# Beyond Raw Videos: Understanding Edited Videos with Large Multimodal Model Supplementary Material

Lu Xu<sup>†</sup> Sijie Zhu<sup>†\*</sup> Chunyuan Li Chia-Wen Kuo Fan Chen Xinyao Wang Guang Chen Dawei Du Ye Yuan Longyin Wen Bytedance Intelligent Creation, San Jose, USA

# Abstract

In the supplementary material, we provide the following content for a better understanding of the paper:

- A. Detailed QA Generation Examples.
- B. Evaluation of Concurrent Video LMMs.
- C. Sorry-style Answers Distribution.
- D. More Visualization Results.

# A. Detailed QA Generation Examples

Fig. 1, 2 show the QA generation prompts for reasoning and temporal categories of Panda-WebVid30K. Fig. 3, 4, 5, 6 show the example QAs and prompts for QA generation of TikTok-CapCut2K. We first generate 5 short QAs using GPT-4V, then each question is feed to GPT4-V to get a more detailed answer. During the annotation, the annotators select all the correct QAs from the GPT-4V outputs. If all the 5 questions are annotated as incorrect, the annotators will select one template question from a pre-defined list (examples in Table 1) and provide an answer for this question.

# **B. Evaluation of Concurrent Video LMMs.**

There are several very recent works on arxiv [6, 7], and Github release [2], which are considered as concurrent works in this paper. We benchmark their performance on the proposed EditVid-QA evaluation set in Table 2 and the academic datasets in Table 3. The detailed evaluation prompt of the EditVid-QA benchmark is provided in Fig. 7. We find that LLaVA-Hound-DPO has a significant performance improvement over other methods, indicating the importance of preference-based LLM tuning data. The other LMMs extended from image model, e.g., LLaVA-NeXT-Video-7B [1] and PLLaVA [6], perform on par with video models like VideoChat2-Mistral [2].

### C. Sorry-style Answers Distribution.

Table 4 shows the detailed list of all the sorry-style answers and we use the first one for "sorry-attack" in tables. The percentage of sorry-style answers in GPT-4V outputs is provided in Table 5. The percentage for consistency is the highest (20.6%) among all the categories.

#### **D.** More Visualization Results.

In Fig. 8, 9, we provide four cases on EditVid-QA benchmark to illustrate the prediction of LLaMA-VID [3] trained with our data, denoted as "Ours". We add the results of GPT-4V for reference.

# References

- [1] https://llava-vl.github.io/blog/2024-04-30-llava-next-video/. 1, 7
- [2] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. arXiv preprint arXiv:2311.17005, 2023. 1, 7
- [3] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. *arXiv* preprint arXiv:2311.17043, 2023. 1
- [4] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv preprint arXiv:2306.05424, 2023. 7
- [5] OpenAI. Gpt-4v(ision) system card. https://cdn.openai.com/papers/gptv\_system\_card.pdf. 2023. 7
- [6] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. arXiv preprint arXiv:2404.16994, 2024. 1, 7
- [7] Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, et al. Direct preference optimization of video large multimodal models from language model reward. arXiv preprint arXiv:2404.01258, 2024. 1, 7

<sup>†</sup> Equal contribution

<sup>\*</sup> Corresponding author, sijiezhu@bytedance.com

f 'You are an AI visual assistant that can analyze video, which is provided as frames extracted from video. Apart from the frames, some short captions are also provided as follows: \"{captions}\"

The task is to understand the video, create 5 questions about the video, and provide the answer in detail. \n The questions should be related to the video content but not only describing the scene in the video.

For example, one question-answer pair for another video can be \" Question: How many people are there in the video? What are they doing? === Answer: Two. There is a man in a blue shirt playing with a Wii. The same man also plays ping pong in a different scene. Another young boy wearing a red shirt is dancing. \"

Make the question challenging by involving time-related activities and not including the visual details in the question so that the user needs to reason about that first. Always answer as if you have directly watched the video. Output questions and answers directly.'

