These supplemental materials encompass details of investigation of user search behavior feedback (§A), more related works (§B), the details of the dataset (§C), the details of the baselines (§E), more visualization quality (§D), response time of video retrieval (§F), limitations (§G), and quantitative results (§H).

## A DETAILS OF INVESTIGATION OF USER SEARCH BEHAVIOR FEEDBACK

To verify users' search needs for interactive retrieval, we design a user search behavior feedback survey. It aims to gain in-depth insights into users' real experiences during the search process, identifying potential issues and areas for improvement. The survey covers several aspects: user demographics (such as age, gender, occupation), video search habits, video search usage (whether users prefer to directly search for video moment), evaluation of search effectiveness (including the accuracy and usability of current video search systems, user satisfaction, and expectations for conversational search tools), and issues and improvement suggestions (open-ended questions for users to share problems and suggestions when using interactive search). To better accommodate the habits and preferences of users from different countries, we implemented a diversified survey approach. On the Amazon Mechanical Turk[3] platform, we conducted an online survey specifically designed to target an international audience, successfully engaging 500 participants from various regions. In China, we utilized the Wenjuanxing[4] platform to accurately reach and collect feedback from 500 Chinese users. Additionally, to gather more comprehensive and in-depth data, we organized offline paper surveys. Over the weekends, we collected insights from another 500 participants in high-traffic areas such as university campuses, shopping malls, parks, popular tourist destinations, and subway stations.

## B MORE RELATED WORK

### B.1 CROSS-MODAL VIDEO RETRIEVAL

**Video Retrieval.** Since TRECVid (Smeaton et al., 2006), the task of cross-modal video retrieval based on text queries has undergone a long development process. The goal of this task is to retrieve relevant videos from a set of video candidates given a text query. In this section, we focus on methods using the deep learning-based paradigm (Chen et al., 2020; Dong et al., 2019; Song et al., 2021; Luo et al., 2022). The basic idea is to encode texts and videos into embeddings, and then learn a common embedding space to do the matching between them. For text encoding, BERT-based models (Devlin et al., 2018) have become the mainstream. While for the video encoding in T2VR, a typical solution has recently evolved from Convolutional Neural Networks (CNN) (e.g., I3D (Carreira & Zisserman, 2017) and 3D ResNets (Hara et al., 2018)) to visual transformers (ViT) (e.g., ViT of CLIP (Radford et al., 2021) and TimeSformer (Bertasius et al., 2021). We can divide related methods into three categories. 1) *Pre-extracted multi-modality fusion* methods (Mithun et al., 2018; Liu et al., 2019; Gabeur et al., 2020; Wang et al., 2021) integrate rich multi-modality information (eg., motion, audio, and face) to improve the performance of T2VR. 2) *Joint text-video pre-training* methods (Wang et al., 2022a; Yan et al., 2023; Li et al., 2022; Ge et al., 2022) train the model with raw video and paired text in an end-to-end manner. 3) *Pre-trained CLIP-based* methods (Luo et al., 2022; Fang et al., 2021; Gorti et al., 2022; Liu et al., 2022b) use CLIP as a text-video backbone and adapt it to T2VR tasks. However, the pre-extracted multi-modality fusion methods are limited by the pre-extracted single-modal features, since these features are not particularly learnt for the target downstream tasks. The joint text-video pre-training methods achieve marginal improvements using a joint text-video pre-training paradigm due to the lack of large-scale text-video datasets. The pre-trained CLIP-based methods largely benefit from the pre-learned vision-text associations inherited frm CLIP and rely on naive mean-pooling or text-conditioned pooling to aggregate visual features.

**Video Moment Retrieval.** As an extension of video retrieval, video moment retrieval task aims to identify specific clips or moments within a video based on a given textual query (Gao et al., 2017; He et al., 2019). Pioneering works have explored various technical avenues, including attention-based retrieval, reinforcement learning, visual-language pretraining, among others. Liu et al. maintain

---

[3] https://www.mturk.com/

[4] https://www.wjx.cn/

a focus on retrieval paradigm, incorporating window segmentation as a preprocessing step, then retrieving moments. (He et al., 2019) devised an agent to pinpoint the start and end timestamps of moments based on reinforcement learning. Following this, some researchers expanded RL-based methods into the spatiotemporal or semantic domains (Cao et al., 2020). Moreover, leveraging pretraining techniques (Zeng et al., 2021; Pan & Zeng, 2023), several prompt-based models have emerged (Zeng, 2022; Liu et al., 2022a), facilitating timestamp prediction through regression.

### B.2 INTERACTIVE RETRIEVAL

The concept of interactive retrieval has long been proposed in the context of combining human-machine learning techniques for multimedia content search. Some studies(Thomee & Lew, 2012; Snoek et al., 2008)interactive have demonstrated that interactive retrieval can significantly improve search performance by enabling users to review search results and refine queries. With the significant progress of deep learning technology in the field of cross-modal video retrieval, interactive video retrieval has re-attracted the attention of researchers. Currently, only a few works(Madasu et al., 2022; Maeoki et al., 2020; Ma & Ngo, 2022; Liang & Albanie, 2023)simple have explored this task. For example, Madasu et al.(Madasu et al., 2022) and Maeoki et al.(Maeoki et al., 2020)adopt a dialogue-based approach, utilizing a series of video-related questions and answers generated by different models as retrieval queries. Meanwhile, addressing the issue that the above methods did not directly involve video question answering, Liang et al.(Liang & Albanie, 2023) employed a video question answering model to generate question-and-answer information in order to improve the retrieval accuracy. Furthermore, Ma et al.(Ma & Ngo, 2022)develop a user simulation for intelligent multimedia applications, leveraging advanced techniques in multimedia content analysis, including concept detection and cross-modal embedding, to enable precise video segment search through human-computer interaction. The aforementioned works are limited to achieving single-task interactive retrieval through methods such as reconstructing retrieval text, using visual question answering models, and simulating users. The technical challenges in modeling multi-turn dialogue retrieval have contributed to the slow development in this direction.

## C DETAILS OF DATASET

### C.1 DATASET INSTANCE

The instances in our dataset consist of four fields: (1) id: a unique identifier generated using the video name and a random number; (2) type: retrieval intent categorized into seven types—0: chat intent, 1: video retrieval intent, 2: video moment retrieval intent, 3: video to video moment retrieval intent, 4: video moment to video retrieval intent, 5: abstract search intent, 6: analogous search intent, 7: context-independent intent; (3) split: the dataset is divided into training, testing, and validation sets; (4) conversations: multi-turn retrieval formats where "from: human" indicates a query from a human, with the corresponding "value" and "Chinese_value" representing the query content in English and Chinese, respectively. "From: gpt" indicates feedback from GPT, with "gt" representing the retrieved video or video moment—note that "gt_se: [-1,1]" indicates video retrieval, while other values indicate video moment retrieval. Additionally, "video_source" indicates the dataset from which the video is sourced, with the corresponding value providing an interpretable description. Figure 8 shows an example in a unified format. These clearly defined fields allow benchmark users to flexibly construct the necessary training instances and easily evaluate the model.

### C.2 DIVERSITY QUALITY

We conducted an analysis of our video sources, the different types of videos (Figure 9), and performed a frequency analysis of annotated sentences (Figure 10 and 11) to ensure a comprehensive diversity. In addition, as shown in Figure 12, we present the statistics for video retrieval cases.

**Table 5: Comparing InterLLaVA performance across different data distributions**

| N training samples | Category | R@1 ↑ | R@1 IoU=0.5 ↑ | R@1 IoU=0.7 ↑ |
|---|---|---|---|---|
| 2K | Movies | 21.36 | 6.21 | 2.24 |
| 20K | TV shows | 26.73 | 8.24 | 3.98 |
| 1K | ALL | 36.6 | 8.49 | 4.94 |
| 6K | ALL | 45.7 | 11.31 | 5.8 |
| 20K | ALL | 47.96 | 11.52 | 6.45 |
| 60K | ALL | 54.86 | 12.28 | 7.13 |

### C.3 TRAINING DATA DISTRIBUTION ANALYSIS.

Our InterLLaVA model is trained on four different data modes, with the number of training samples ranging from 1K to 60K. Table 5 summarizes the performance evaluation results for all training samples. We observed the following points: 1) In low-sample scenarios, particularly when the sample size is less than 6K, InterLLaVA's accuracy is significantly limited (36.6 vs 54.86 for R@1), showing much lower performance compared to conditions with larger sample sizes. 2) We further explored training the model using samples from a single category (e.g., movies). The experimental results indicate that compared to training on data of the same scale but with more diverse categories, InterLLaVA's video and moment retrieval performance decreased by 21.23% and 3.28% in R@1 and R@1 IoU=0.5, respectively. This result aligns with expectations, as training on more diverse categories allows the model to capture richer features and enhance its generalization ability.

## D MORE VISUALIZATION QUALITY

We present more examples from our IVCR-200K dataset in Figures 13-21.

## E DETAILS OF BASELINES

For video retrieval, we selected the following five state-of-the-art models as benchmarks. We adopt their original setup, using both video and text as model inputs for the video retrieval task. CLIP4Clip(Luo et al., 2022) uses CLIP to extract the frame features and the text features, and then uses the mean pooling to aggregate the feature of all frames for video representation. X-Pool(Gorti et al., 2022) adopts text-conditioned pooling to aggregate visual features. TS2-Net(Liu et al., 2022b) proposes different token shift operations in ViT to learn short-term temporal dependencies across locally adjacent frames. T-MASS(Wang et al., 2024) proposes a stochastic modeling approach to achieve expressive and flexible text embeddings, enhancing the alignment of text and video semantics in the joint space. Furthermore, we reimplement a video retrieval model named BLIP-2, utilizing the video and text features encoded by BLIP-2(Li et al., 2023a), with X-Pool(Gorti et al., 2022) serving as the base model.

For video moment retrieval, we selected six methods as benchmarks. We utilize BLIP-2(Li et al., 2023a) as the encoder to extract video and text features, which are then used as inputs for all models in video moment retrieval. 2D-TAN(Zhang et al., 2020) proposes a novel two-dimensional temporal matrix for moment localization. MMN(Wang et al., 2022b) introduces a mutual matching network that directly models the similarity between language queries and video moments within a joint embedding space. UMT(Liu et al., 2022a) proposes a unified framework for solving joint moment retrieval and highlight detection. CG-DETR(Moon et al., 2023) explores the provision of cues for query-associated video clips within cross-modal attention. MomentDiff(Li et al., 2024a) utilizes diffusion models to diffuse real span to random noise, and then learns to denoise the random noise back to the original span under the guidance of text and video similarity. Moreover, we chose a model based on multi-modal large language models as additional benchmarks. TimeChat(Ren et al., 2023)

**Table 6: Comparing average response times of the model across different test data scales**

| N Testing Samples | Average Response Time of Video Retrieval(s) | Average Response Time of MLLM (s) | Total Average Response Time (s) |
|---|---|---|---|
| 1008 | 0.06 | 0.67 | 0.73 |
| 1534 | 0.06 | 0.59 | 0.65 |
| 1878 | 0.06 | 0.73 | 0.79 |
| 2436 | 0.06 | 0.8 | 0.86 |

proposes a time-sensitive multimodal large language model for long video understanding and precise temporal localization.

## F RESPONSE TIME OF VIDEO RETRIEVAL

We compared the average response times of video retrieval and Multi-Modal Large Language Model (MLLM) inference, and the results are summarized in the Table 6. Our observations are as follows: 1) The average response time for video retrieval is notably lower compared to the MLLM inference. However, the average response time of the MLLM remains below 1 second, which is acceptable. 2) Testing with different numbers of video samples revealed a slight increase in the average response time for video retrieval. Meanwhile, the average response time of the MLLM remains almost constant, demonstrating that our model does not introduce significant delays even when handling larger volumes of video data.

