

VideoChat: Chat-Centric Video Understanding

KunChang Li^{*1,4}, Yanan He^{*1}, Yi Wang^{*†1}, Yizhuo Li^{1,3}, Wenhai Wang¹
Ping Luo³, Yali Wang^{4,1}, Limin Wang^{2,1}, Yu Qiao

Presented by Justin, Ziyang and Feihong



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

Introduction

- Video-centric multimodal instruction fine-tuning dataset
- Video-centric multimodal dialogue system



Spatialtemporal reasoning:



Is an accident in the video?



Yes, there is an accident in the video. You can see a silver car on the road that appears to have been involved in an accident.

What to do if someone is in the car?



In case someone is in the car, you should immediately call for emergency services and wait for the police to arrive. You should not try to move the car or remove the person from the car unless they are in danger.



Event Localization:



What is she doing in the video?

She is doing yoga on the roof in the video.

What is she wearing?

She is wearing a black shirt and pants, and a black swimsuit.

Where is she doing yoga?

She is doing yoga on a roof.

Why do you think it's on a roof?

It appears to be on a roof because of the clear sky and the large green tree in the background.

Is she safe when doing yoga? Will she fall down?

Yes, she appears to be safe as she is on a flat surface and there is no one else around. It is unlikely that she will fall down as she is on a flat surface and appears to be taking the necessary safety precautions.

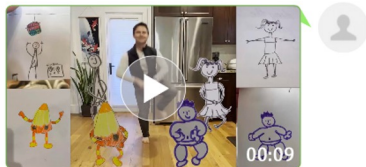
When was this video shot? Was it taken at night?

This video appears to have been shot during the day, as the sky is clear and there is a large green tree in the background.

What is her purpose in doing this?

Her purpose in doing this appears to be to practice yoga and to enjoy the view from the rooftop.



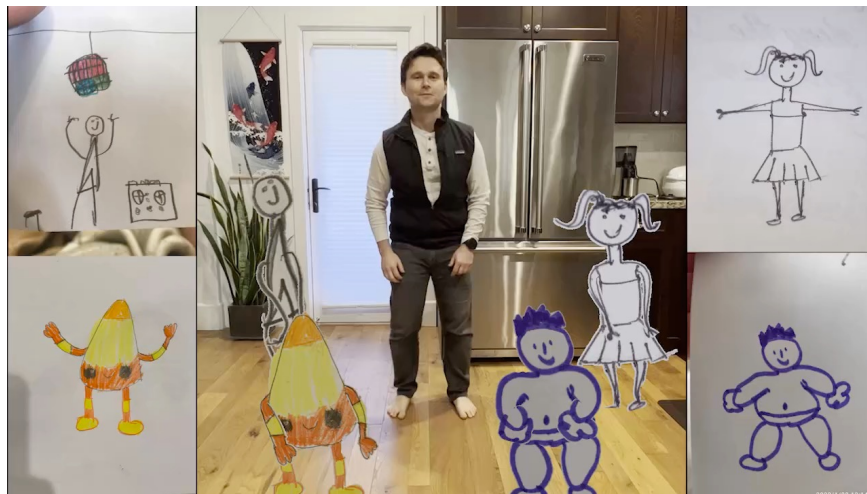


Describe the video.

The video shows a young man **dancing in a kitchen**, surrounded by **colorful cartoon characters**. He wears a white shirt and blue pants and is dancing in front of the **refrigerator**. The video also shows the kitchen's interior with a wooden floor and appliances. In some frames, the man is dancing in front of the cereal box, cookies, and a can of peas. There are also other items, like a juice bottle, in the kitchen. The man dances **joyfully**, with a big smile on his face, and enjoys the music. Overall, the video captures the energy and enthusiasm of the dancer and creates a **fun and lively atmosphere**.

Explain why this video is funny.

