Vista-LLaMA: Reliable Video Narrator via Equal Distance to Visual Tokens

Supplementary Material

6. Additional Results

Visualization Results. We present additional visualization results for different video questions in Figures 8 and 9. In comparison to Video-ChatGPT [15], our Vista-LLaMA provides more reasonable answers and descriptions that align better with the video content. Video-ChatGPT often responds with irrelevant information, resulting in hallucinations. For example, in the first video where the baby appears tired, Video-ChatGPT incorrectly states that the baby was eating a snack, even though there is no eating action shown in the video. We have more examples that demonstrate the improvement of our method on NExT-QA [28]. We only showcase a small portion of these cases to highlight the differences. Our methods achieve much better performance and offer more reliable replies due to the proposed EDVT-Attention, which maintains an equal distance to all visual tokens.

Attention Weights in Different Layers. In Figure 10, we present the attention weights in different layers. Different from Figure 5, here we sum instead of average the attention weights of 32 heads to present clear comparison. From the figure, we show that the attention weights between text tokens in the EDVT-Attention are larger than attention weights in Vanilla attention. It indicates that the EDVT-Attention strengthen the impact of visual tokens on generating text.

In Figure 10, we visualize the attention weights in different layers. In contrast to Figure 5, where we averaged the attention weights of 32 heads, here we present the sum for a clearer comparison. The figure reveals that the attention weights between text tokens in the EDVT-Attention are greater than those in Vanilla attention. This suggests that the EDVT-Attention enhances the influence of visual tokens on text generation.

Positional Embedding Study. We explored various strategies for positional embedding in the attention layer, focusing on the query and key vectors. According to Tab. 5, the model achieves the highest accuracy when only text tokens have rotary positional embedding applied to both the query and key vectors. When only the query vectors have RoPE applied and the key vectors do not, the performance decreases significantly. This is because the relative distance is compromised when only the query has RoPE. We also attempted to use fixed positional embedding on all visual tokens. Unlike in DEVT, all visual tokens have RoPE applied with the same position index of 0. Compared to the baseline, this modification also improves performance on different question types. However, it is still inferior to our design. This demonstrates that the proposed EDVT design truly enhances video understanding in LLMs.

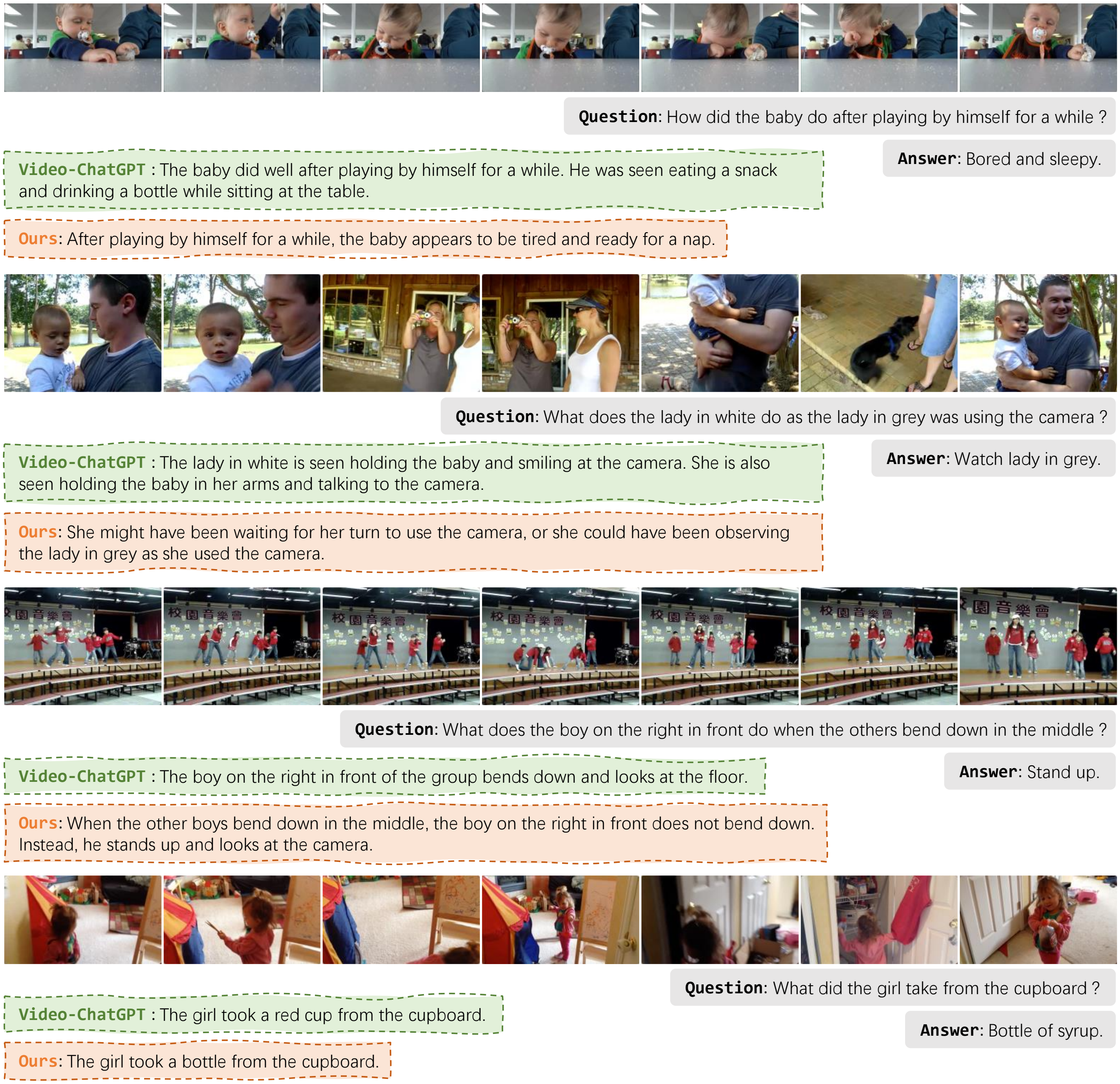
Query	**Key**	**NExT-QA [28]**			
		Tem.	**Cau.**	**Des.**	**Avg.**
RoPE	RoPE	34.3	65.8	55.9	54.1
FixVPE	FixVPE	37.0	70.5	56.7	57.6
RoPE	EDVT	32.2	48.1	41.8	42.0
EDVT	EDVT	40.7	72.3	57.0	59.7

