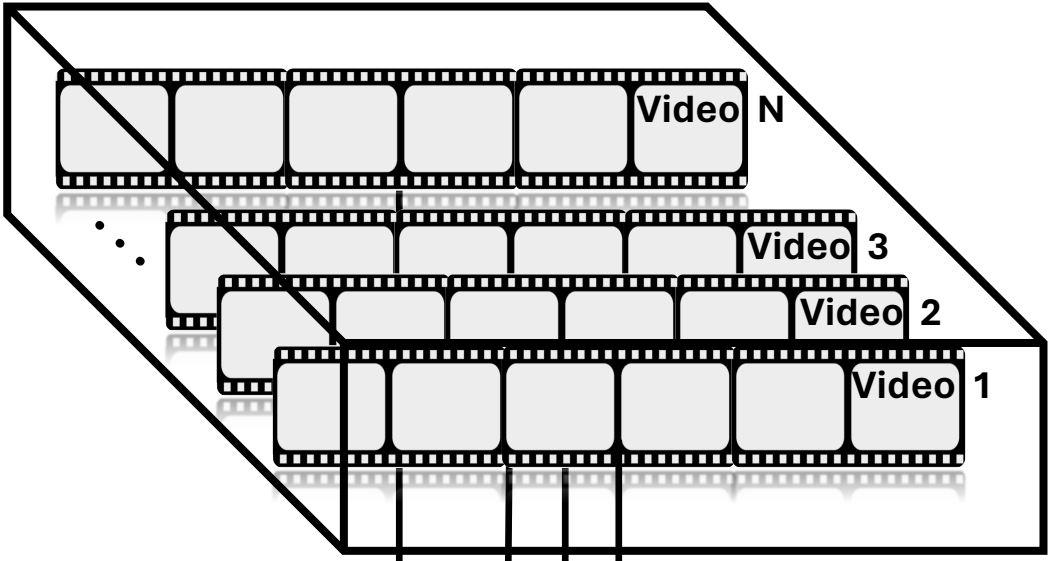

V2T-B (Video Caption Model)


```
"applejuice_N": [
  {
"applejuice_03": [
"applejuice_02": [
"applejuice_01": [
    {
      "start_sec": 0,
      "end_sec": 6,
      "image_caption": "a glass of juice next to two apples on a table",
      "dense_caption": "the blue vase on the left: [0.26, 0.1, 0.6, 0.99]; a purple onion on a table: [0.69, 0.0, 0.95, 0.39]; green apple in white bowl: [0.75, 0.3, 1.0, 0.81]; the top of the blue vase: [0.29, 0.1, 0.57, 0.45]; the vase is blue: [0.31, 0.26, 0.57, 0.84]",
      "audio_transcript": ""
    },.
.....
    {
      "start_sec": 87,
      "end_sec": 93,
      "image_caption": "a glass of juice next to two apples on a table",
      "dense_caption": "a purple onion on a table: [0.69, 0.0, 0.95, 0.39]; green apple in white bowl: [0.75, 0.29, 1.0, 0.81]; the vase is blue: [0.31, 0.26, 0.56, 0.83]; the top of the blue vase: [0.28, 0.1, 0.57, 0.45]",
      "audio_transcript": ""
    }
  ]
```


LLM

Fusion of Captioning


```
"ranked_captions_and_task_procedures": {
    "1": {
      "video_caption": "A person is pouring a glass of lemonade into a pitcher",
      "task_procedures": "<person> is <verb> pouring <action> a glass of lemonade into a pitcher."
    }
    "2": {
      "video_caption": "A person is pouring water into a blender with some apples in it",
      "task_procedures": "<person> is <verb> pouring <action> water into a blender with some apples in it."

.....

    "9": {
      "video_caption": "A person is pouring a liquid into a strainer on a counter",
      "task_procedures": "<person> is <verb> pouring <action> a liquid into a strainer on a counter."
    },
    "10": {
      "video_caption": "A person is holding a blender with their hand",
      "task_procedures": "<person> is <verb> holding <action> a blender with their hand."
    }
```