Figure 1. The detailed prompt for reasoning QA generation of Panda-WebVid30K.

f 'You are an AI visual assistant that can analyze video, which is provided as frames extracted from video. Apart from the frames, some short captions are also provided as follows: \"{captions}\"

Design 5 question-answer pairs between you and a person specifically related to temporal understanding from the video content. You need to first play the role of a human who asks a question about the temporal sequence or timing of events in the video and then play the role of an AI assistant that provides information based on the video content. The answers should be in a tone that a visual AI assistant is watching the video and answering the question. \n Ask diverse questions and give corresponding answers.\n

The answer must be detailed and descriptive, and should directly reference the information provided with respect to the temporal sequence of events in the video. Only include questions that have definite answers: \n (1) one can see the content in the video that the question asks about and can answer confidently; \n (2) one can determine confidently from the video that it is not in the video. \n Do not ask any question that cannot be answered confidently. \n Also include complex questions that are relevant to the content in the video, for example, asking about background knowledge of the objects in the video, asking to discuss about events happening in the video, etc. Again, do not ask about uncertain details. \n Provide detailed answers when answering complex questions. For example, give detailed examples or reasoning steps to make the content more convincing and well-organized. \n Generate the question-answer pairs in \"Question: <generated question> \n Answer: <generated answer>\" format.'

Figure 2. The detailed prompt for temporal QA generation of Panda-WebVid30K.



**Prompt:** The input images are frames from a video. Generate 5 high-level reasoning questions about the visual effect and the corresponding answers based on the video. For example, question 1 is 'what is the visual effect applied in this video?', answer is 'The visual {} in this video is a heart-shaped effect to show love. The lights are shimmering and appear to be hearts, giving the impression that they are floating in the water. The purpose of this effect is likely to create a beautiful and romantic atmosphere, perhaps depicting a serene night scene with the stars reflected in the water.

**Generated QA:** Based on the provided images, which show a progression of frames from a video, here are five high-level reasoning questions and their corresponding answers regarding the visual effect applied:\\n\\nQuestion 1: What is the visual effect applied in this video?\\nAnswer 1: The visual effect applied in this video appears to be an edge-detection filter combined with color inversion, creating an outline-based representation of the scene with a neon glow effect on the edges, giving it a visually striking, graphic quality that emphasizes motion and form.\\n\nQuestion 2: How does the visual effect impact the viewer's perception of movement within the video?\\nAnswer 2: The edge detection and glowing effect applied to the visual accents the movement of the cyclist, converting the dynamic motion into bright streaks and contours. This effect enhances the viewer's perception of speed and fluidity of movement as the bright outlines shift rapidly against the dark background. . . .

Question: What is the visual effect applied in this video?

**Answer:** The visual effect applied in this video appears to be an "Edge Detection" or "Glowing Edges" filter. This effect emphasizes the edges of the objects within the video frame, creating an outline or glowing line around shapes and contours. The result is a graphic, almost sketch-like appearance where the edges of objects are highlighted to stand out against a darker, more subdued background. This type of effect can be achieved through various video editing or compositing software programs that possess the capability to analyze the luminance values of the pixels within an image or video frame and then apply a stylized edge enhancement based on those values. The intensity, colors, and width of the edges can often be adjusted to achieve the desired artistic effect. The frames suggest that the filter has been applied with a neon or glowing color palette, creating a vivid high-contrast look that gives the video a more dramatic and stylized appearance.

**Question:** How does the visual effect impact the viewer's perception of movement within the video? **Answer:** The visual effect applied to these frames gives them a neon outline or edge-detection style look. This visual style enhances the contours and edges of objects and figures within the video, making shapes more defined and pronounced against a darker background.\n\nWith regards to movement, this effect can make the direction and speed of motion more visually striking. The glowing outlines can trace the path of a moving object or person, creating a trail that emphasizes the trajectory of the movement. For instance, as the main subject moves through the scene, the neon edges follow dynamically, which can make the motion appear more fluid and add a sense of swift motion. \n\nThis effect might also add a sense of rhythm to repetitive actions, enhancing the viewer's perception of speed by highlighting the outlines more sharply with each frame. In rapidly changing scenes, this could create a strobe-like impression, making the action feel more intense or energetic. \n\nHowever, it's important to use this effect thoughtfully as it could be distracting if overused or if the contrast against the background is too stark, potentially hindering the viewer's ability to smoothly follow the action, especially in complex scenes with multiple moving elements.

Figure 3. Example of QA generation for effect category.