Table 5. **Comparison of positional embedding strategies** on NExT-QA [28]. We provide a list of various positional embedding strategies used for query and key vectors in the attention layer. The "RoPE" indicates the use of rotary positional embedding for all visual and text tokens. The "FixVPE" refers to the fixed position rotary positional embedding used for all visual tokens. Lastly, "EDVT" indicates that the rotary positional embedding is exclusively applied to text tokens.

7. Movie Evaluation

Dataset Collection. In this paper, we introduce a new dataset named CineClipQA. The CineClipQA dataset encompasses a collection of 153 curated video clips, derived from five movies that span diverse genres and storytelling styles. Each clip, representing one or more distinct segment of the movie plot, is accompanied by a set of 16 tailored questions, thereby totaling 2,448 questions in various dimensions, as is presented in Figure 11. The question consists of two parts: System Prompt and Questions. The System Prompt contains basic information about key characters in the current video clip and provides prompts for the initial actions of characters when necessary. The Questions are primarily divided into five categories: Recognizing, Temporality (Prediction), Spaciality (Interaction), Intention, and Perception. Specifically, Recognizing includes questions about Location and Action; Temporality encompasses questions about the next action, the previous action, simultaneous actions, and predicted actions; Spaciality involves questions about spatial information between Object and Human; Intention involves three similar types of questions about the purpose of actions; finally, Perception examines the recognition of emotions and inquiries about the "how" (approaches, manners...). At last, we provide a detailed explanation and corresponding case for all 16 types, shown in Figure 12, Figure 13, Figure 14, Figure 15, Figure 16.

Method Comparison. In our study, we evaluated the performance of several SOTA methods, including MovieChat, VideoLLAMA, Video-ChatGPT, and our own Vista-LLaMA, on the newly proposed CineClipQA dataset. This dataset,



Question: How did the baby do after playing by himself for a while ?

Answer: Bored and sleepy.

Video-ChatGPT : The baby did well after playing by himself for a while. He was seen eating a snack and drinking a bottle while sitting at the table.

Ours: After playing by himself for a while, the baby appears to be tired and ready for a nap.

Question: What does the lady in white do as the lady in grey was using the camera ?

Answer: Watch lady in grey.

Video-ChatGPT : The lady in white is seen holding the baby and smiling at the camera. She is also seen holding the baby in her arms and talking to the camera.

Ours: She might have been waiting for her turn to use the camera, or she could have been observing the lady in grey as she used the camera.

Question: What does the boy on the right in front do when the others bend down in the middle ?

Answer: Stand up.

Video-ChatGPT : The boy on the right in front of the group bends down and looks at the floor.

Ours: When the other boys bend down in the middle, the boy on the right in front does not bend down. Instead, he stands up and looks at the camera.

Question: What did the girl take from the cupboard ?

Answer: Bottle of syrup.

Video-ChatGPT : The girl took a red cup from the cupboard.