The video is funny because the man's dancing is random and **spontaneous**, and he adds an element of humor to the ordinary kitchen setting. His dancing is unpredictable and silly, and it adds an element of surprise to the video. The man's joyful facial expressions and **childlike gestures** make the video amusing and entertaining to watch. Additionally, the juxtaposition of the man's dancing with the serious and mundane objects in the kitchen creates an interesting contrast and adds to the humor. The fact that the man is dancing in front of food items in a kitchen is also a nod to our daily lives and the fact that we often find ourselves in the kitchen dancing and singing along to music, making the video relatable and funny to watch.



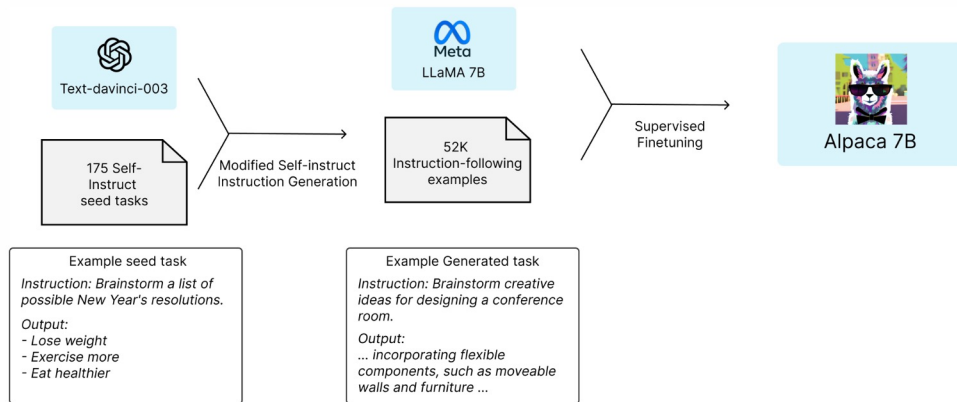
Casual relationship inference



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

Related Work

- Large Language Model (instruction-tuning)



- Vision Foundation Model



Related Work

LLM for Multimodal Understanding

Llava(image)

Context type 1: Captions

A group of people standing outside of a black vehicle with various luggage. Luggage surrounds a vehicle in an underground parking area. People try to fit all of their luggage in an SUV.

The sport utility vehicle is parked in the public garage, being packed for a trip. Some people with luggage near a van that is transporting it.

Context type 2: Boxes

person: [0.681, 0.242, 0.774, 0.694], person: [0.63, 0.222, 0.686, 0.516], person: [0.444, 0.233, 0.487, 0.34], backpack: [0.384, 0.696, 0.485, 0.914], backpack: [0.755, 0.413, 0.846, 0.692], suitcase: [0.758, 0.413, 0.845, 0.69], suitcase: [0.1, 0.497, 0.173, 0.579], bicycle: [0.282, 0.363, 0.327, 0.442], car: [0.786, 0.25, 0.848, 0.322], car: [0.783, 0.27, 0.827, 0.335], car: [0.86, 0.254, 0.891, 0.3], car: [0.261, 0.101, 0.787, 0.626]

Response type 1: conversation

Question: What type of vehicle is featured in the image?

Answer: The image features a black sport utility vehicle (SUV).

Question: Where is the vehicle parked?

Answer: The vehicle is parked in an underground parking area, likely in a public garage.

Question: What are the people in the image doing?

Answer: The people are trying to fit all of their luggage into the SUV, likely preparing for a trip.

Response type 2: detailed description

The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip.

In the parking area, various luggage items are scattered around the vehicle. There are two backpacks, one located near the left rear wheel and the other closer to the right side of the vehicle. Additionally, there are two suitcases, one on the right side of the car and another further away near the center of the parking area. A bicycle can also be seen on the left side of the vehicle.

Other cars are parked around the main SUV, with one car positioned behind it and slightly to the left, another behind and slightly to the right, and the third car further behind on the right side.

Response type 3: complex reasoning

Question: What challenges do these people face?

Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings to accommodate. They might have to strategize and arrange the luggage efficiently to ensure that everything fits properly into the vehicle. Additionally, they need to consider the comfort of the passengers and visibility while driving, so the placement of the luggage must not obstruct the driver's view or make the passengers uncomfortable during the trip.