## G LIMITATIONS

The IVCR-200K dataset is constrained by the depth of manual annotation and the diversity of real-world data types. It needs to be expanded to cover a wider array of interactive retrieval scenarios, including complex analogy searches, diverse contextual searches, and fine-grained interactive search requirements. Additionally, the current model does not achieve seamless integration of video retrieval and moment retrieval into a unified, efficient end-to-end system. There is considerable potential for improvement in areas such as temporal video modeling, accurate capture of user retrieval intent, and the natural and fluid execution of multi-round dialogues.

## H QUANTITATIVE RESULTS

Figures22-25 presents a qualitative comparison between InterLLaVA and other video large language models. Our observations are as follows: 1) Video-LLaVA(Lin et al., 2023) has limitations in handling video retrieval, as it is limited to describing the direct content relationship between video and text. It lacks intent analysis for text retrieval and cannot provide interpretable feedback for identifying relevant videos and moment. 2) In contrast, TimeChat(Ren et al., 2023) merely offers the start and end timestamps of videos, lacking any form of feedback on the retrieval outcomes, thus insufficient to cater to users' personalized search demands. 3) InterLLaVA excels at accurately matching the desired retrieval video, precisely locating specific moments within videos, and providing interpretable feedback. However, there is still potential to further enhance the richness and detail of the feedback generated by our model.

```
{
    "id": "v_n9eNF1WqFu",
    "type": [1,2,4], #video-moment-similar
    "split": "train",
    "conversations": [
      {
        "from": "human",
        "value": "Person␣in␣white␣clothes␣are␣dancing.",
        "Chinese_value": "穿白色衣服的人在跳舞。",
        "text_id": "247775f9"
      },
      {
        "from": "gpt",
        "gt": ["s234d9"],
        "gt_se": [-1, -1],
        "video_source": "ActivityNet",
        "value": "The␣intent␣of␣this␣query␣is␣to␣find␣a␣complete␣video.",
        "Chinese_value": "这个查询的意图是找到一个完整的视频。"
      },
      {
        "from": "human",
        "value": "Locate␣a␣video␣moment␣of␣the␣person␣spinning␣in␣circles.",
        "Chinese_value": "定位视频中该人转圈的瞬间。",
        "text_id": "0c1dce9c"
      },
      {
        "from": "gpt",
        "gt": ["s234d9"],
        "gt_se": [1.3, 5.6],
        "video_source": "ActivityNet",
        "value": "The␣intent␣of␣this␣query␣is␣to␣further␣locate␣a␣spinning␣
            moment.",
        "Chinese_value": "这个查询的意图是进一步定位转圈的瞬间。"
      },
      {
        "from": "human",
        "value": "Is␣there␣any␣similar␣moment?",
        "Chinese_value": "有没有类似的瞬间？",
        "text_id": "5553fa8a"
      },
      {
        "from": "gpt",
        "gt": ["45866g"],
        "gt_se": [-1, -1],
        "video_source": "MSR-VTT",
        "value": "The␣’similar’␣in␣the␣query␣may␣refer␣to␣videos␣of␣dancing␣
            in␣circles.",
        "Chinese_value": "查询中的类似可能指的是在圆圈中跳舞的视频。"
      }
    ]
}
```

**Figure 8: Example JSON of our IVCR-200K dataset.**

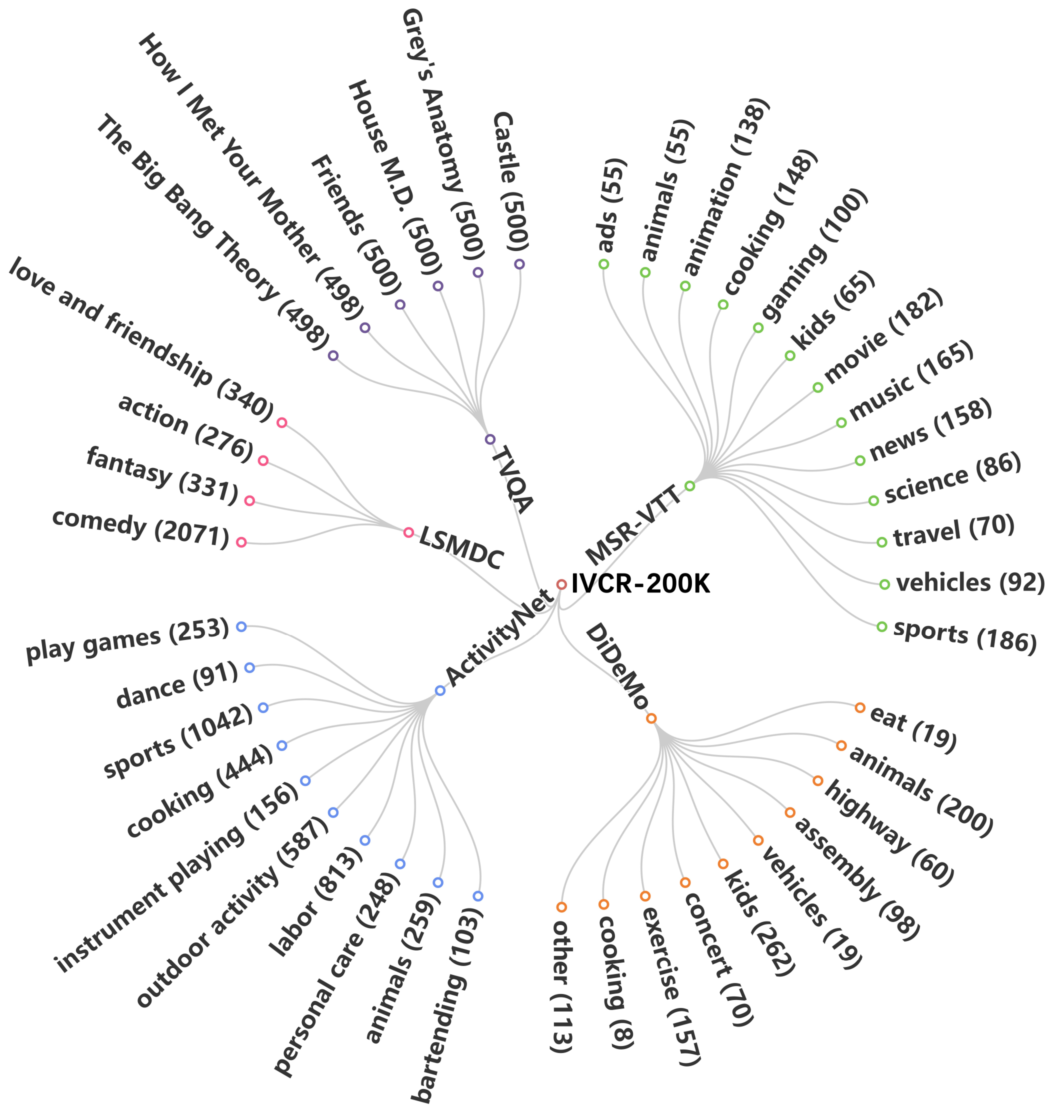


**Figure 9: The hierarchical structure of the sources and categories of IVCR-200K.**

man
woman
people
talking
camera
person
House
Beckett
patient
water
room
shown
girl
Castle
table
playing
sitting
walks
boy
Ross
Sheldon
Chandler
holding
SOMEONE
showing
speaking
men
puts
video
door
Barney
lady
Marshall
Leonard
Monica
Howard
demonstrates
hospital
riding
walking
walks
face
picks
something
throws
team
movie
cutting
doctor
Richard
mouth
Bernadette
gymnast
watch
Robin
dog
front
head
Derek
Ryan
chair
Lily
bed
paper
Penny
gets
game
Ted
Raj
phone
George
someone
Burke
hands
bag
Cameron
court
dress
driving
discussing
going
orange
walked
says
times
places
children
asks
get
mixing
office
taking
flips
song
moves
pound
begins
looking
scene
arms
counter
plays
points
falls
enters
baby
news
watching
road
child
stage
Alex
adding
explains
comes
moving
dirt
street
audience
running
pours
Wilson
pizza
jumps
walk
tells
cup
box
Later
does
players
cake
Chase
match
adds
bike
Bailey
beam
pot
pan
uses
crowd
runs
was
food
drink
music
mat
beach
ground
cartoon
piece
talks
credits
Foreman
jacket
bottle
Cristina
conversation
starts
do
talks
seen
takes
car
bar
shirt
hand
cat
Joey
stands
ingredients
kitchen
time
opens
guy
others
making
performing
dancing
horse
Meredith
standing
screen
clip
sits
hands
group
Rachel
holds
glass
board
Phoebe
women
followed
appears
leaves
boat
brush
makes
including
coffee
gives
bowl
show
doing
Amy
arm
speaks
back
sits
rope
knife
getting
glasses
jumping
bars
tries
ends
tricks
stops
title
performs
put
track
player
Esposito
plate
Cuddy
continues
goes
Addison
couch
make
removes
pool
towel
ending
seat
giving
discusses
have
shots
gun
machine
talk
People
tree
coat
see
ice
top
tv
air
turns
looks
wearing
kids
side
singing
wall
desk
field
sink
end
putting
suit
jumps
girls
prepares
floor
tire
using
ball
bow
oil
Dr.
hair
having
toy
apartment
seen
being

**Figure 10: Word cloud of textual queries from IVCR-200K.**

男人
正在
他们
展示
视频
同时
男子
女人
一些
开始
交谈
说话
穿着
如何
使用
那个
我们
房间
最后
人们
看到
镜头
女孩
贝克特
女士
现在
进行
放在
两个
一起
孩子
离开
讨论
游戏
一边
比赛
走进
桌子
旁边
提供
结束
屏幕
自己
完成
通过
谈论
病人
继续
随后
谢尔顿
看着
打开
走向
其他
这些
巴尼
它们
那里
椅子
妇女
后面
以下
食材
互动
产品
家伙
面前
绿色
之前
递给
过程
观众
角色
技巧
照片
再次
地板
用户
自行车
工作
创建
运动员
制作
衬衫
显示
内容
沿着
衣服
工具
乔伊
乌歇尔
加入
莫斯
各种
黑色
房子
告诉
男孩
罗宾
偏妮
移动
饮料
检查
那位
文件
平台
动画
活动
埃斯波
汽车
可能
泰德
罗斯
莉莉
解释
瑞秋
年轻
白色
几个
场景
东西
克里斯蒂娜
拥抱
床上
聊天
放下
进入
微笑
其中
问题
公寓
需要
红色
帮助
动作
电影
准备
关于
某人
主要
包括
以及
马歇尔
电话
女性
舞蹈
摄像机
伦纳德
一组
之后
沙发
食物
演示
接着
旋转
下来
体操
片段
出来
互相
画面
不同
乔治
医生
餐厅
城堡
蛋糕
杯子
盘子
飞盘
一下
厨房
菲比
搅拌
瑞安
讲话
手里
走开
头发
倒入
华盛顿
看起来
霍华德
一次
麦克
音乐
起来