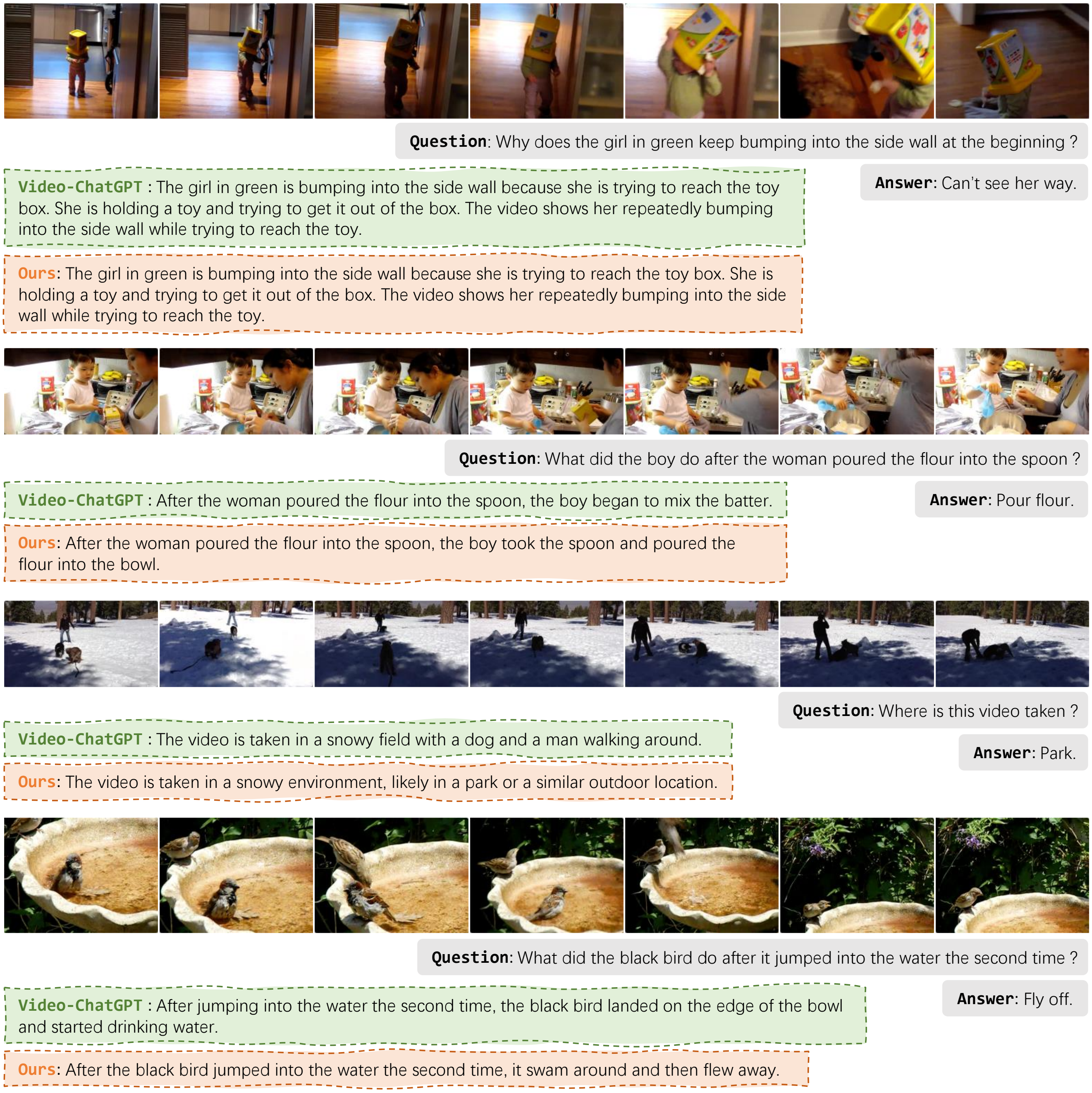
Ours: The girl took a bottle from the cupboard.

Figure 8. **Visualization results** on different video questions.

Method	Overall		Description		Temporality		Spaciality		Intention		Perception	
	Score	Accuracy	Score	Accuracy	Score	Accuracy	Score	Accuracy	Score	Accuracy	Score	Accuracy
MovieChat	2.11	20.86	2.41	23.67	1.97	16.32	1.98	16.40	2.41	30.19	1.97	21.80
Video-LLAMA	2.27	23.17	2.31	19.30	2.12	16.35	2.19	21.95	2.47	31.94	2.35	27.70
Video-ChatGPT	2.60	34.11	2.55	26.24	2.60	34.11	2.50	30.62	2.94	46.36	2.43	31.77
VISTA-LLAMA (Ours)	**2.98**	**44.90**	**2.79**	**31.46**	**2.92**	**46.22**	**2.73**	**35.63**	**3.38**	**61.89**	**3.12**	**47.49**

Table 6. **Performance Comparison on CineClipQA** of different methods on various classifications.

designed to assess comprehension of complex movie content, encompasses challenges in high content complexity, extensive scene variety, and prolonged temporal dimensions. The empirical results from our evaluations demonstrate that



Question: Why does the girl in green keep bumping into the side wall at the beginning ?

Answer: Can't see her way.

Video-ChatGPT : The girl in green is bumping into the side wall because she is trying to reach the toy box. She is holding a toy and trying to get it out of the box. The video shows her repeatedly bumping into the side wall while trying to reach the toy.