What are the meals that I can cook with these?

With the variety of food items stored in the refrigerator, you can prepare several meals. For example, you can create a fruit salad using the strawberries, blueberries, and carrots. Additionally, you can make a delicious smoothie using the fruits and vegetables, which would be a healthy and nutritious snack or breakfast option. The milk and eggs can be used to create various dishes, such as a milk-based smoothie, a baked casserole, or scrambled eggs. The juice can also be used to make a fresh fruit and juice mix or as a refreshing homemade popsicle. The possibilities are vast, and the meals you can create will depend on your preferences and dietary needs.

I would like to do the fruit salad. Show me the recipe.

Certainly! Here's a simple fruit salad recipe using the items in your refrigerator:

Ingredients:

- 2 cups fresh strawberries, sliced
- 1 cup fresh blueberries
- 1 large carrot, grated
- Juice of half a lemon
- A handful of chopped fresh parsley or mint
- A few ice cubes
- Optional: honey or maple syrup (adjust to taste)

Instructions:

1. In a large bowl, combine the sliced strawberries, blueberries, and carrot. Gently toss them together to mix.
2. Drizzle the juice of half a lemon over the fruit and vegetables to prevent them from oxidizing and add a refreshing citrus flavor.
3. Sprinkle the chopped fresh parsley or mint over the fruit and gently toss once more.
4. If desired, add a drizzle of honey or maple syrup for a touch of sweetness. Adjust the amount according to your taste preferences.
5. Cover the bowl and refrigerate the fruit salad for at least 30 minutes or up to a few hours. This allows the flavors to meld and creates a more refreshing salad.
6. Serve the fruit salad in cups or bowls, garnished with a sprinkle of fresh parsley or mint, if desired.

This fruit salad recipe is perfect for a healthy snack or side dish, and with the variety of ingredients available in your refrigerator, it should be easy to put together quickly.