**Figure 11: Chinese Word cloud of textual queries from IVCR-200K.**

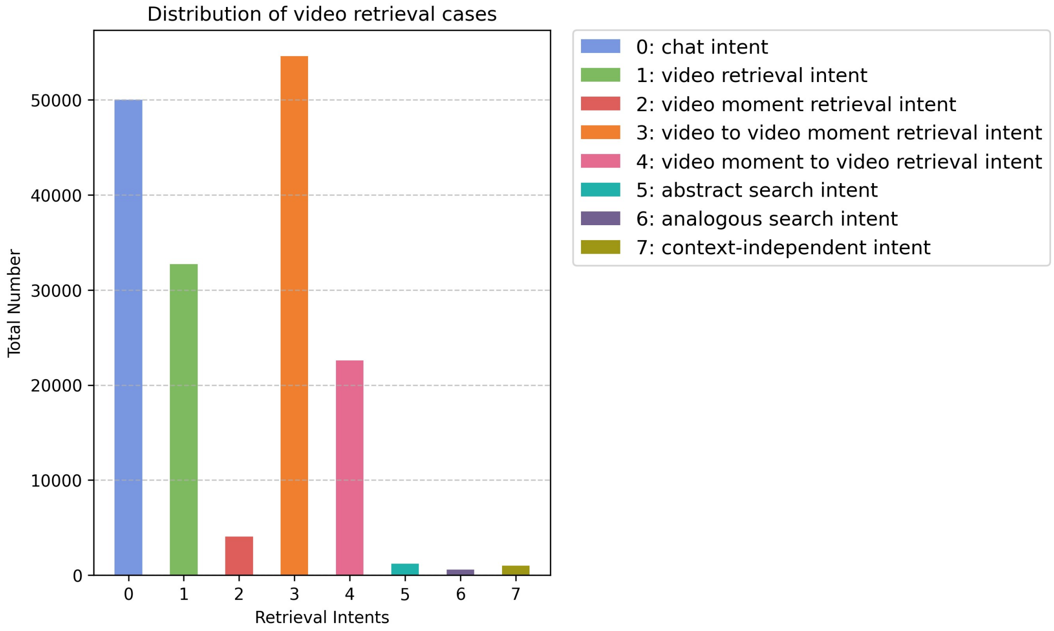
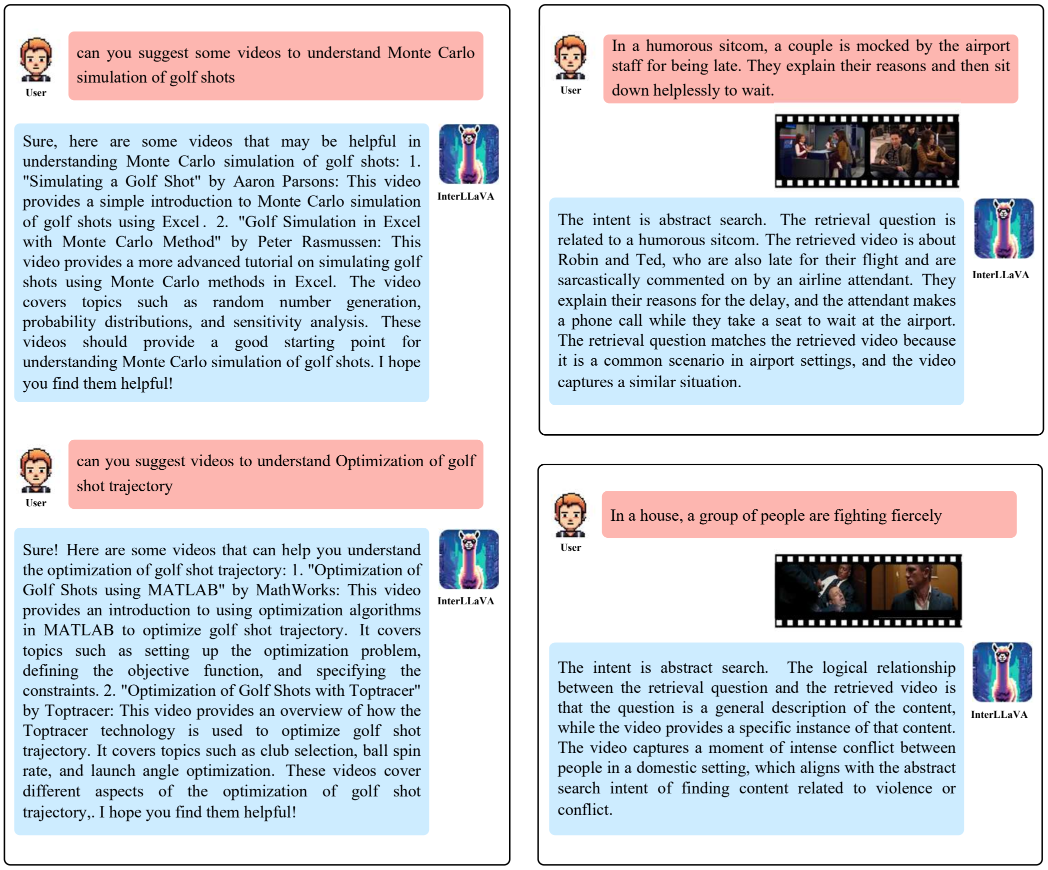


**Figure 12: Distribution of video retrieval cases.**



can you suggest some videos to understand Monte Carlo simulation of golf shots

Sure, here are some videos that may be helpful in understanding Monte Carlo simulation of golf shots: 1. "Simulating a Golf Shot" by Aaron Parsons: This video provides a simple introduction to Monte Carlo simulation of golf shots using Excel . 2. "Golf Simulation in Excel with Monte Carlo Method" by Peter Rasmussen: This video provides a more advanced tutorial on simulating golf shots using Monte Carlo methods in Excel. The video covers topics such as random number generation, probability distributions, and sensitivity analysis. These videos should provide a good starting point for understanding Monte Carlo simulation of golf shots. I hope you find them helpful!

can you suggest videos to understand Optimization of golf shot trajectory

Sure! Here are some videos that can help you understand the optimization of golf shot trajectory: 1. "Optimization of Golf Shots using MATLAB" by MathWorks: This video provides an introduction to using optimization algorithms in MATLAB to optimize golf shot trajectory. It covers topics such as setting up the optimization problem, defining the objective function, and specifying the constraints. 2. "Optimization of Golf Shots with Toptracer" by Toptracer: This video provides an overview of how the Toptracer technology is used to optimize golf shot trajectory. It covers topics such as club selection, ball spin rate, and launch angle optimization. These videos cover different aspects of the optimization of golf shot trajectory,. I hope you find them helpful!

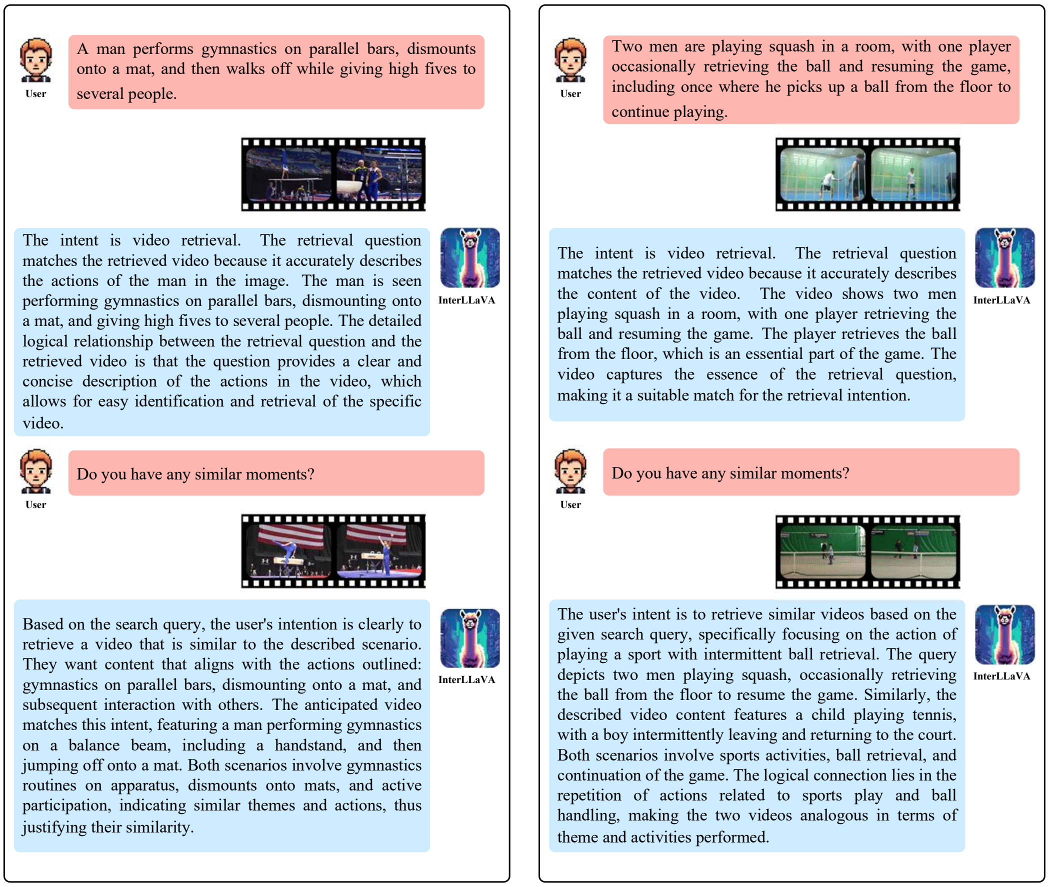
In a humorous sitcom, a couple is mocked by the airport staff for being late. They explain their reasons and then sit down helplessly to wait.

The intent is abstract search. The retrieval question is related to a humorous sitcom. The retrieved video is about Robin and Ted, who are also late for their flight and are sarcastically commented on by an airline attendant. They explain their reasons for the delay, and the attendant makes a phone call while they take a seat to wait at the airport. The retrieval question matches the retrieved video because it is a common scenario in airport settings, and the video captures a similar situation.

In a house, a group of people are fighting fiercely

The intent is abstract search. The logical relationship between the retrieval question and the retrieved video is that the question is a general description of the content, while the video provides a specific instance of that content. The video captures a moment of intense conflict between people in a domestic setting, which aligns with the abstract search intent of finding content related to violence or conflict.

**Figure 13: An example of abstract concept retrieval and dialogue from the IVCR-200K dataset.**



A man performs gymnastics on parallel bars, dismounts onto a mat, and then walks off while giving high fives to several people.

The intent is video retrieval. The retrieval question matches the retrieved video because it accurately describes the actions of the man in the image. The man is seen performing gymnastics on parallel bars, dismounting onto a mat, and giving high fives to several people. The detailed logical relationship between the retrieval question and the retrieved video is that the question provides a clear and concise description of the actions in the video, which allows for easy identification and retrieval of the specific video.

Do you have any similar moments?

Based on the search query, the user's intention is clearly to retrieve a video that is similar to the described scenario. They want content that aligns with the actions outlined: gymnastics on parallel bars, dismounting onto a mat, and subsequent interaction with others. The anticipated video matches this intent, featuring a man performing gymnastics on a balance beam, including a handstand, and then jumping off onto a mat. Both scenarios involve gymnastics routines on apparatus, dismounts onto mats, and active participation, indicating similar themes and actions, thus justifying their similarity.