Ours: The girl in green is bumping into the side wall because she is trying to reach the toy box. She is holding a toy and trying to get it out of the box. The video shows her repeatedly bumping into the side wall while trying to reach the toy.

Question: What did the boy do after the woman poured the flour into the spoon ?

Answer: Pour flour.

Video-ChatGPT : After the woman poured the flour into the spoon, the boy began to mix the batter.

Ours: After the woman poured the flour into the spoon, the boy took the spoon and poured the flour into the bowl.

Question: Where is this video taken ?

Answer: Park.

Video-ChatGPT : The video is taken in a snowy field with a dog and a man walking around.

Ours: The video is taken in a snowy environment, likely in a park or a similar outdoor location.

Question: What did the black bird do after it jumped into the water the second time ?

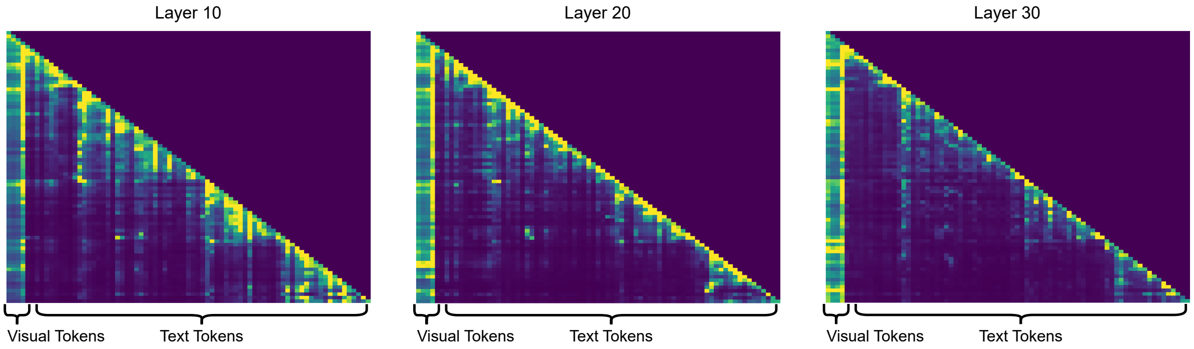
Answer: Fly off.

Video-ChatGPT : After jumping into the water the second time, the black bird landed on the edge of the bowl and started drinking water.

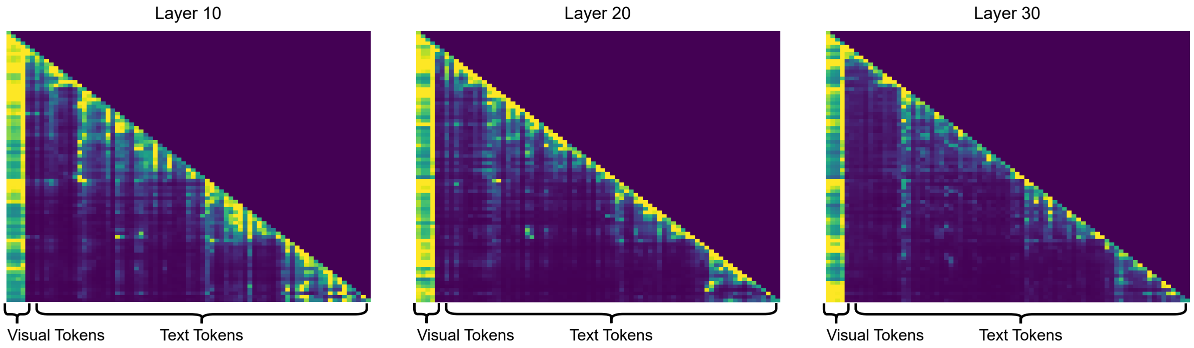
Ours: After the black bird jumped into the water the second time, it swam around and then flew away.

Figure 9. **Visualization results** on different video questions.

all methods, including our Vista-LLaMA, achieved state-of-the-art (SOTA) performance on the CineClipQA dataset. This outcome unequivocally confirms the superior ability of our approach in understanding videos with intricate content, wide-ranging scenes, and extended time frames.Notably, across all tested models, the highest accuracy was observed in the Intention category of the CineClipQA dataset. This suggests a particularly effective grasp of human behavioral reasoning, likely attributed to the rich prior knowledge embedded within these large language models (LLMs). The Intention category, by its nature, demands an in-depth analysis of purpose and motive behind actions depicted in the video clips, a task which seems to align well with the inherent strengths of current LLMs.Furthermore, this finding underscores the potential of LLMs in bridging the gap between mere visual recognition and deeper narrative under-



(a) Vanilla attention



(b) EDVT-Attention (Ours)

Figure 10. **Comparison of attention weights** for varing context lengths in different layers. Lighter colors represent higher weights. To improve clarity, we have combined visual token weights into the first four tokens. We recommend zooming in for optimal viewing.