**Prompt:** The input images are frames from a video. Generate 5 high-level reasoning questions and corresponding answers based on the funny video. For example, question 1 is 'what is funny about the person in this video?', answer is 'The person and the dog are both surprised by each other with shouting. The person's face looks funny.'

**Generated QA**: Based on the series of images provided, which appear to show frames from a humorous video involving a dog and a long stick, here are five high-level reasoning questions along with their corresponding answers:\\n\\n\*Question 1:\*\* What challenge is the dog trying to overcome in the video, and why is this scenario found to be amusing?\\n\*\*Answer:\*\* The dog is trying to pass through a narrow gate while carrying a long stick in its mouth. The scenario is amusing because the stick's length is much greater than the width of the gate, causing the dog to bump into the gateposts and struggle to figure out how to get through.\\n\\n\*\*Question 2:\*\* How does the dog's determination add to the humor of the video?\\n\*\*Answer:\*\* The dog's determination is humorous because it persistently attempts to pass through the gate without understanding the problem, showing a kind of single-mindedness that is both endearing and comical....

**Question:** What challenge is the dog trying to overcome in the video, and why is this scenario found to be amusing?

Answer: The dog in the video is trying to overcome the challenge of carrying a long stick through a narrow gate opening. It appears that the dog has found a stick that is almost as long as the width of the gate and is struggling to figure out how to get it through the opening. The dog is seen trying multiple times, turning and twisting the stick, getting stuck, and even falling over in the process.\n\n This scenario is found to be amusing for several reasons. Firstly, it's funny to watch the dog's determination and persistence in trying to fit the stick through the gate, despite it being a seemingly impossible task. Secondly, the dog's actions are quite relatable, as humans often find themselves in similar situations where we try to fit something through a space that is just too small. Lastly, the dog's antics and behavior are inherently cute and entertaining, making the video enjoyable to watch for anyone who loves animals and appreciates a good chuckle.

Question: How does the dog's determination add to the humor of the video?

Answer: The dog's determination in the video adds to the humor because, despite the seemingly insurmountable obstacle it faces in the form of a stick much larger than the entrance gate, the dog refuses to give up. The stick is too wide to fit through the bars of the gate, and in its attempt to enter, the dog comically hits the gate multiple times, causing it to bounce back. The humor is further amplified by the dog's persistence in repeatedly trying to get through the gate with the stick, showcasing the classic \"dogged\" determination dogs are known for. The situation is humorous because the viewer can see the futility of the dog's efforts, yet the dog continues to try, oblivious to the fact that the stick will not fit through. The dog's actions provide a light-hearted moment as we witness its unwavering commitment to getting the stick through the gate, no matter how impossible the task may seem.

Figure 4. Example of QA generation for funny category.



**Prompt:** The input images are frames from a video. Generate 5 high-level reasoning questions and corresponding answers based on the meme video. For example, question 1 is 'what is ironic about the person in the video?', answer is 'The person is a beginner in the gym, watching other people drink some kind of powder with water. So he buys lots of powdered tea to drink, which would not help muscle building at all. Actually, other people were drinking protein powder.'

**Generated QA**: Based on the provided frames from a meme video, here are five high-level reasoning questions along with their corresponding answers:\\n\\n1. Question: What aspect of gameplay is the meme referring to with \"My gang beast character climbing out of incinerator last second\"?\\n Answer: The meme refers to a dramatic or unexpected moment in the game \"Gang Beasts,\" where the player\\'s character narrowly escapes elimination by climbing out of an incinerator just as they are about to lose.\\n\\n2. Question: What role does the \"Nobody:\" line play in the meme\\'s format?\\n Answer: The \"Nobody:\" line is used to set up a contrast between the absence of a prompt or challenge and the exaggerated and unprovoked response of the Gang Beasts character\\'s action. This implies humor derived from the character\\'s over-the-top reaction to a non-existent call to action. . . .

**Question:** What aspect of gameplay is the meme referring to with \"My gang beast character climbing out of incinerator last second?