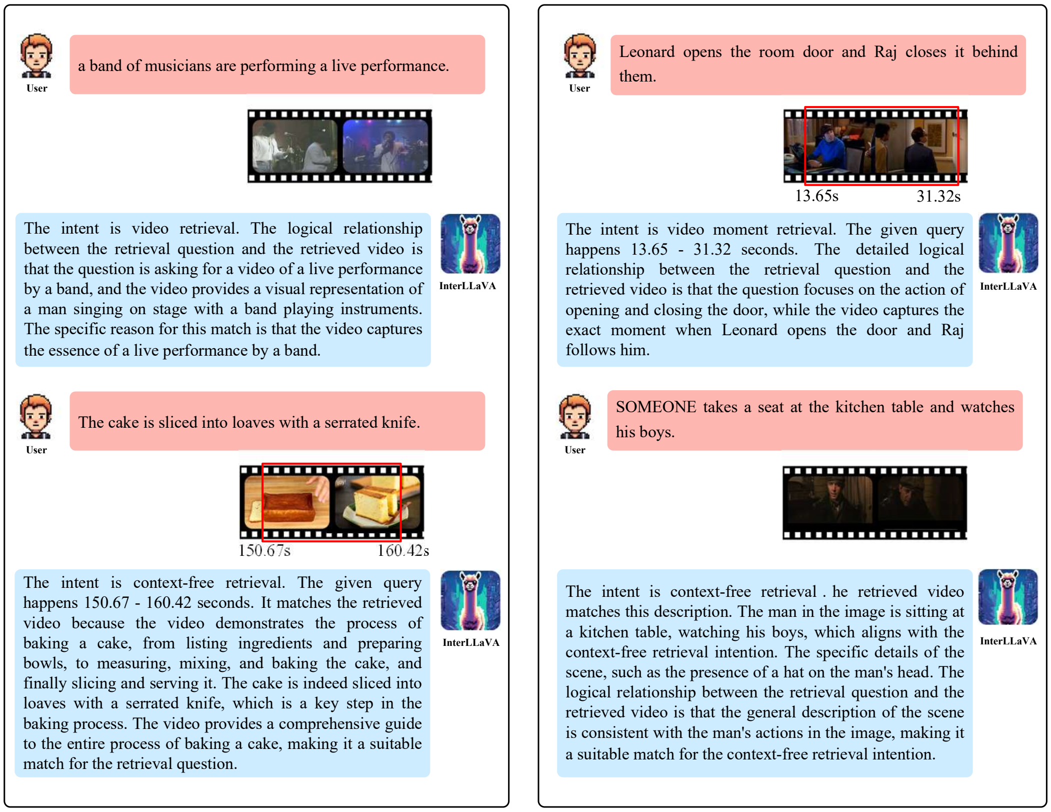
Two men are playing squash in a room, with one player occasionally retrieving the ball and resuming the game, including once where he picks up a ball from the floor to continue playing.

The intent is video retrieval. The retrieval question matches the retrieved video because it accurately describes the content of the video. The video shows two men playing squash in a room, with one player retrieving the ball and resuming the game. The player retrieves the ball from the floor, which is an essential part of the game. The video captures the essence of the retrieval question, making it a suitable match for the retrieval intention.

Do you have any similar moments?

The user's intent is to retrieve similar videos based on the given search query, specifically focusing on the action of playing a sport with intermittent ball retrieval. The query depicts two men playing squash, occasionally retrieving the ball from the floor to resume the game. Similarly, the described video content features a child playing tennis, with a boy intermittently leaving and returning to the court. Both scenarios involve sports activities, ball retrieval, and continuation of the game. The logical connection lies in the repetition of actions related to sports play and ball handling, making the two videos analogous in terms of theme and activities performed.

**Figure 14: An example of analogy retrieval from the IVCR-200K dataset.**



a band of musicians are performing a live performance.

The intent is video retrieval. The logical relationship between the retrieval question and the retrieved video is that the question is asking for a video of a live performance by a band, and the video provides a visual representation of a man singing on stage with a band playing instruments. The specific reason for this match is that the video captures the essence of a live performance by a band.

The cake is sliced into loaves with a serrated knife.

The intent is context-free retrieval. The given query happens 150.67 - 160.42 seconds. It matches the retrieved video because the video demonstrates the process of baking a cake, from listing ingredients and preparing bowls, to measuring, mixing, and baking the cake, and finally slicing and serving it. The cake is indeed sliced into loaves with a serrated knife, which is a key step in the baking process. The video provides a comprehensive guide to the entire process of baking a cake, making it a suitable match for the retrieval question.

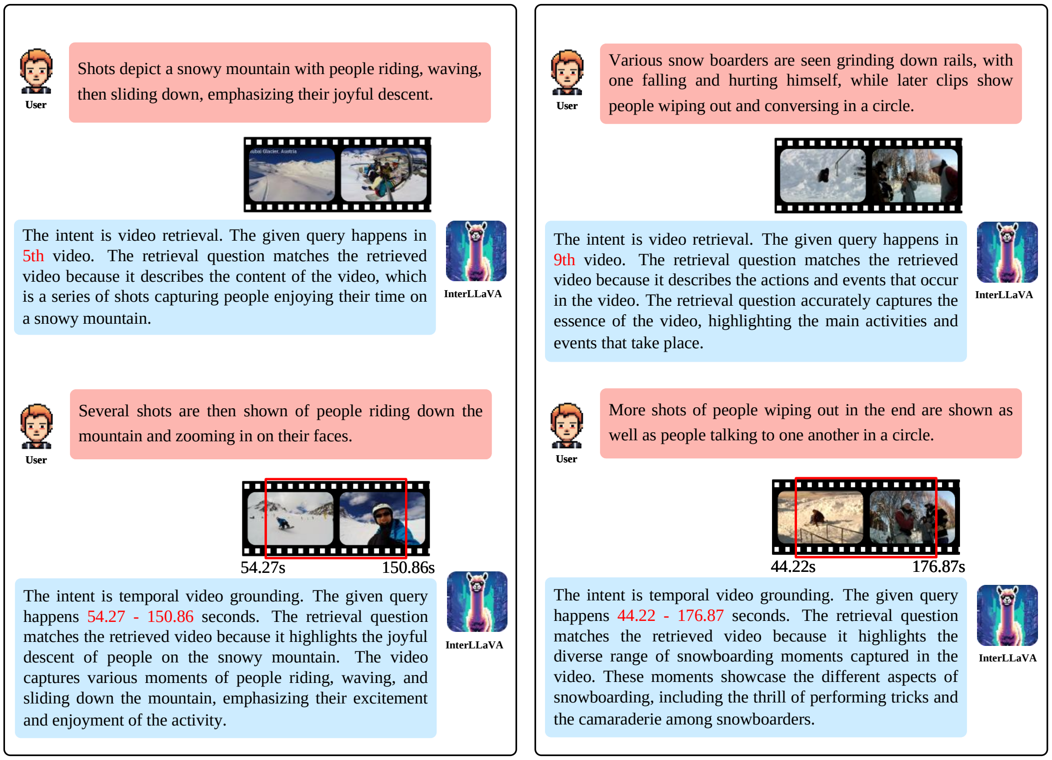
Leonard opens the room door and Raj closes it behind them.

The intent is video moment retrieval. The given query happens 13.65 - 31.32 seconds. The detailed logical relationship between the retrieval question and the retrieved video is that the question focuses on the action of opening and closing the door, while the video captures the exact moment when Leonard opens the door and Raj follows him.

SOMEONE takes a seat at the kitchen table and watches his boys.

The intent is context-free retrieval . he retrieved video matches this description. The man in the image is sitting at a kitchen table, watching his boys, which aligns with the context-free retrieval intention. The specific details of the scene, such as the presence of a hat on the man's head. The logical relationship between the retrieval question and the retrieved video is that the general description of the scene is consistent with the man's actions in the image, making it a suitable match for the context-free retrieval intention.

**Figure 15: An example of context-independent retrieval from the IVCR-200K dataset.**



Shots depict a snowy mountain with people riding, waving, then sliding down, emphasizing their joyful descent.

The intent is video retrieval. The given query happens in 5th video. The retrieval question matches the retrieved video because it describes the content of the video, which is a series of shots capturing people enjoying their time on a snowy mountain.

Several shots are then shown of people riding down the mountain and zooming in on their faces.

The intent is temporal video grounding. The given query happens 54.27 - 150.86 seconds. The retrieval question matches the retrieved video because it highlights the joyful descent of people on the snowy mountain. The video captures various moments of people riding, waving, and sliding down the mountain, emphasizing their excitement and enjoyment of the activity.

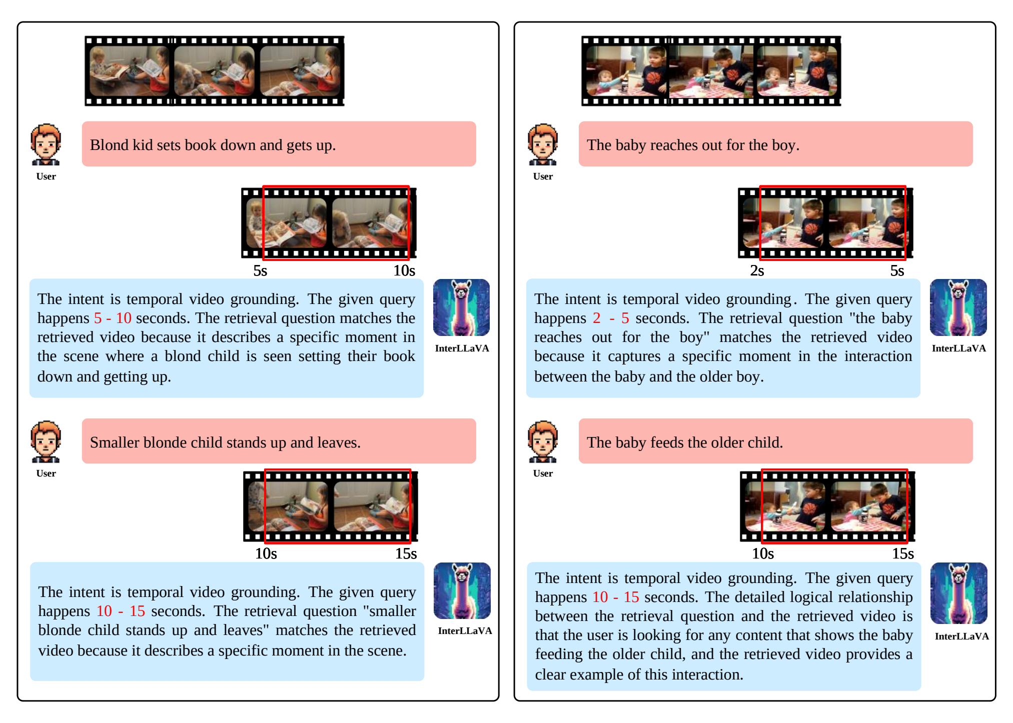
Various snow boarders are seen grinding down rails, with one falling and hurting himself, while later clips show people wiping out and conversing in a circle.

The intent is video retrieval. The given query happens in 9th video. The retrieval question matches the retrieved video because it describes the actions and events that occur in the video. The retrieval question accurately captures the essence of the video, highlighting the main activities and events that take place.

More shots of people wiping out in the end are shown as well as people talking to one another in a circle.

The intent is temporal video grounding. The given query happens 44.22 - 176.87 seconds. The retrieval question matches the retrieved video because it highlights the diverse range of snowboarding moments captured in the video. These moments showcase the different aspects of snowboarding, including the thrill of performing tricks and the camaraderie among snowboarders.

**Figure 16: An example of video-to-video moment retrieval from the IVCR-200K dataset.**



Blond kid sets book down and gets up.

The intent is temporal video grounding. The given query happens 5 - 10 seconds. The retrieval question matches the retrieved video because it describes a specific moment in the scene where a blond child is seen setting their book down and getting up.

Smaller blonde child stands up and leaves.

The intent is temporal video grounding. The given query happens 10 - 15 seconds. The retrieval question "smaller blonde child stands up and leaves" matches the retrieved video because it describes a specific moment in the scene.