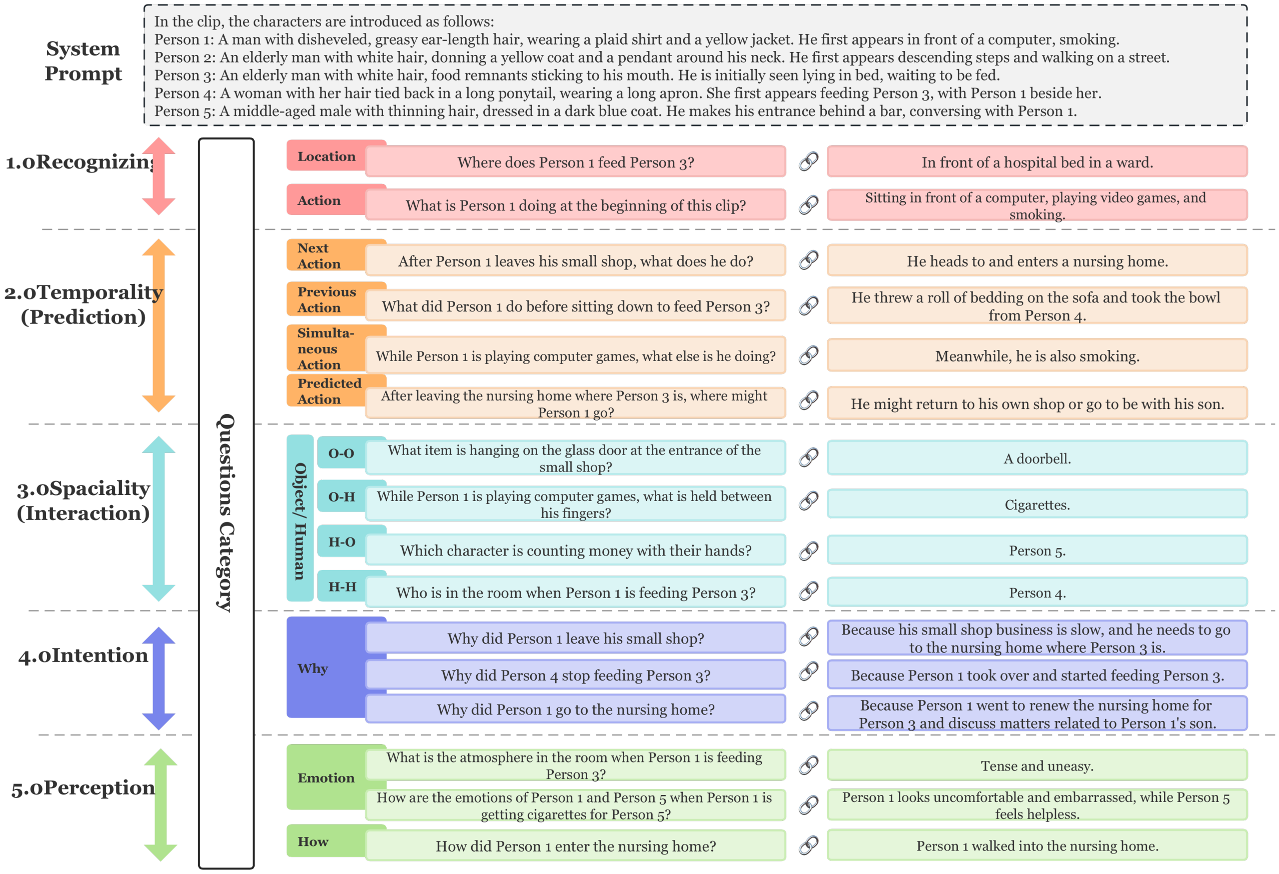
standing. The ability of these models to not only identify characters and actions but also infer underlying intentions is indicative of their advancing sophistication. It highlights a significant stride in the evolution of AI, where models are increasingly capable of nuanced interpretation akin to human-like understanding.

8. Discussion

Advantages. In this study, we present two innovations, namely the EDVT-Attention and the sequential visual projector, aimed at enhancing video comprehension in LLMs. Our evaluation primarily focuses on zero-shot question answering benchmarks. The model used is built upon LLaVA [14], which is pre-trained and then fine-tuned with video instruction data. Vista-LLaMAachieves a notable enhancement in the proposed innovations when tested on NExT-QA [28] and MSRVTT-QA [31]. Additionally, we conduct several ablations to illustrate the effectiveness of our innovations. The outcomes demonstrate the significant potential of our approach to enhance video comprehension with LLMs.

Limitations. There are also limitations in our work. For the VideoQA task, the evaluation process is assisted with GPT-3.5, which may result in some false judgments. GPT-4 might provide more accurate evaluations, but it comes at a higher cost since it is 20 times more expensive than GPT-3.5. Additionally, evaluating with GPT-4 requires the use of huge tokens, further increasing the expense. Furthermore, the evaluation speed is limited by query restrictions, and GPT-4 takes more time compared to training. We have evaluated a few cases using GPT-3.5, and the response has been reasonable and the accuracy has remained stable. When the same results are evaluated on NExT-QA three times, the variance is lower than 0.5 in the experiments.