Answer: The meme refers to an intense and often dramatic moment in the game Gang Beasts, where a player's character is on the brink of being eliminated by falling into the incinerator but manages to climb out just before being knocked out of the game. This scenario captures a common experience for players, where despite the odds being seemingly stacked against them being on the edge of defeat in a high-stakes area like the incinerator they pull off an unexpected comeback. Gang Beasts is known for its humorous and unpredictable physics-based gameplay, and climbing is one of the key mechanics. Characters cling to surfaces with jelly-like movements, adding a sense of urgency and comedy to scenarios where they're dangling on the verge of doom. The meme plays on the often last-minute, chaotic energy of the game, highlighting the sensation of narrowly avoiding a loss, something many players can relate to. The situation is exaggerated for comic effect, as escaping the incinerator in the last possible moment is both thrilling and typically highly amusing due to the game's quirky physics.

Question: What role does the "Nobody:" line play in the meme's format?

**Answer:** The "Nobody:" line in meme formats typically functions as a setup for a non sequitur or punchline that follows. It implies that without any prompt or reason, something unexpected or absurd happens. In this particular meme format, where the phrase "Nobody:" is followed by a specific example, like "My gang beast character climbing out of incinerator last second," it suggests that the example is occurring spontaneously and without any external influence or provocation. The humor often comes from the juxtaposition of the implied silence or inactivity suggested by "Nobody:" with the sudden, often exaggerated action that is described immediately after. This format is commonly used to highlight actions or responses perceived as over-the-top or amusingly dramatic in common situations or scenarios.

Figure 5. Example of QA generation for meme category.



The input images are frames from a video. Generate 5 high-level reasoning questions about the game tool and corresponding answers based on the video. For example, question 1 is 'what is the game tool in this video?', answer is 'The game tool controls the driving of a motorcycle with the nose of the person in the camera. The driving score is displayed at the top of the screen. This game tool is usually used by young people for interactive funny videos.

Based on the sequence of images provided, here are five high-level reasoning questions about the game tool and corresponding answers:\\n\\n1. Question: What is the objective of the game as shown in the video?\\n Answer: The objective of the game is to navigate a car through a road course and reach the destination, which appears to be marked with a \"Perfect score!\" text and a score of 1000 at a gas station.\\n\n2. Question: How is the player\\'s progress and success measured in this game?\\n Answer: The success in the game appears to be measured by the ability to navigate the car through the course successfully and reach the destination without crashing or going off the road. A numerical score, which shows up as \"Perfect score! 1000\" in the last frames, indicates the player\\'s performance. . . .

Q: What is the objective of the game as shown in the video?

A: Based on the input frames provided, the objective of the game appears to be to drive a car along a road and reach a specific destination or checkpoint. The frames show the car approaching a gas station with an arrow indicating the direction to go. It seems that the player needs to navigate the car to the gas station, possibly to refuel or to complete a task related to the gas station.

#### Q: How is the player's progress and success measured in this game?

A: Based on the input frames, it appears that the player's progress and success in this game are measured by completing specific tasks or challenges within the game environment. \n\n The frames show a virtual car driving towards a gas station with an arrow pointing downwards, indicating that the player needs to reach that location. Once the car reaches the gas station, the player may need to refuel the car or perform other interactions to complete the task. There is also a visible score counter on one of the frames, which suggests that the player's success may also be measured by the points they accumulate throughout the game.\n\n Additionally, the presence of other cars and obstacles in the environment implies that the player may need to navigate through traffic and avoid collisions to progress in the game successfully. The overall goal of the game may involve reaching certain destinations, completing tasks at those locations, and earning points to increase the player's score.\n\n Overall, the player's progress and success are likely measured by their ability to complete objectives, navigate the game environment skillfully, and accumulate a high score.

Figure 6. Example of QA generation for game category.

Category	Example Question				
Effect	What is the visual effect applied in this video? What visual transition is applied in this video?				
Funny	What is the humorous element introduced in this video? What might be happening in the video and why is it humorous				
Meme	What is the central theme of this meme video? What is the main theme of the meme video?				
Game	What is the game tool in this video? What is the objective of the game as depicted in the video?				

Table 1. Template questions for TikTok-CapCut2K when all GPT4-V generated questions for one video are incorrect.