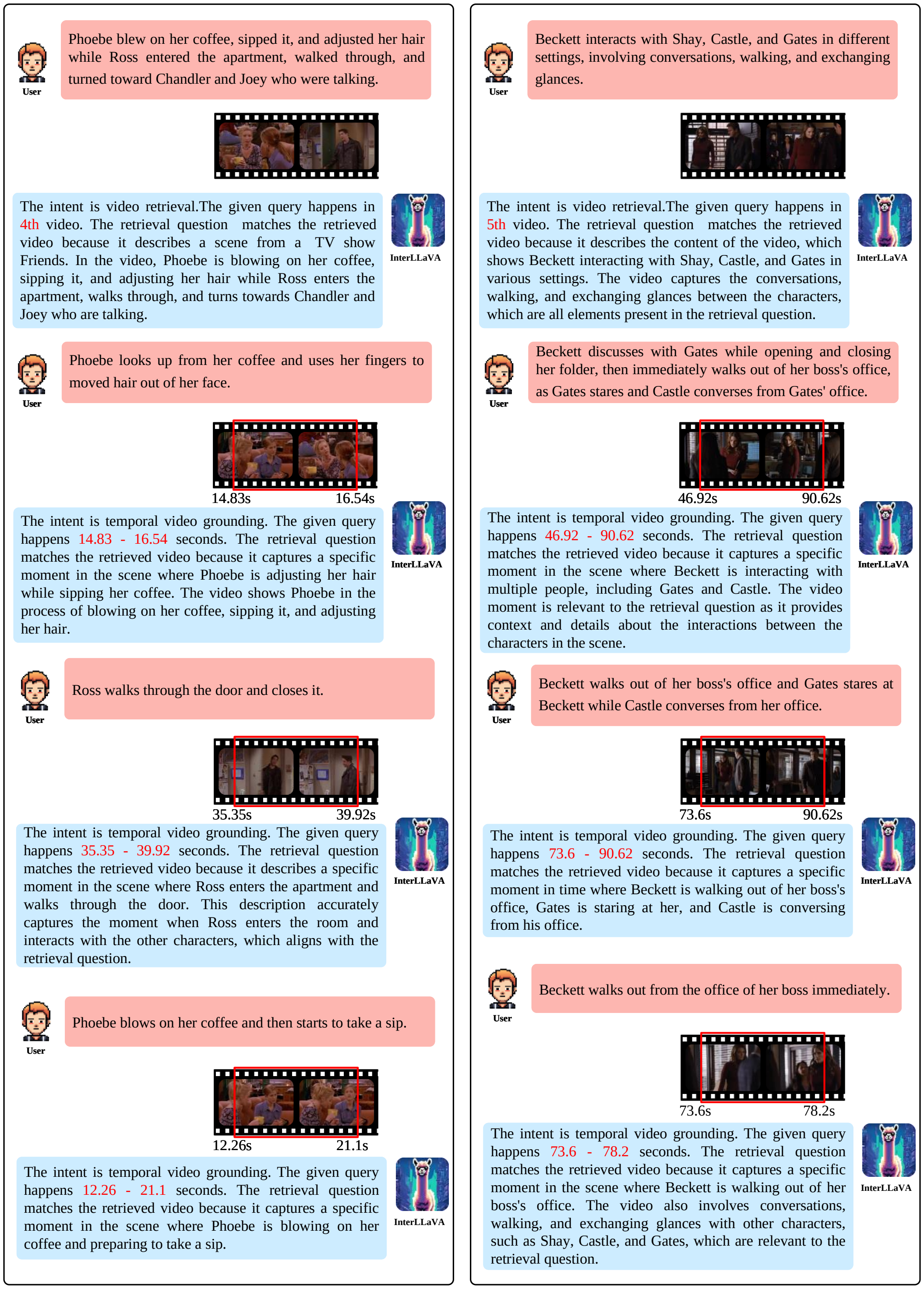
The baby reaches out for the boy.

The intent is temporal video grounding. The given query happens 2 - 5 seconds. The retrieval question "the baby reaches out for the boy" matches the retrieved video because it captures a specific moment in the interaction between the baby and the older boy.

The baby feeds the older child.

The intent is temporal video grounding. The given query happens 10 - 15 seconds. The detailed logical relationship between the retrieval question and the retrieved video is that the user is looking for any content that shows the baby feeding the older child, and the retrieved video provides a clear example of this interaction.

**Figure 17: An example of video-to-video moment retrieval from the IVCR-200K dataset.**



Phoebe blew on her coffee, sipped it, and adjusted her hair while Ross entered the apartment, walked through, and turned toward Chandler and Joey who were talking.

The intent is video retrieval.The given query happens in 4th video. The retrieval question matches the retrieved video because it describes a scene from a TV show Friends. In the video, Phoebe is blowing on her coffee, sipping it, and adjusting her hair while Ross enters the apartment, walks through, and turns towards Chandler and Joey who are talking.

Phoebe looks up from her coffee and uses her fingers to moved hair out of her face.

The intent is temporal video grounding. The given query happens 14.83 - 16.54 seconds. The retrieval question matches the retrieved video because it captures a specific moment in the scene where Phoebe is adjusting her hair while sipping her coffee. The video shows Phoebe in the process of blowing on her coffee, sipping it, and adjusting her hair.

Ross walks through the door and closes it.

The intent is temporal video grounding. The given query happens 35.35 - 39.92 seconds. The retrieval question matches the retrieved video because it describes a specific moment in the scene where Ross enters the apartment and walks through the door. This description accurately captures the moment when Ross enters the room and interacts with the other characters, which aligns with the retrieval question.

Phoebe blows on her coffee and then starts to take a sip.

The intent is temporal video grounding. The given query happens 12.26 - 21.1 seconds. The retrieval question matches the retrieved video because it captures a specific moment in the scene where Phoebe is blowing on her coffee and preparing to take a sip.

Beckett interacts with Shay, Castle, and Gates in different settings, involving conversations, walking, and exchanging glances.

The intent is video retrieval.The given query happens in 5th video. The retrieval question matches the retrieved video because it describes the content of the video, which shows Beckett interacting with Shay, Castle, and Gates in various settings. The video captures the conversations, walking, and exchanging glances between the characters, which are all elements present in the retrieval question.

Beckett discusses with Gates while opening and closing her folder, then immediately walks out of her boss's office, as Gates stares and Castle converses from Gates' office.

The intent is temporal video grounding. The given query happens 46.92 - 90.62 seconds. The retrieval question matches the retrieved video because it captures a specific moment in the scene where Beckett is interacting with multiple people, including Gates and Castle. The video moment is relevant to the retrieval question as it provides context and details about the interactions between the characters in the scene.

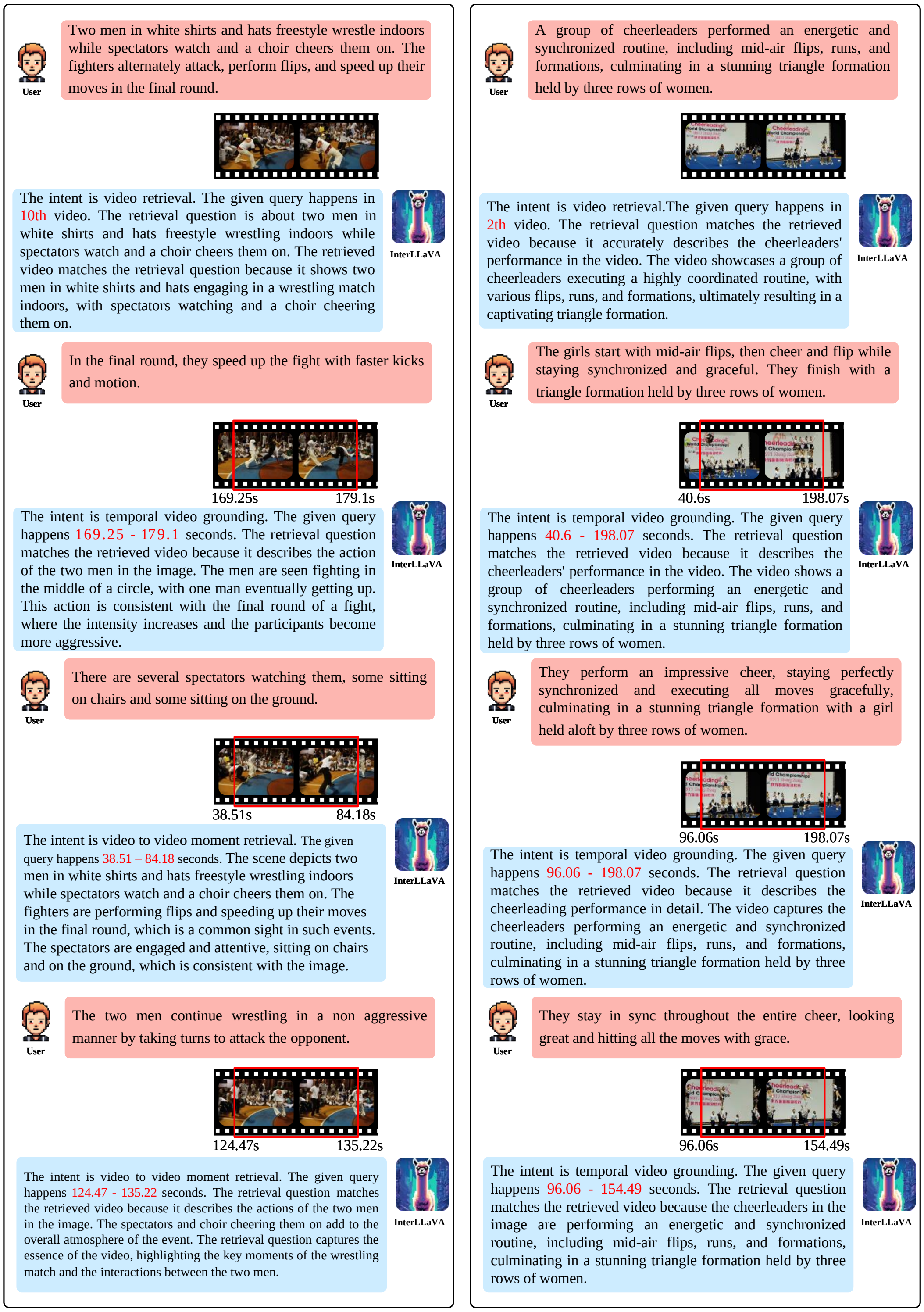
Beckett walks out of her boss's office and Gates stares at Beckett while Castle converses from her office.

The intent is temporal video grounding. The given query happens 73.6 - 90.62 seconds. The retrieval question matches the retrieved video because it captures a specific moment in time where Beckett is walking out of her boss's office, Gates is staring at her, and Castle is conversing from his office.

Beckett walks out from the office of her boss immediately.

The intent is temporal video grounding. The given query happens 73.6 - 78.2 seconds. The retrieval question matches the retrieved video because it captures a specific moment in the scene where Beckett is walking out of her boss's office. The video also involves conversations, walking, and exchanging glances with other characters, such as Shay, Castle, and Gates, which are relevant to the retrieval question.

**Figure 18: An example of video-to-video moment retrieval from the IVCR-200K dataset.**



Two men in white shirts and hats freestyle wrestle indoors while spectators watch and a choir cheers them on. The fighters alternately attack, perform flips, and speed up their moves in the final round.

InterLLaVA: The intent is video retrieval. The given query happens in 10th video. The retrieval question is about two men in white shirts and hats freestyle wrestling indoors while spectators watch and a choir cheers them on. The retrieved video matches the retrieval question because it shows two men in white shirts and hats engaging in a wrestling match indoors, with spectators watching and a choir cheering them on.