Since this work only focuses on fine-tuning rather than pre-training, the full potential of EDVT-Attention may not be fully explored. EDVT-Attention can also be utilized for image-text related tasks. However, the impact of EDVT-Attention on pre-training, image-text related tasks, or other multi-modal tasks is not investigated in this manuscript. Additionally, the use of rotary positional embedding in some LLMs restricts the applicability of the current design. In this work, the rotary positional embedding is removed to ensure the same distance to visual tokens in decoder layers of LLMs. There may be alternative dynamic designs that can achieve this objective without eliminating the positional embedding. All these aspects are worth considering. Although the number of hallucination cases is reduced with our



System Prompt

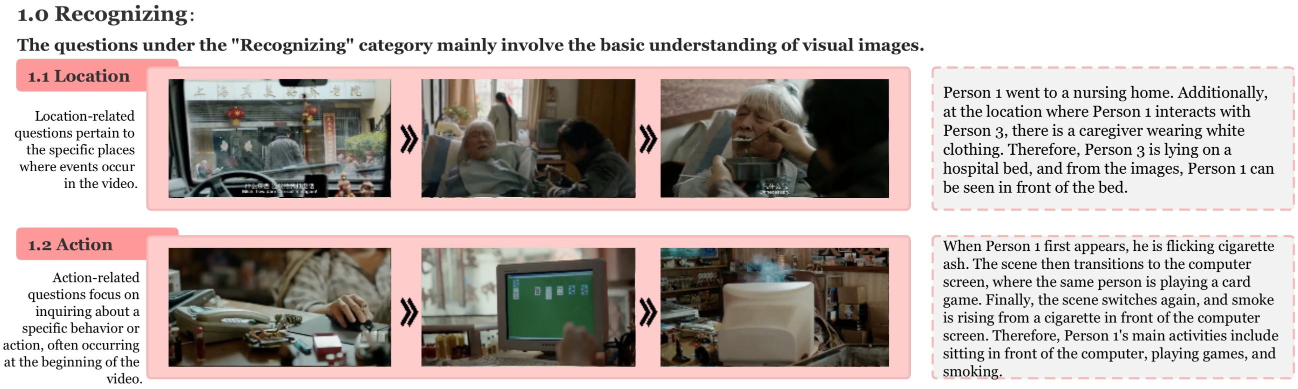
In the clip, the characters are introduced as follows:
Person 1: A man with disheveled, greasy ear-length hair, wearing a plaid shirt and a yellow jacket. He first appears in front of a computer, smoking.
Person 2: An elderly man with white hair, donning a yellow coat and a pendant around his neck. He first appears descending steps and walking on a street.
Person 3: An elderly man with white hair, food remnants sticking to his mouth. He is initially seen lying in bed, waiting to be fed.
Person 4: A woman with her hair tied back in a long ponytail, wearing a long apron. She first appears feeding Person 3, with Person 1 beside her.
Person 5: A middle-aged male with thinning hair, dressed in a dark blue coat. He makes his entrance behind a bar, conversing with Person 1.

Questions Category

1.0Recognizing		Location	Where does Person 1 feed Person 3?	In front of a hospital bed in a ward.
		Action	What is Person 1 doing at the beginning of this clip?	Sitting in front of a computer, playing video games, and smoking.
2.0Temporality (Prediction)		Next Action	After Person 1 leaves his small shop, what does he do?	He heads to and enters a nursing home.
		Previous Action	What did Person 1 do before sitting down to feed Person 3?	He threw a roll of bedding on the sofa and took the bowl from Person 4.
		Simultaneous Action	While Person 1 is playing computer games, what else is he doing?	Meanwhile, he is also smoking.
		Predicted Action	After leaving the nursing home where Person 3 is, where might Person 1 go?	He might return to his own shop or go to be with his son.
3.0Spaciality (Interaction)	Object/ Human	O-O	What item is hanging on the glass door at the entrance of the small shop?	A doorbell.
		O-H	While Person 1 is playing computer games, what is held between his fingers?	Cigarettes.
		H-O	Which character is counting money with their hands?	Person 5.
		H-H	Who is in the room when Person 1 is feeding Person 3?	Person 4.
4.0Intention		Why	Why did Person 1 leave his small shop?	Because his small shop business is slow, and he needs to go to the nursing home where Person 3 is.
			Why did Person 4 stop feeding Person 3?	Because Person 1 took over and started feeding Person 3.
			Why did Person 1 go to the nursing home?	Because Person 1 went to renew the nursing home for Person 3 and discuss matters related to Person 1's son.
5.0Perception		Emotion	What is the atmosphere in the room when Person 1 is feeding Person 3?	Tense and uneasy.
			How are the emotions of Person 1 and Person 5 when Person 1 is getting cigarettes for Person 5?	Person 1 looks uncomfortable and embarrassed, while Person 5 feels helpless.
		How	How did Person 1 enter the nursing home?	Person 1 walked into the nursing home.