Method	LLM	Effect	Funny	Meme	Game	Avg.
GPT-4V(Azure) [5]	NA	44.6/2.3	49.3 / 2.5	80.6 / 4.1	61.8 / 3.0	59.1 / 3.0
LLaVA-NeXT-Video-7B [1] VideoChat2-Mistral [2] PLLaVA [6] LLaVA-Hound-DPO [7]	Vicuna-1.5-7B Mistral-Ins-0.2-7B LLaVA-v1.6-Vicuna-7b-hf [1] Vicuna-7B-1.5	23.9 / 1.4 28.9 / 1.4 25.6 / 1.4 38.0 / 2.0	29.7 / 1.8 37.6 / 2.0 37.7 / 2.1 48.6 / 2.5	26.0 / 1.9 18.3 / 1.3 33.7 / 2.1 39.4 / 2.1	28.9 / 1.7 26.3 / 1.5 28.9 / 1.8 39.5 / 2.4	27.1 / 1.7 27.8 / 1.6 31.5 / 1.9 41.4 / 2.3

Table 2. Performance of concurrent video LMMs on the proposed EditVid-QA evaluation set. The performance is reported in the form of x/y indicating accuracy and score from GPT-4 judge with keyword filtering, i.e., removing invalid answers with "sorry" or "apologize". The same GPT-4 version (0613) is used for all methods in this table.

Judge	 Method	LLM	AN-OA	VideoChatGPT				
14481				CI	DO	CU	TU	СО
GPT-4	GPT-4V(Azure) [5]	NA	43.1 / 2.1	2.5	2.6	2.8	1.6	2.6
	VideoChat2-Mistral [2]	Mistral-Ins-0.2-7B	48.0 / 2.4	2.2	2.0	2.6	1.9	2.7
	PLLaVA [6]	LLaVA-v1.6-Vicuna-7b-hf [1]	53.1 / 2.6	2.1	2.3	2.5	1.5	2.5
	LLaVA-Hound-DPO [7]	Vicuna-7B-1.5	49.4 / 2.5	2.3	2.1	2.7	2.0	2.7
GPT-3.5	GPT-4V(Azure) [5]	NA	55.4 / 2.8	3.0	2.8	3.1	2.1	2.7
	VideoChat2-Mistral [2]	Mistral-Ins-0.2-7B	53.3 / 3.6	2.4	2.1	2.7	2.0	2.7
	PLLaVA [6]	LLaVA-v1.6-Vicuna-7b-hf [1]	66.8 / 3.6	2.1	2.3	2.5	1.5	2.6
	LLaVA-Hound-DPO [7]	Vicuna-7B-1.5	61.6/3.6	2.4	2.1	2.7	2.0	2.7

Table 3. Performance comparison of concurrent video LMMs on ActivityNet-QA (AN-QA) datasets and the VideoChatGPT benchmark [4]. The abbreviations CI, DO, CU, TU, CO denote the correctness of information, detail orientation, context understanding, temporal understanding, and consistency. The performance on ActivityNet-QA datasets is reported in the form of x/y indicating accuracy and score from GPT-4/3.5 judge with keywords filtering, i.e., removing invalid answers with "sorry" or "apologize". The same GPT-4 (0613) and GPT-3.5-turbo (0301) versions are used for all methods in this table.

{"role": "system", "content":

"You are an intelligent chatbot designed for evaluating the correctness of generative outputs for question-answer pairs."

"Your task is to compare the predicted answer with the correct answer and determine if they match meaningfully. Here's how you can accomplish the task:"

"\_\_\_\_\_"

"##INSTRUCTIONS:"

"- Focus on the meaningful match between the predicted answer and the correct answer.\n"

"- Consider synonyms or paraphrases as valid matches.\n"

"- Evaluate the correctness of the prediction compared to the answer."}

{"role": "user", "content":

"Please evaluate the following video-based question-answer pair:\n\n"

f"Question: {question}\n"

f"Correct Answer: {answer}\n"

f"Predicted Answer: {pred}\n\n"

"Provide your evaluation only as a yes/no and score where the score is an integer value between 0 and 5, with 5 indicating the highest meaningful match. "

"Please generate the response in the form of a Python dictionary string with keys 'pred' and 'score', where value of 'pred' is a string of 'yes' or 'no' and value of 'score' is in INTEGER, not STRING." "DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string. "

"For example, your response should look like this: {'pred': 'yes', 'score': 4.8}."}

Figure 7. The input prompt for evaluation with GPT-4 judge.