In the final round, they speed up the fight with faster kicks and motion.

InterLLaVA: The intent is temporal video grounding. The given query happens 169.25 - 179.1 seconds. The retrieval question matches the retrieved video because it describes the action of the two men in the image. The men are seen fighting in the middle of a circle, with one man eventually getting up. This action is consistent with the final round of a fight, where the intensity increases and the participants become more aggressive.

There are several spectators watching them, some sitting on chairs and some sitting on the ground.

InterLLaVA: The intent is video to video moment retrieval. The given query happens 38.51 – 84.18 seconds. The scene depicts two men in white shirts and hats freestyle wrestling indoors while spectators watch and a choir cheers them on. The fighters are performing flips and speeding up their moves in the final round, which is a common sight in such events. The spectators are engaged and attentive, sitting on chairs and on the ground, which is consistent with the image.

The two men continue wrestling in a non aggressive manner by taking turns to attack the opponent.

InterLLaVA: The intent is video to video moment retrieval. The given query happens 124.47 - 135.22 seconds. The retrieval question matches the retrieved video because it describes the actions of the two men in the image. The spectators and choir cheering them on add to the overall atmosphere of the event. The retrieval question captures the essence of the video, highlighting the key moments of the wrestling match and the interactions between the two men.

A group of cheerleaders performed an energetic and synchronized routine, including mid-air flips, runs, and formations, culminating in a stunning triangle formation held by three rows of women.

InterLLaVA: The intent is video retrieval.The given query happens in 2th video. The retrieval question matches the retrieved video because it accurately describes the cheerleaders' performance in the video. The video showcases a group of cheerleaders executing a highly coordinated routine, with various flips, runs, and formations, ultimately resulting in a captivating triangle formation.

The girls start with mid-air flips, then cheer and flip while staying synchronized and graceful. They finish with a triangle formation held by three rows of women.

InterLLaVA: The intent is temporal video grounding. The given query happens 40.6 - 198.07 seconds. The retrieval question matches the retrieved video because it describes the cheerleaders' performance in the video. The video shows a group of cheerleaders performing an energetic and synchronized routine, including mid-air flips, runs, and formations, culminating in a stunning triangle formation held by three rows of women.

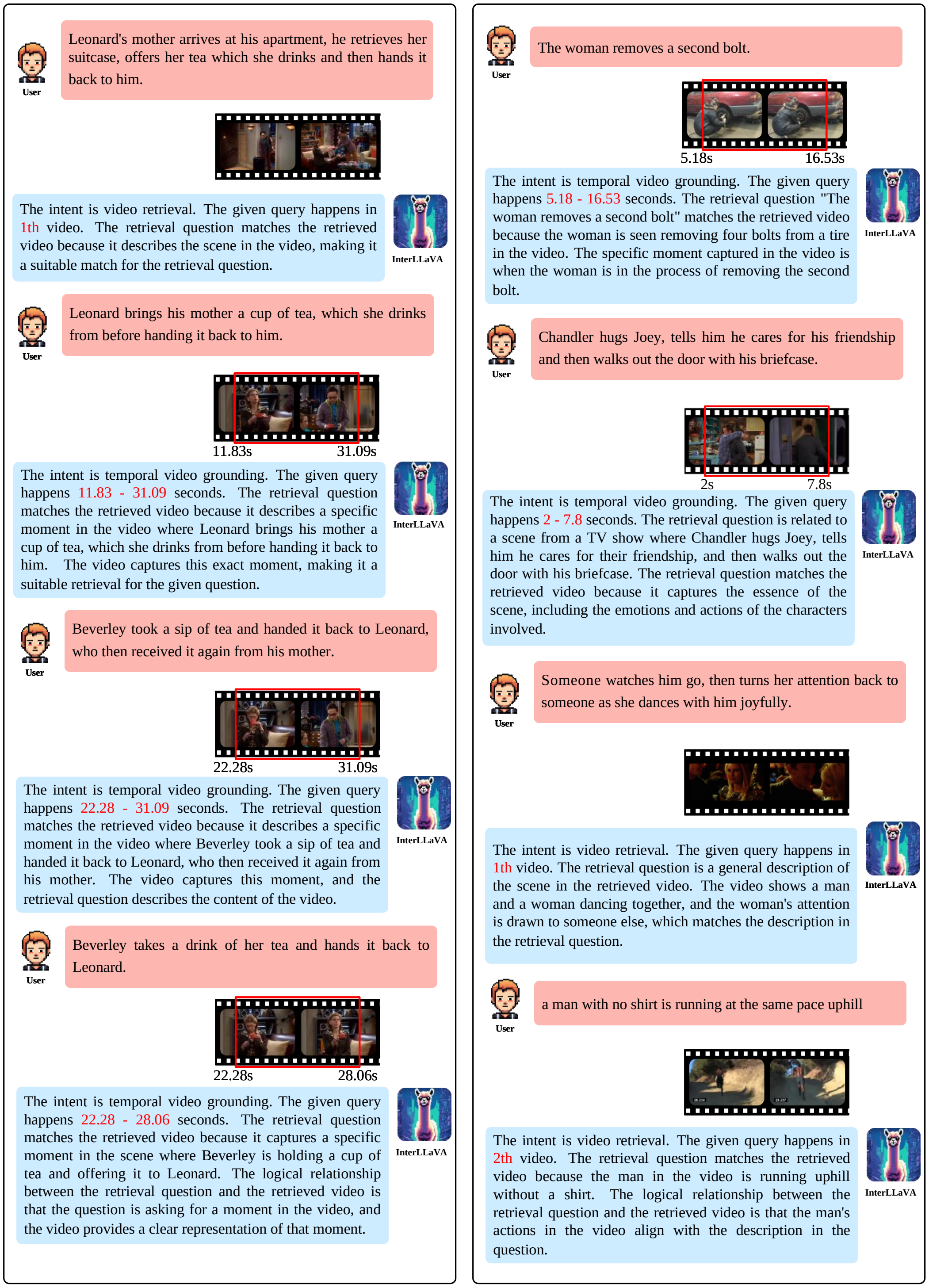
They perform an impressive cheer, staying perfectly synchronized and executing all moves gracefully, culminating in a stunning triangle formation with a girl held aloft by three rows of women.

InterLLaVA: The intent is temporal video grounding. The given query happens 96.06 - 198.07 seconds. The retrieval question matches the retrieved video because it describes the cheerleading performance in detail. The video captures the cheerleaders performing an energetic and synchronized routine, including mid-air flips, runs, and formations, culminating in a stunning triangle formation held by three rows of women.

They stay in sync throughout the entire cheer, looking great and hitting all the moves with grace.

InterLLaVA: The intent is temporal video grounding. The given query happens 96.06 - 154.49 seconds. The retrieval question matches the retrieved video because the cheerleaders in the image are performing an energetic and synchronized routine, including mid-air flips, runs, and formations, culminating in a stunning triangle formation held by three rows of women.

**Figure 19: An example of video-to-video moment retrieval from the IVCR-200K dataset.**



Leonard's mother arrives at his apartment, he retrieves her suitcase, offers her tea which she drinks and then hands it back to him.

The intent is video retrieval. The given query happens in 1th video. The retrieval question matches the retrieved video because it describes the scene in the video, making it a suitable match for the retrieval question.

Leonard brings his mother a cup of tea, which she drinks from before handing it back to him.

The intent is temporal video grounding. The given query happens 11.83 - 31.09 seconds. The retrieval question matches the retrieved video because it describes a specific moment in the video where Leonard brings his mother a cup of tea, which she drinks from before handing it back to him. The video captures this exact moment, making it a suitable retrieval for the given question.

Beverley took a sip of tea and handed it back to Leonard, who then received it again from his mother.

The intent is temporal video grounding. The given query happens 22.28 - 31.09 seconds. The retrieval question matches the retrieved video because it describes a specific moment in the video where Beverley took a sip of tea and handed it back to Leonard, who then received it again from his mother. The video captures this moment, and the retrieval question describes the content of the video.

Beverley takes a drink of her tea and hands it back to Leonard.

The intent is temporal video grounding. The given query happens 22.28 - 28.06 seconds. The retrieval question matches the retrieved video because it captures a specific moment in the scene where Beverley is holding a cup of tea and offering it to Leonard. The logical relationship between the retrieval question and the retrieved video is that the question is asking for a moment in the video, and the video provides a clear representation of that moment.

The woman removes a second bolt.

The intent is temporal video grounding. The given query happens 5.18 - 16.53 seconds. The retrieval question "The woman removes a second bolt" matches the retrieved video because the woman is seen removing four bolts from a tire in the video. The specific moment captured in the video is when the woman is in the process of removing the second bolt.

Chandler hugs Joey, tells him he cares for his friendship and then walks out the door with his briefcase.

The intent is temporal video grounding. The given query happens 2 - 7.8 seconds. The retrieval question is related to a scene from a TV show where Chandler hugs Joey, tells him he cares for their friendship, and then walks out the door with his briefcase. The retrieval question matches the retrieved video because it captures the essence of the scene, including the emotions and actions of the characters involved.

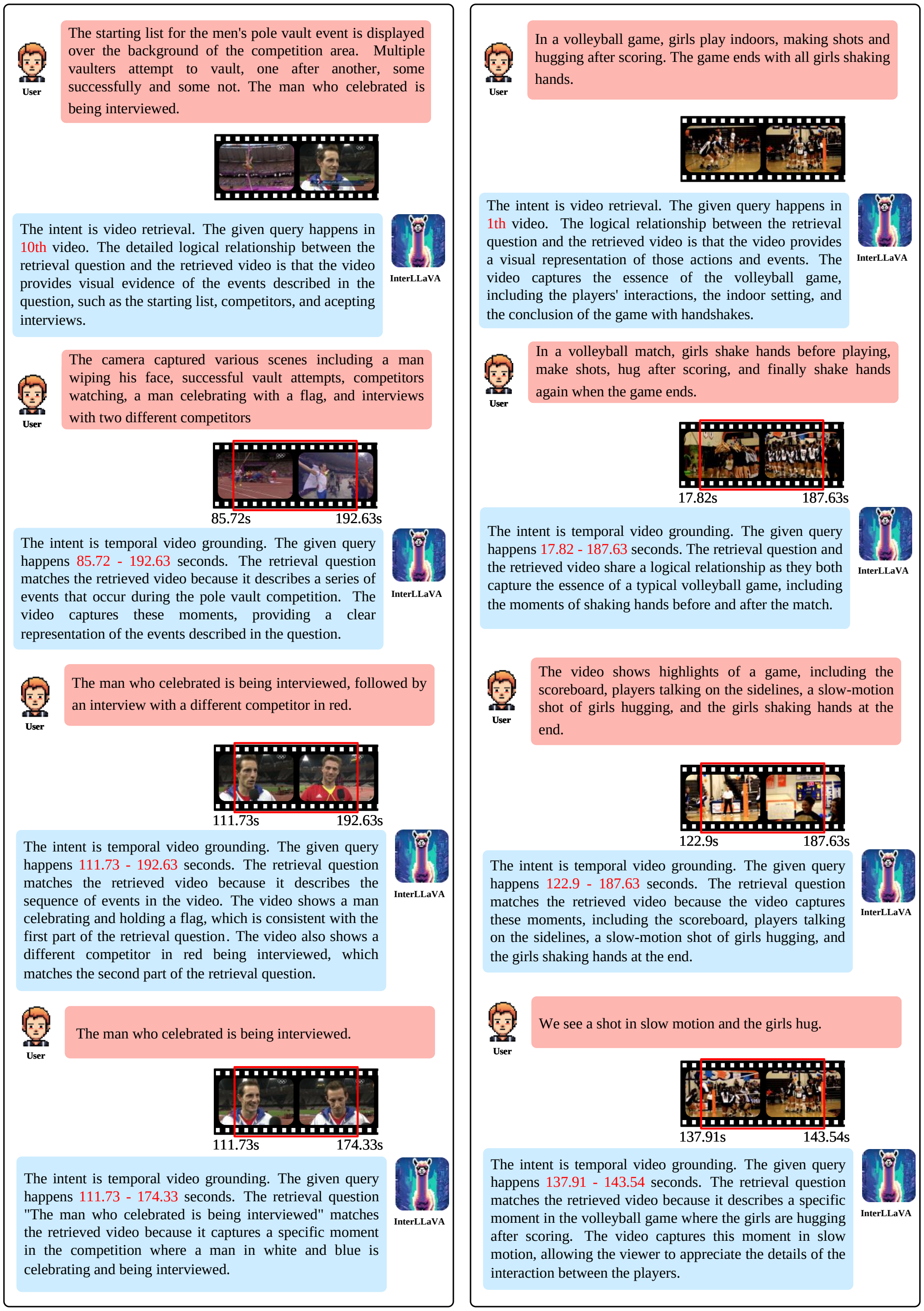
Someone watches him go, then turns her attention back to someone as she dances with him joyfully.