The question consists of two parts: **System Prompt** and **Questions**. The **System Prompt** contains basic information about key characters in the current video clip and provides prompts for the initial actions of characters when necessary. The **Questions** are primarily divided into five categories: *Recognizing*, *Temporality (Prediction)*, *Spaciality (Interaction)*, *Intention*, and *Perception*. Specifically, *Recognizing* includes questions about Location and Action; *Temporality* encompasses questions about the next action, the previous action, simultaneous actions, and predicted actions; *Spaciality* involves questions about spatial information between Object and Human; *Intention* involves three similar types of questions about the purpose of actions; finally, *Perception* examines the recognition of emotions and inquiries about the "how" (approaches, manners ...).

Figure 11. **CineClipQA**, a novel dataset meticulously crafted to probe the capabilities of visual language models in comprehending and interpreting plot-driven video content.



1.0 Recognizing:

The questions under the "Recognizing" category mainly involve the basic understanding of visual images.

1.1 Location

Location-related questions pertain to the specific places where events occur in the video.

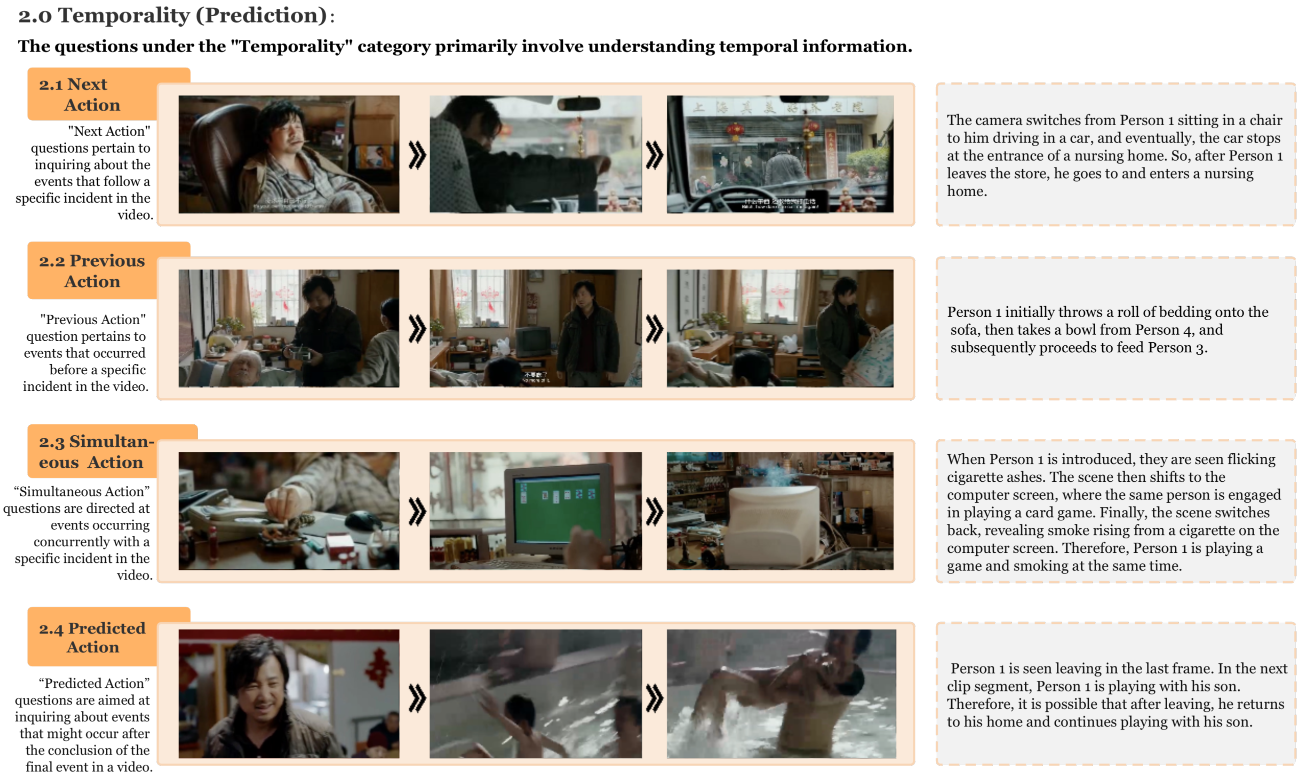
Person 1 went to a nursing home. Additionally, at the location where Person 1 interacts with Person 3, there is a caregiver wearing white clothing. Therefore, Person 3 is lying on a hospital bed, and from the images, Person 1 can be seen in front of the bed.

1.2 Action

Action-related questions focus on inquiring about a specific behavior or action, often occurring at the beginning of the video.

When Person 1 first appears, he is flicking cigarette ash. The scene then transitions to the computer screen, where the same person is playing a card game. Finally, the scene switches again, and smoke is rising from a cigarette in front of the computer screen. Therefore, Person 1's main activities include sitting in front of the computer, playing games, and smoking.