Framework

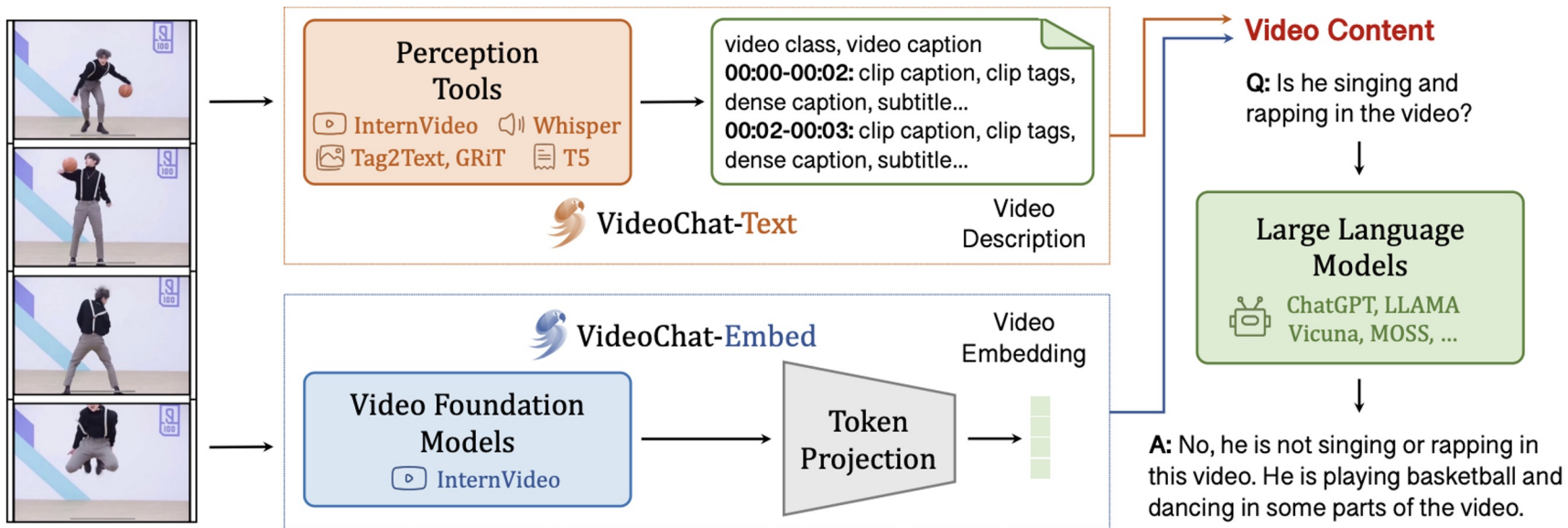


Figure 1: **The framework of VideoChat.** VideoChat-Text textualizes videos in stream. VideoChat-Embed encodes videos as embeddings. Both video content will be input in LLMs for multimodal understanding.

Framework

- VideoChat unifies video-related tasks into the formulation of multiple-round video question answering.
- Tasks are defined by words in a live inference and no or a few instances are given for learning
- Treat an LLM as a universal video task decoder, turning video-related descriptions or embeddings into human-understandable text.

VideoChat-Text

- Convert visual data from videos into textual format
- Perception Models
 - Video Analysis: actions, objects, object annotations with positions and more
 - Pretrained T5 language model to refine their descriptions for improved clarity
 - Integrate the Whisper speech recognition model to capitalize on audio data within videos
- Prompt System

You are a chatbot that conducts conversations based on video contexts. You mainly answer based on the given contexts, and you can also modify the content according to the tag information, and you can also answer the relevant knowledge of the person or object contained in the video. **The timing description is a description every $1/FPS$ second, so that you can convert it into time. When describing, please mainly refer to the timing description. Dense caption is to give content every five seconds, you can disambiguate them in timing.** But you don't create a video plot out of nothing.

Begin!

Video contexts in temporal order: `textualizing_videos`

Question: `question`

Video description with perception models

Video Class, Video Caption

00:00-00:02 Clip Caption, Clip Tag, Dense Caption, Video Subtitle...

00:02-00:03 Clip Caption, Clip Tag, Dense Caption, Video Subtitle...

00:03-00:06 Clip Caption, Clip Tag, Dense Caption, Video Subtitle...

...



answering questions, a man and a woman sitting on a couch in a living room with a table in front of them.

00:00-00:11 a man and a girl sitting on a couch in a living room.

a lamp with a white shade a woman sitting at a table: [446, 155, 710, 476]; man wearing a plaid shirt: [361, 44, 581, 337]; man sitting on couch: [10, 63, 324, 350]; the tie is grey: [441, 150, 486, 280]; a glass of beer: [38, 305, 77, 367]; a stack of magazines: [28, 350, 180, 394]; a white tablecloth: [0, 334, 626, 476]; stainless steel oven: [1, 55, 150, 142]; a brown tie on a man: [144, 168, 191, 270]; the couch is white: [0, 119, 730, 472]; a gray binder: [0, 377, 157, 411]; a white couch: [768, 350, 848, 477]; a lamp with a white shade: [582, 26, 713, 195];

00:00-00:02: Hey, Pheeb, you gonna have the rest of that Pop-Tart?

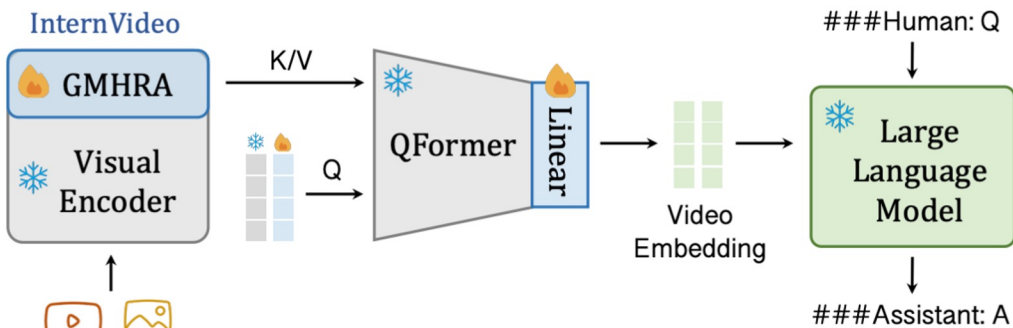
00:02-00:03: Pheeb?