The intent is video retrieval. The given query happens in 1th video. The retrieval question is a general description of the scene in the retrieved video. The video shows a man and a woman dancing together, and the woman's attention is drawn to someone else, which matches the description in the retrieval question.

a man with no shirt is running at the same pace uphill

The intent is video retrieval. The given query happens in 2th video. The retrieval question matches the retrieved video because the man in the video is running uphill without a shirt. The logical relationship between the retrieval question and the retrieved video is that the man's actions in the video align with the description in the question.

**Figure 20: An example of video-to-video moment retrieval and moment-to-video retrieval from the IVCR-200K dataset.**



User: The starting list for the men's pole vault event is displayed over the background of the competition area. Multiple vaulters attempt to vault, one after another, some successfully and some not. The man who celebrated is being interviewed.

InterLLaVA: The intent is video retrieval. The given query happens in 10th video. The detailed logical relationship between the retrieval question and the retrieved video is that the video provides visual evidence of the events described in the question, such as the starting list, competitors, and accepting interviews.

User: The camera captured various scenes including a man wiping his face, successful vault attempts, competitors watching, a man celebrating with a flag, and interviews with two different competitors

InterLLaVA: The intent is temporal video grounding. The given query happens 85.72 - 192.63 seconds. The retrieval question matches the retrieved video because it describes a series of events that occur during the pole vault competition. The video captures these moments, providing a clear representation of the events described in the question.

User: The man who celebrated is being interviewed, followed by an interview with a different competitor in red.

InterLLaVA: The intent is temporal video grounding. The given query happens 111.73 - 192.63 seconds. The retrieval question matches the retrieved video because it describes the sequence of events in the video. The video shows a man celebrating and holding a flag, which is consistent with the first part of the retrieval question. The video also shows a different competitor in red being interviewed, which matches the second part of the retrieval question.

User: The man who celebrated is being interviewed.

InterLLaVA: The intent is temporal video grounding. The given query happens 111.73 - 174.33 seconds. The retrieval question "The man who celebrated is being interviewed" matches the retrieved video because it captures a specific moment in the competition where a man in white and blue is celebrating and being interviewed.

User: In a volleyball game, girls play indoors, making shots and hugging after scoring. The game ends with all girls shaking hands.

InterLLaVA: The intent is video retrieval. The given query happens in 1th video. The logical relationship between the retrieval question and the retrieved video is that the video provides a visual representation of those actions and events. The video captures the essence of the volleyball game, including the players' interactions, the indoor setting, and the conclusion of the game with handshakes.

User: In a volleyball match, girls shake hands before playing, make shots, hug after scoring, and finally shake hands again when the game ends.

InterLLaVA: The intent is temporal video grounding. The given query happens 17.82 - 187.63 seconds. The retrieval question and the retrieved video share a logical relationship as they both capture the essence of a typical volleyball game, including the moments of shaking hands before and after the match.

User: The video shows highlights of a game, including the scoreboard, players talking on the sidelines, a slow-motion shot of girls hugging, and the girls shaking hands at the end.

InterLLaVA: The intent is temporal video grounding. The given query happens 122.9 - 187.63 seconds. The retrieval question matches the retrieved video because the video captures these moments, including the scoreboard, players talking on the sidelines, a slow-motion shot of girls hugging, and the girls shaking hands at the end.

User: We see a shot in slow motion and the girls hug.

InterLLaVA: The intent is temporal video grounding. The given query happens 137.91 - 143.54 seconds. The retrieval question matches the retrieved video because it describes a specific moment in the volleyball game where the girls are hugging after scoring. The video captures this moment in slow motion, allowing the viewer to appreciate the details of the interaction between the players.

**Figure 21: An example of video-to-video moment retrieval and moment-to-video retrieval from the IVCR-200K dataset.**

**User**

Query 1#: A large snowy mountain is seen as people gather on the ice. They hit several pucks, sweeping them in a game.

Query 2#: A large snowy mountain is seen as people gather on the ice.

**Video-LLaVA**

Response 1#: The retrieved video matches the retrieval question because it shows a group of people playing a game on ice, hitting pucks, and sweeping them. The logical relationship between the retrieval question and the retrieved video is that the video provides a visual representation of the game being played, which aligns with the description provided in the retrieval question.

Response 2#: The retrieved video matches the retrieval question because it shows a group of people playing curling on an ice rink, which is a winter sport that is often played in snowy mountainous regions. The video also shows the players interacting with each other, which adds to the overall experience of the game.

**TimeChat**

Response 1#: ---

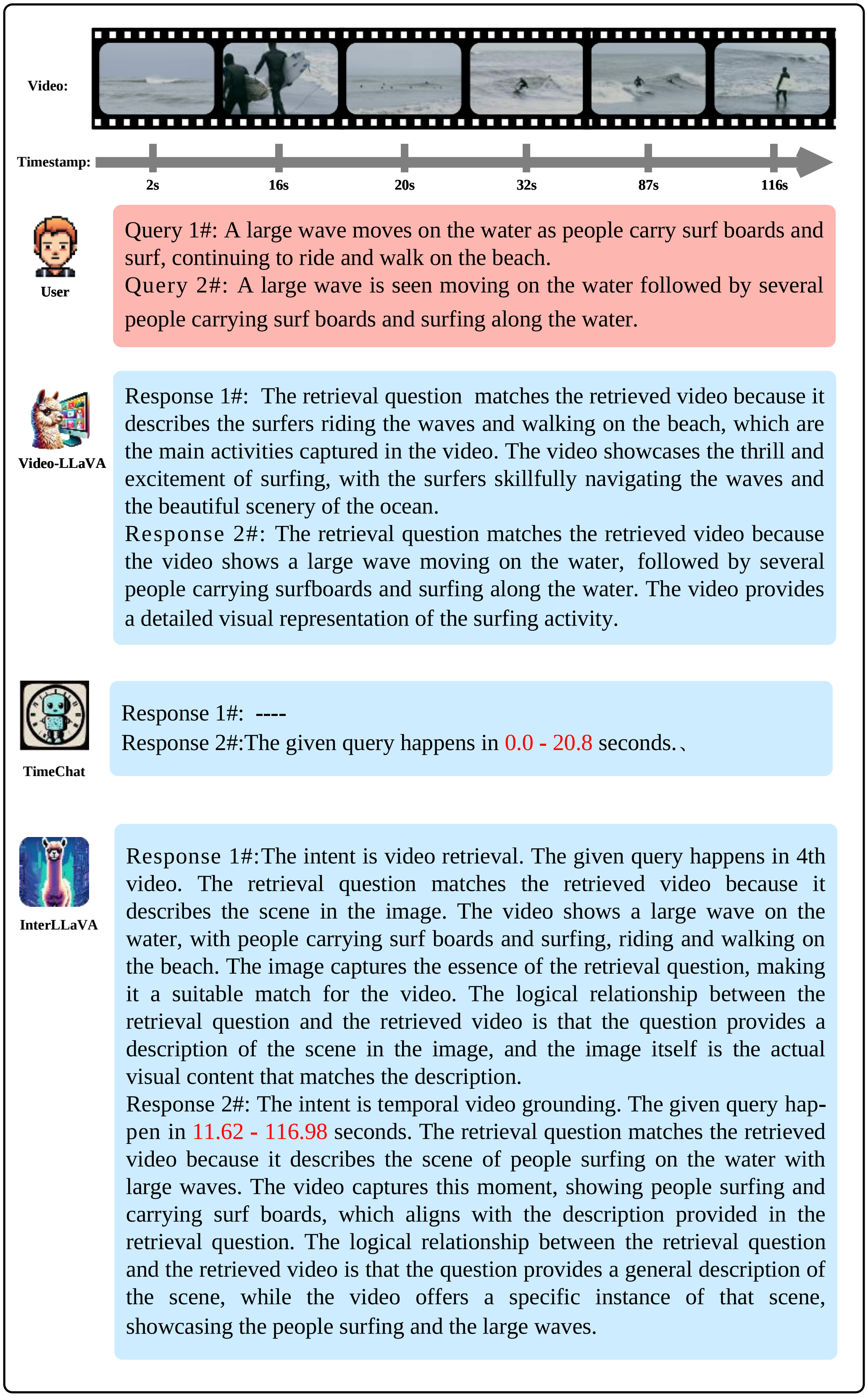
Response 2#:The given query happens in the 0.0 - 21.6 seconds.

**InterLLaVA**

Response 1#: The intent is video retrieval. The given query happens in 2th video. The retrieval question matches the retrieved video because it describes the scene in the image. The presence of a large snowy mountain in the background adds to the overall context of the scene, making it a fitting description of the video. The logical relationship between the retrieval question and the retrieved video is that the question provides a clear and concise description of the scene, while the video provides the actual visual content that matches the description.

Response 2#: "The intent is temporal video grounding. The given query happens 0.0 - 3.84 seconds. The retrieval question matches the retrieved video because it describes the scene of people engaging in a winter sport on a snowy mountain. The video shows people playing curling, which is a popular winter sport involving sliding stones on a sheet of ice. The presence of a large snowy mountain in the background further supports the idea that this is a winter sports scene. The retrieval question and the retrieved video share a common theme of winter sports and outdoor activities, making them a perfect match.

Figure 22: Comparing InterLLaVA with other video large language models.