Sorry, I can't help with identifying or making assumptions about content in videos.

I'm sorry, I cannot provide information about the image content in this context.

Sorry, I can't provide assistance with identifying or making assumptions about content within videos.

I'm sorry, I can't provide assistance with that request.

I'm sorry, but I cannot provide assistance with the content as presented.

Sorry, I can't provide information about this video as the main content is not visible.

Sorry, I can't provide assistance with identifying or making assumptions about content from video as it goes beyond my current capabilities.

I'm sorry, but I cannot provide real-time analysis or commentary on videos or sets of images that are meant to be sequential frames from a video, as I cannot view videos or understand the continuity in a series of still images.

Sorry, I can't provide information on the tool used to create content in a video.

I'm sorry, but I cannot assist with requests involving video content or provide information on the objectives of a game based on video stills.

Sorry, I can't help with identifying or making assumptions about people in images.

I'm sorry, but I'm unable to provide real-time analysis on videos, visuals, or any other media that changes over time since I can only evaluate static images that are provided to me.

I'm sorry, but I cannot provide assistance with videos or live motion images.

I'm sorry, but I cannot provide real-time analysis or commentary on video content.

I'm sorry, I can't provide information about the contents of these images.

Table 4. Sorry-style answers we observed in GPT4-V responses. We use the first answer in the table for "sorry attack" evaluations.

EditVid-QA				VideoChatGPT			
Effect	Funny	Meme	Game	General Temporal		Consistency	
3.3%	8.1%	4.9%	5.3%	6.3%	6.4%	20.6%	

Table 5. GPT4-V sorry-style response ratio on EditVid-QA and VideoChatGPT benchmarks.



#### **Q:** What visual effect is applied in this video?

**Ground-truth:** In the video there is a visual effect called "glitch," where the video appears distorted or broken, with colors appearing to "glitch" or shift rapidly and unexpectedly. It is commonly used in digital art and music videos for its surreal and futuristic aesthetic.

**GPT4-V:** The visual effect applied in this video sequence appears to be data moshing. Data moshing is a digital technique that manipulates the data of media files in order to achieve visual or auditory effects when the file is decoded. In video data moshing, frames may be blended or overlaid in unintended ways that can create a glitchy or surreal effect, often with smearing or ghosting of moving objects. This specific effect often results from errors in the compression and decompression process of digital video, where predictive frames (P-frames and B-frames) that rely on previous data for rendering can become corrupted, leading to the incorrect rendering of certain parts of the video. The images show a cat moving with parts of its body and the background smeared and distorted, which is typical of a data moshing artifact.



**Ours:** The visual effect applied in this video is called "glitch" or "digital distortion." This effect creates a sense of disruption or error in the digital image, often used to convey a sense of chaos, technical failure, or to add a stylistic, edgy look to the video. The glitch effect can be achieved through various methods, such as data moshing, where the pixels are randomly shuffled, or by manipulating the digital signal of the video. It's a popular effect in music videos, film, and digital art to create a sense of tension or to add a modern, edgy aesthetic.

**Q:** What causes the comedic effect in the sequence of images?

**Ground-truth:** The video shows two dogs with their leashes tied to a tree. As the golden dog moves around, its leash wraps around the neck of the other dog and gets tightened. This unexpected consequence is considered comedic by the viewers.

**GPT4-V:** The comedic effect in the sequence of images is achieved through the unexpected reveal of a fake dog leg attached to a rope, giving the illusion that a dog is standing on its hind legs holding the other end of the rope with its front paws. The first image sets up the scene with what appears to be a dog's hind leg and paw holding the rope. As the sequence progresses, the perspective shifts, revealing that the 'leg' is actually a static fake appendage, and a real dog then enters the scene from behind it, completing the visual joke by feigning to be the one holding the rope. The humor comes from the initial deception and the surprising twist when the actual dog appears, playfully subverting the viewers' expectations.