00:03-00:09: Does anyone want the rest of this Pop-Tart?

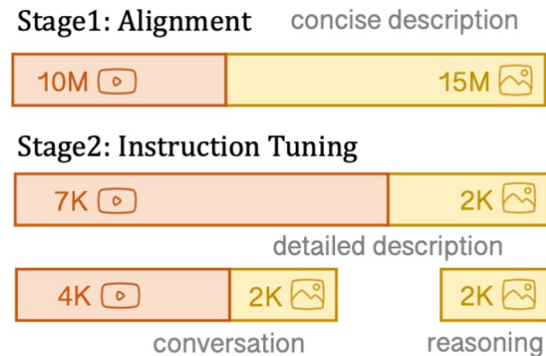
00:09-00:11: Hey, I might.

VideoChat-Embed

- Learnable Video-Language Token Interface (VLTF)
- Using cross-attention to compress the video tokens
- Tuned with video-text data



(a) Architecture



(b) Data

Figure 2: **Architecture and training paradigm of VideoChat-Embed.** The architecture is built on BLIP-2 [18] and StableVicuna [10]. The training contains two-stage alignment and instruction tuning.

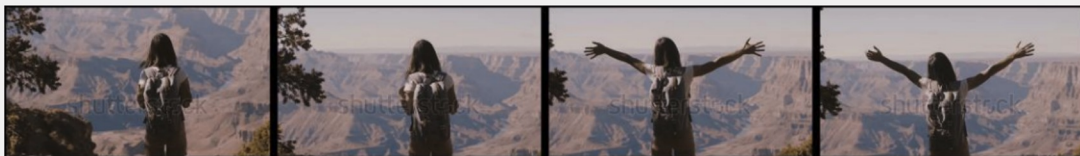
VideoChat-Embed Training

- Freeze all parameter except GMHRA, queries and linear projection.
- Two stage training
 - Stage1: align the video encoder with LLM via large-scale video-text fine-tuning
 - Stage2: tune the system with two types of video instruction data.
- Instruction Data
 - Detailed Video Descriptions: condense the provided video description into a video narrative employing GPT-4 (highlights the temporal aspects of the video by illustrating its progression over time)
 - Video Conversations: given video description, generate multi-round dialogues with three types of prompts concerning descriptive, temporal, and causal content for videos with ChatGPT

Prompts and Examples for detailed video descriptions.

Give you a video of [origin_caption](#). The content of the video in temporal order is: [textualizing_videos](#). Please use the sequence adverbs "first", "next", "then" and "finally" to describe this video in detail, but don't mention the specific time. Give as many details as possible. Say everything you see. The description should be more than 150 words and less than 200 words.

Fix the error in the given paragraph. Remove any repeating sentences, meaningless characters, not English sentences, and so on. Remove unnecessary repetition. Rewrite any incomplete sentences. Return directly the results without explanation. Return directly the input paragraph if it is already correct without explanation.



###Human: <Video>[video_embed](#)</Video> The video contains **T** frames sampled at t_0, t_1, \dots, t_T seconds.

###Human: Portray the video with a rich, descriptive narrative, capturing its atmosphere and events.

###Assistant: **First**, we see a young woman with long brown hair wearing a backpack standing on the edge of a cliff, looking at the beautiful mountain scenery. She is in awe and stands there for a few seconds. **Next**, we see her turning around slowly, with her arms open wide. **Then**, we see her taking a few steps back, still facing the canyon, with her arms still open wide. She seems to be soaking in the grandeur of the view. **Finally**, she stands confidently at the edge of the cliff, her backpack still on her back, as if she has conquered the mountain. Throughout the video, we see the vast expanse of the canyon with the mountains in the background. The sky is overcast in some parts and clear blue in others, creating a beautiful contrast. **Overall**, the young woman seems to be thrilled with the view and enjoys every bit of it with open arms.