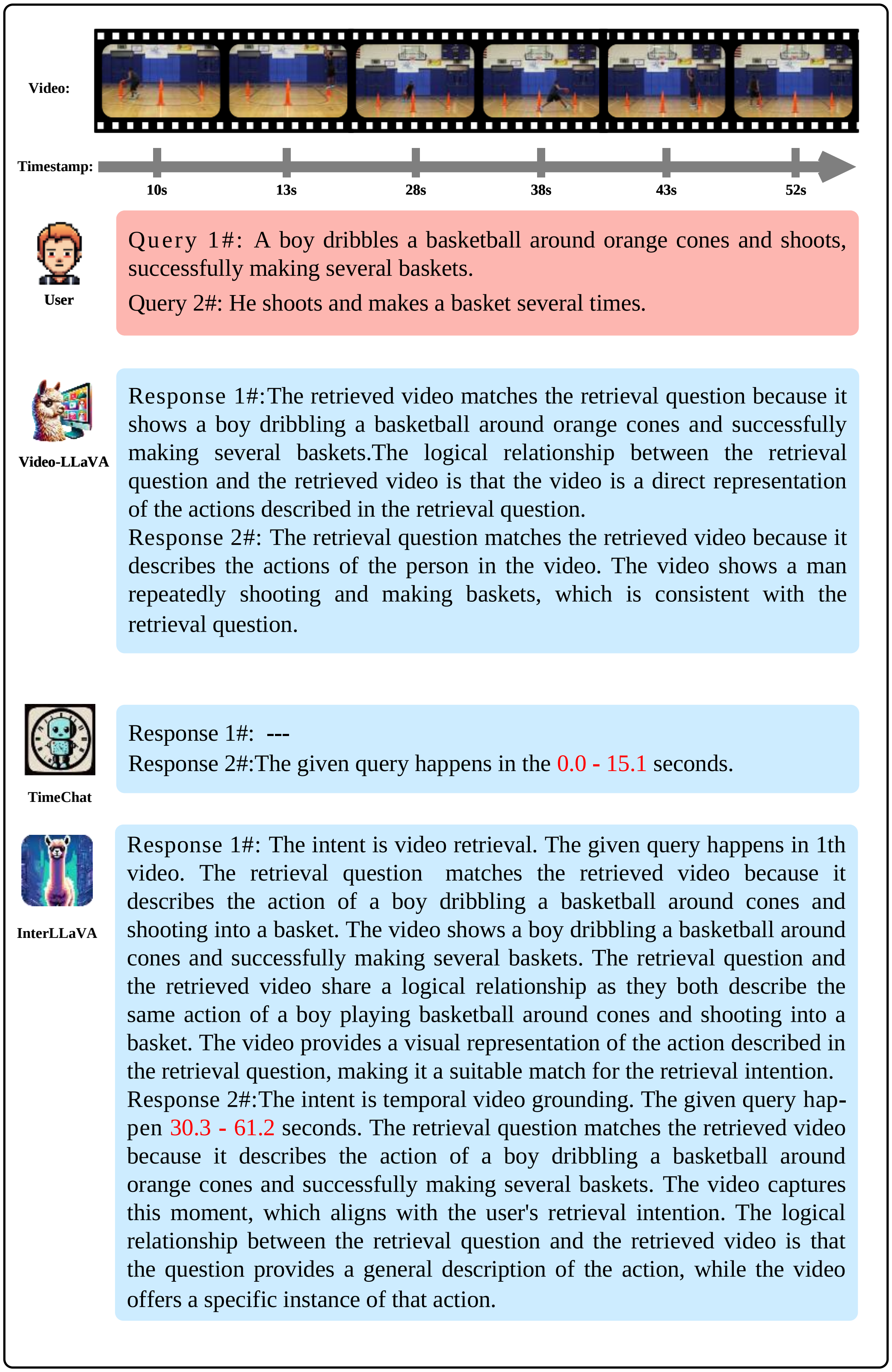
Query 1#: A large wave moves on the water as people carry surf boards and surf, continuing to ride and walk on the beach.
Query 2#: A large wave is seen moving on the water followed by several people carrying surf boards and surfing along the water.

Response 1#: The retrieval question matches the retrieved video because it describes the surfers riding the waves and walking on the beach, which are the main activities captured in the video. The video showcases the thrill and excitement of surfing, with the surfers skillfully navigating the waves and the beautiful scenery of the ocean.
Response 2#: The retrieval question matches the retrieved video because the video shows a large wave moving on the water, followed by several people carrying surfboards and surfing along the water. The video provides a detailed visual representation of the surfing activity.

Response 1#: ----
Response 2#:The given query happens in 0.0 - 20.8 seconds.、

Response 1#:The intent is video retrieval. The given query happens in 4th video. The retrieval question matches the retrieved video because it describes the scene in the image. The video shows a large wave on the water, with people carrying surf boards and surfing, riding and walking on the beach. The image captures the essence of the retrieval question, making it a suitable match for the video. The logical relationship between the retrieval question and the retrieved video is that the question provides a description of the scene in the image, and the image itself is the actual visual content that matches the description.
Response 2#: The intent is temporal video grounding. The given query happen in 11.62 - 116.98 seconds. The retrieval question matches the retrieved video because it describes the scene of people surfing on the water with large waves. The video captures this moment, showing people surfing and carrying surf boards, which aligns with the description provided in the retrieval question. The logical relationship between the retrieval question and the retrieved video is that the question provides a general description of the scene, while the video offers a specific instance of that scene, showcasing the people surfing and the large waves.

**Figure 23: Comparing InterLLaVA with other video large language models.**



**User**

Query 1#: A boy dribbles a basketball around orange cones and shoots, successfully making several baskets.

Query 2#: He shoots and makes a basket several times.

**Video-LLaVA**

Response 1#:The retrieved video matches the retrieval question because it shows a boy dribbling a basketball around orange cones and successfully making several baskets.The logical relationship between the retrieval question and the retrieved video is that the video is a direct representation of the actions described in the retrieval question.

Response 2#: The retrieval question matches the retrieved video because it describes the actions of the person in the video. The video shows a man repeatedly shooting and making baskets, which is consistent with the retrieval question.

**TimeChat**

Response 1#: ---

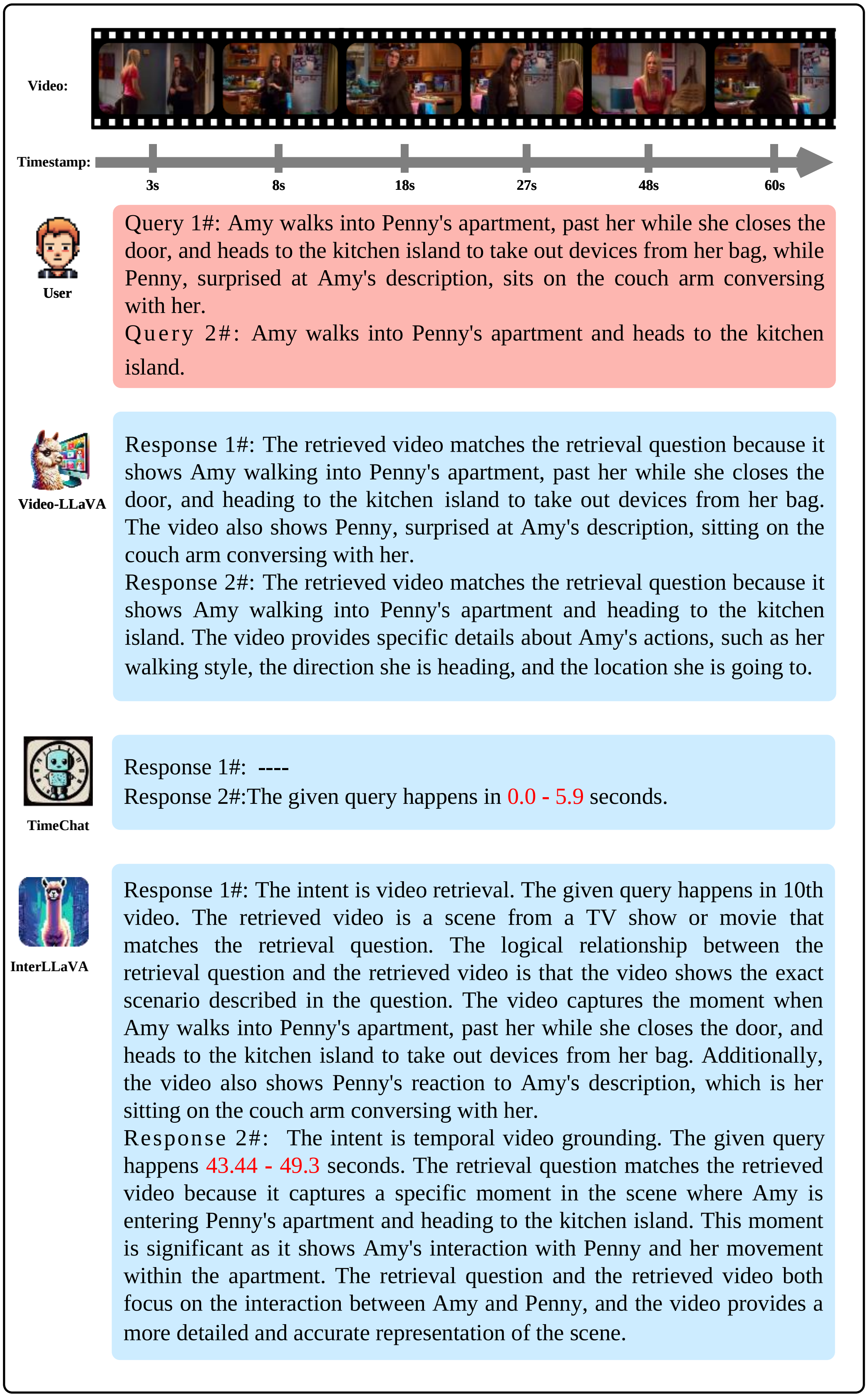
Response 2#:The given query happens in the 0.0 - 15.1 seconds.

**InterLLaVA**

Response 1#: The intent is video retrieval. The given query happens in 1th video. The retrieval question matches the retrieved video because it describes the action of a boy dribbling a basketball around cones and shooting into a basket. The video shows a boy dribbling a basketball around cones and successfully making several baskets. The retrieval question and the retrieved video share a logical relationship as they both describe the same action of a boy playing basketball around cones and shooting into a basket. The video provides a visual representation of the action described in the retrieval question, making it a suitable match for the retrieval intention.

Response 2#:The intent is temporal video grounding. The given query happen 30.3 - 61.2 seconds. The retrieval question matches the retrieved video because it describes the action of a boy dribbling a basketball around orange cones and successfully making several baskets. The video captures this moment, which aligns with the user's retrieval intention. The logical relationship between the retrieval question and the retrieved video is that the question provides a general description of the action, while the video offers a specific instance of that action.

**Figure 24: Comparing InterLLaVA with other video large language models.**



Query 1#: Amy walks into Penny's apartment, past her while she closes the door, and heads to the kitchen island to take out devices from her bag, while Penny, surprised at Amy's description, sits on the couch arm conversing with her.
Query 2#: Amy walks into Penny's apartment and heads to the kitchen island.

Response 1#: The retrieved video matches the retrieval question because it shows Amy walking into Penny's apartment, past her while she closes the door, and heading to the kitchen island to take out devices from her bag. The video also shows Penny, surprised at Amy's description, sitting on the couch arm conversing with her.
Response 2#: The retrieved video matches the retrieval question because it shows Amy walking into Penny's apartment and heading to the kitchen island. The video provides specific details about Amy's actions, such as her walking style, the direction she is heading, and the location she is going to.

Response 1#: ----
Response 2#:The given query happens in 0.0 - 5.9 seconds.

Response 1#: The intent is video retrieval. The given query happens in 10th video. The retrieved video is a scene from a TV show or movie that matches the retrieval question. The logical relationship between the retrieval question and the retrieved video is that the video shows the exact scenario described in the question. The video captures the moment when Amy walks into Penny's apartment, past her while she closes the door, and heads to the kitchen island to take out devices from her bag. Additionally, the video also shows Penny's reaction to Amy's description, which is her sitting on the couch arm conversing with her.
Response 2#: The intent is temporal video grounding. The given query happens 43.44 - 49.3 seconds. The retrieval question matches the retrieved video because it captures a specific moment in the scene where Amy is entering Penny's apartment and heading to the kitchen island. This moment is significant as it shows Amy's interaction with Penny and her movement within the apartment. The retrieval question and the retrieved video both focus on the interaction between Amy and Penny, and the video provides a more detailed and accurate representation of the scene.

Figure 25: Comparing InterLLaVA with other video large language models.