**Ours:** The comedic effect in the sequence of images is primarily caused by the unexpected and exaggerated actions of the dog. The dog appears to be engaging in a playful or mock-aggressive behavior, possibly mimicking the actions of a larger animal like a bear. This anthropomorphic behavior, where the dog is standing on its hind legs and appears to be "boxing" with its front paws, is humorous because it is an unusual and unexpected sight. Additionally, the dog's facial expressions and body language, such as the wide-eyed stare and the flailing arms, add to the comedic effect by giving the impression that the dog is taking the playful interaction very seriously. The contrast between the dog's small size and the exaggerated boxing motions also contributes to the humor, as it creates a visual gag that is amusing to the viewer. Overall, the comedic effect is achieved through the combination of the dog's unexpected behavior, the anthropomorphic nature of its actions, and the exaggerated physical comedy of the interaction.

Figure 8. Qualitative results of our model and GPT-4V on the proposed EditVid-QA benchmark. Watermarks are removed for anonymity.



**Q:** What type of humor is being used in this video?

**Ground-truth:** The video uses a type of humor known as "anthropomorphism," where animals are given human-like characteristics or behaviors. In this video, the creator is using a raccoon to depict a comical scenario of a boyfriend getting into the shower without his partner. The raccoon, labeled as "BF" (boyfriend), acts as if it is taking a shower while the other raccoon, labeled as "Me," amusingly tries to get in. The humor comes from the unexpected and exaggerated portrayal of a typical couple's interaction, using cute and mischievous raccoons to play out the scene.

**GPT4-V:** The humor in this video appears to be slapstick and visual gag-based. Slapstick humor is characterized by physical comedy and exaggerated actions, often leading to unexpected and humorous outcomes. In this case, the gag seems to involve a cat attempting to enter a shower, but encountering an obstacle that prevents it from doing so. The humor comes from the cat's persistent and futile efforts, as well as the unexpected nature of the barrier. The video likely includes a build-up of the cat's attempts, followed by a sudden and surprising reveal of the obstacle, which is a common structure in slapstick comedy. The use of text overlay ("When you're trying to get in the shower but your bf is in there") adds a relatable and humorous context to the situation, making it more amusing to the audience. Overall, the humor is physical, unexpected, and relatable, which are key elements of slapstick comedy.



**GPT4-V:** The images show a virtual representation of a pool table with a cue stick, balls, and pockets. It appears to be a mobile or augmented reality (AR) game simulating the game of billiards or pool. The "WIN" trophy graphic and the "Reset" button suggest that the player has completed a game or challenge within the application. The gameplay elements such as the pool cue and balls are digitally superimposed over a real-world background of a stone wall, which is characteristic of AR gaming experiences where digital and real-world elements are combined.

**Ours:** The humor being used in this collection of images from a video is a form of situational comedy and anthropomorphism, where animals are humorously positioned in a scenario typically involving humans. The images present a sequence from a point of view (POV) perspective supposedly humorously depicting what happens when a person's boyfriend gets into the shower without them, represented by raccoons in the shower with captions identifying one raccoon as "BF" (boyfriend) and the other as "Me." The incongruity of using animals to play out a human relationship scenario, combined with the unexpectedness and the playful anthropomorphic portrayals, creates a humorous effect. This form of humor might be particularly popular on social media platforms, where cute or unexpected animal behavior is used to create relatable or amusing content.

**Q:** What is the game tool displayed in the video frames?

**Ground-truth:** The input images are frames from a video that showcases an Augmented Reality (AR) pool game tool. The game tool displays virtual elements overlaid on the real-world environment. Specifically, the frames show a digital pool table with cues and pool balls rendered in 3D graphics, as if they existed in the physical space. This AR pool game tool uses the camera of a smartphone or a tablet to scan the surrounding environment and places virtual objects like the pool table, cues, and balls in the real world through the device's screen.

**Ours:** The game tool displayed in the video frames appears to be a billiards table, specifically a pool table. The table has a green cloth surface with a black border and a white cue ball on it. The presence of a cue stick and a pool cue ball indicates that the game being played is pool or billiards. The table also has a triangle rack with red and yellow billiard balls, which are commonly used in the game. The game involves striking the cue ball with the cue stick to hit other balls into pockets on the table. The video frames show the progression of the game, with the balls being hit and moved around the table. The game is typically played for recreation or as a competitive sport, and requires skill and strategy to successfully pot the balls.

Figure 9. Qualitative results of our model and GPT-4V on the proposed EditVid-QA benchmark. Watermarks are removed for anonymity.