Prompts for Video Conversations

As an AI visual assistant, you are observing a single video. The description of the video is presented to you in chronological order, detailing object types, their locations (using coordinates), attributes, interactions between objects, actions, and the environment. Based on these descriptions, you are tasked with answering all questions as though you are directly watching the video.

Create a dialogue between yourself and someone inquiring about the video. Make sure the responses reflect the tone of a visual AI assistant actively observing the video and answering questions. Include diverse queries and corresponding answers.

Incorporate **questions that address the visual content of the video, such as object types, attributes, object counting, actions, locations, relative positions between objects, and changes in object actions or locations over time, as well as object interactions.** Only include questions with definitive answers:

- Questions whose contents can be confidently observed and answered based on the video.
- Questions whose absence from the video can be confidently determined.

Next, encompass **questions related to temporal perception and reasoning, such as inquiring about what a person did before or after an event, or asking for specific timestamps of certain events or actions**

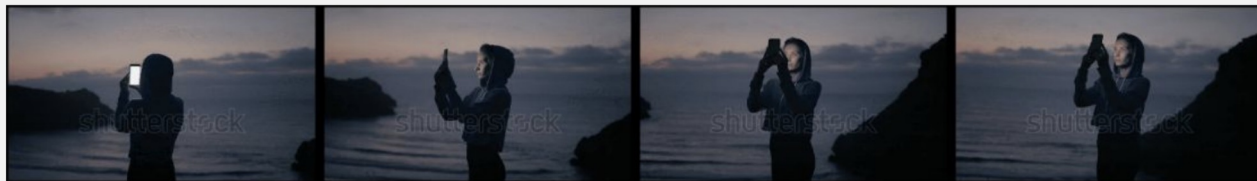
Also include **complex questions relevant to the video's content, like those asking about the background knowledge of objects or actions in the video, discussing events occurring in the video, delving into counterfactual topics (e.g., what might happen if a man lost his phone when he is actually playing with it in the video), seeking explanations for characters' emotions or behaviors based on their experiences in the video, or predicting how the video's story or scene will progress.**

Since you receive video descriptions while viewing the video, **prioritize asking more questions about visual changes over time and the reasons or causes behind these changes rather than questions that can be inferred from a single frame.**

Remember not to inquire about uncertain details. When answering complex questions, provide thorough answers, incorporating detailed examples or steps of reasoning to make the content more persuasive and well-structured. Use multiple paragraphs if necessary. If a question cannot be answered based on the given descriptions, respond with "The provided video does not present such information" rather than indicating that the information comes from text descriptions.

Table 6: **Prompt for video conversations.** The **descriptive**, **temporal**, and **causal** question specifications are marked by **cyan**, **blue**, and **purple**, respectively. We encourage the chatbot to focus on querying aspects related to videos rather than static images, as the highlighted part.

Examples for Video Conversations



###Human: <Video>[video_embed](#)</Video> The video contains T frames sampled at t_0, t_1, \dots, t_T seconds.

###Human: What is happening in the video?

###Assistant: In the video, we see a silhouette of a young woman holding a smartphone and taking pictures of the sunset over the sea on the beach at night. Then, there are several other silhouettes of people taking pictures and looking at their phones.

###Human: Can you describe the **interactions** between the objects in the video?

###Assistant: The silhouettes of people in the video are holding smartphones, and she is taking pictures of the sunset over the sea on the beach at night.

###Human: What is the **location** of the scene?

###Assistant: The scene is located on a beach at night near the sea.

###Human: Can you tell me about any changes that happen in the video **over time**?

###Assistant: The color of the sky changes from orange to grey as the video progresses. Additionally, the actions of the silhouettes of people change as they take pictures and look at their phones on the beach.

Two-stage joint training

- Image+Video instruction data
- Stage1: Alignment
 - Language models receive corresponding visual descriptions as answers.
 - “###Human: <Video>video_embed</Video> video_instruction ###Assistant:”
 - “###Human: <Image>image_embed</Image> image_instruction ###Assistant:”
- Stage2: Instruction tuning
 - Self-built video instruction data consists of 7K detailed video descriptions and 4K video conversations
 - Include temporal reasoning sampling information for video data: “The video contains T frames sampled at t0, t1, ..., tT seconds.”

Qualitative Analysis - Spatial Perception



What kind of music do you think suits this video?



I believe that this video is most suitable for anime or Japanese-themed music, particularly k-pop or anime openings. The bright and colorful costumes and dancers remind me of a cheerful, upbeat atmosphere. The music could complement that mood by being light and fun.

How many person in the video?



There are 3 people in the video.



Qualitative Analysis - Spatial Perception



Explain why this meme is funny.



This meme is funny because it shows a small dog laying on the floor with a person's feet in the background. The dog looks very relaxed and content, and the caption reads "just Monday". This implies that the dog is feeling the same way as many people do on Mondays, which is a common feeling of exhaustion and lack of motivation. The meme is relatable and humorous because it plays on the idea that Mondays are often seen as a day that **people do not look forward to**, and this dog is no exception. It is also funny because it shows **a small dog in a very human - like position**, which adds to the humor of the meme.



What is unusual about this image?



This is a very unusual image because the man is **standing on the back of a parked car** while wearing a yellow shirt. He seems to be **ironing clothes on the back of the car**, which is parked on the street in front of a tall building. The yellow SUV has its rear door open, possibly allowing the man to reach the clothes he is ironing. The car is parked on a busy street with other cars and a yellow taxi driving by, which adds to the overall scene's busyness and hustle and bustle. The man's unique job, ironing clothes on the back of a car on a busy street, seems to be part of the unique experience of the city's daily life.



Qualitative Analysis - Temporal Perception and Reasoning



How long is this video?



This video lasts 16 seconds.

Is he singing and rapping in the video?



No, he is not singing or rapping in this video. He is playing basketball and dancing in some parts of the video.

In which second, he starts to dance?



The man starts to dance in the second 9.8. He dances while wearing a brown sweater and black pants, and is wearing a black and white striped shirt.

Where did he take this video?



The video was filmed in an indoor gym, on a basketball court. You can see the basketball court in the background of some shots, as well as some large white letters with words painted on the wall. There are also white walls and a grey floor, which suggest that the location is likely an indoor gym.

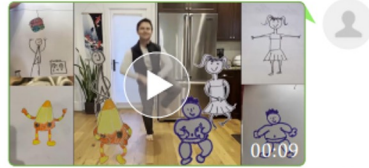
Please describe whether the camera lens moving.



No, the camera lens does not appear to be moving in this video. The camera remains stationary and captures the actions of the man in the gym. In some shots, the camera captures the basketball bouncing on the floor, but this is due to the action of the man in the gym rather than any camera movement. Overall, the camera lens appears to remain fixed throughout the duration of the video.



Qualitative Analysis - Casual Inference



Describe the video.

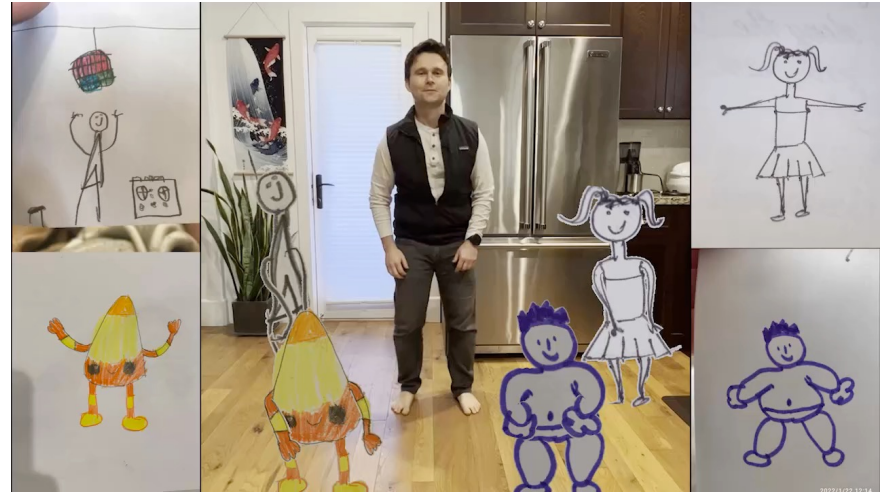


The video shows a young man dancing in a kitchen, surrounded by colorful cartoon characters. He wears a white shirt and blue pants and is dancing in front of the refrigerator. The video also shows the kitchen's interior with a wooden floor and appliances. In some frames, the man is dancing in front of the cereal box, cookies, and a can of peas. There are also other items, like a juice bottle, in the kitchen. The man dances joyfully, with a big smile on his face, and enjoys the music. Overall, the video captures the energy and enthusiasm of the dancer and creates a fun and lively atmosphere.

Explain why this video is funny.



The video is funny because the man's dancing is random and spontaneous, and he adds an element of humor to the ordinary kitchen setting. His dancing is unpredictable and silly, and it adds an element of surprise to the video. The man's joyful facial expressions and childlike gestures make the video amusing and entertaining to watch. Additionally, the juxtaposition of the man's dancing with the serious and mundane objects in the kitchen creates an interesting contrast and adds to the humor. The fact that the man is dancing in front of food items in a kitchen is also a nod to our daily lives and the fact that we often find ourselves in the kitchen dancing and singing along to music, making the video relatable and funny to watch.



Qualitative Analysis - Temporal Perception and Reasoning



What is she doing in the video?

She is doing yoga on the roof in the video.

What is she wearing?

She is wearing a black shirt and pants, and a black swimsuit.

Where is she doing yoga?

She is doing yoga on a roof.

Why do you think it's on a roof?

It appears to be on a roof because of the clear sky and the large green tree in the background.

Is she safe when doing yoga? Will she fall down?

Yes, she appears to be safe as she is on a flat surface and there is no one else around. It is unlikely that she will fall down as she is on a flat surface and appears to be taking the necessary safety precautions.

When was this video shot? Was it taken at night?

This video appears to have been shot during the day, as the sky is clear and there is a large green tree in the background.

What is her purpose in doing this?

Her purpose in doing this appears to be to practice yoga and to enjoy the view from the rooftop.




Comparison to Other Multimodal Dialogue Systems


Language Assistants(3)

00:01:09,836 --> 00:01:11,201
(EXHALES)

00:01:15,442 --> 00:01:18,843
(Eve:)Do you think the guy who did this to me feels bad?


What was House sit on when Eve asked if the guy felt bad? According to the subtitle of a video, and choose the best answer matches the question.
A: The floor in an office.
B: A desk in an office.
C: A couch in a living room.
D: A park bench.
E: The foot of the bed.
Only choose one answer.

 **C: A couch in a living room.**


 00:01:15,442--> 00:01:18,843
(Eve:)Do you think the guy who did this to me feels bad?


Answer: **B: A desk in an office.**


Image Assistants(5)




Choose the best answer matches the question: What was House sit on when Eve asked if the guy felt bad?
A: The floor in an office.
B: A desk in an office.
C: A couch in a living room.
D: A park bench.
E: The foot of the bed.
Only choose one answer.


 The best answer is **A: The floor in an office.**

 The correct answer is **B: A desk in an office.**


 **A couch in a living room.**


 **D: A park bench.**

Video Assistants(3)





Choose the best answer matches the question: What was House sit on when Eve asked if the guy felt bad?
A: The floor in an office.
B: A desk in an office.
C: A couch in a living room.
D: A park bench.
E: The foot of the bed.
Only choose one answer.

 **D: A park bench.**

 **D: A park bench.** In the video, House and Eve are sitting on a park bench in a park when Eve asks if the guy feels bad. There is a lake and a road nearby in the background.

Conclusion

- Pioneering investigation on multimodal dialogue system.
- Two settings:  VideoChat-Text &  VideoChat-Embed
- Propose a video-centric instructional dataset.
- (7K detailed video descriptions and 4K video conversations)
- Only qualitative evaluations.