Figure 12. **CineClipQA**, the detailed description for the dataset.



2.0 Temporality (Prediction):

The questions under the "Temporality" category primarily involve understanding temporal information.

2.1 Next Action

"Next Action" questions pertain to inquiring about the events that follow a specific incident in the video.

The camera switches from Person 1 sitting in a chair to him driving in a car, and eventually, the car stops at the entrance of a nursing home. So, after Person 1 leaves the store, he goes to and enters a nursing home.

2.2 Previous Action

"Previous Action" question pertains to events that occurred before a specific incident in the video.

Person 1 initially throws a roll of bedding onto the sofa, then takes a bowl from Person 4, and subsequently proceeds to feed Person 3.

2.3 Simultaneous Action

"Simultaneous Action" questions are directed at events occurring concurrently with a specific incident in the video.

When Person 1 is introduced, they are seen flicking cigarette ashes. The scene then shifts to the computer screen, where the same person is engaged in playing a card game. Finally, the scene switches back, revealing smoke rising from a cigarette on the computer screen. Therefore, Person 1 is playing a game and smoking at the same time.

2.4 Predicted Action

"Predicted Action" questions are aimed at inquiring about events that might occur after the conclusion of the final event in a video.

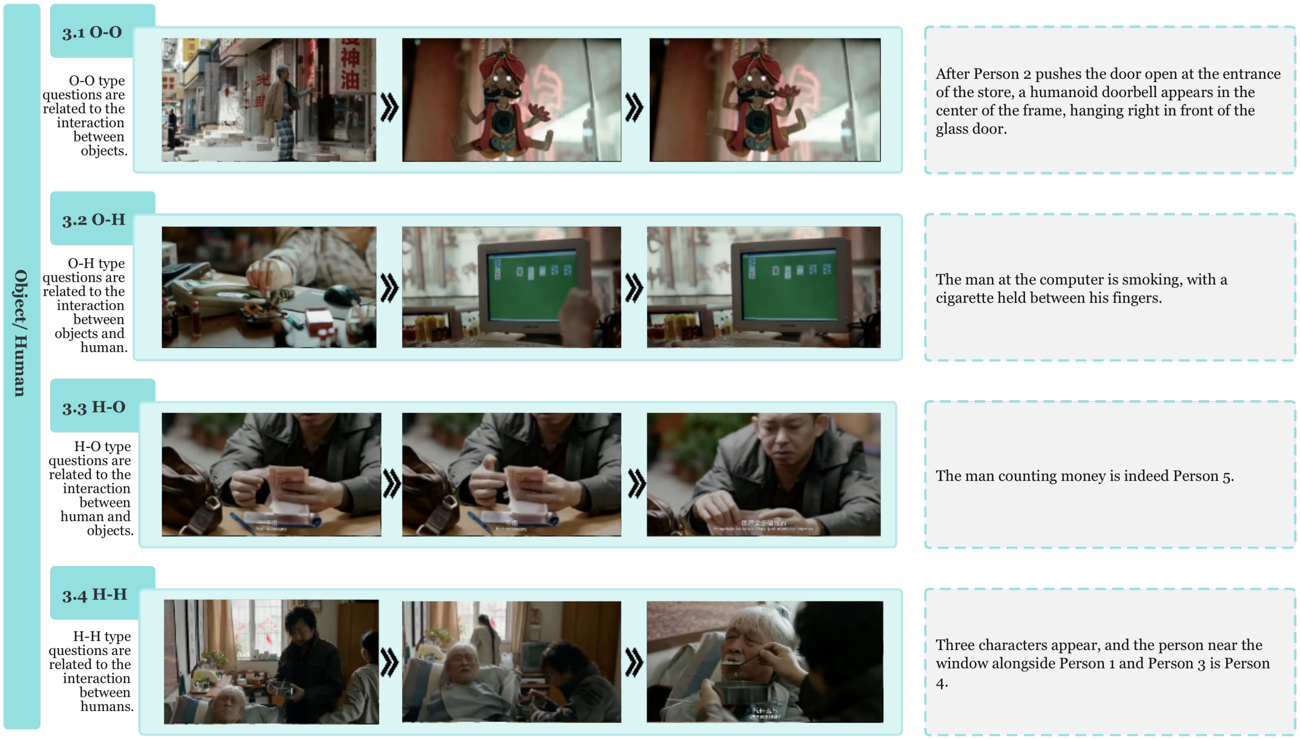
Person 1 is seen leaving in the last frame. In the next clip segment, Person 1 is playing with his son. Therefore, it is possible that after leaving, he returns to his home and continues playing with his son.

Figure 13. **CineClipQA**, the detailed description for the dataset.

method, there are still instances where the model provides irrelevant replies. Further studies are necessary to address this issue. To enhance the current manuscript, our future work will focus on developing more general designs for practical cases.

3.0 Spaciality (Interaction):

The questions under the "Spatiality" category primarily involve understanding spatial interactivity.



3.1 O-O

O-O type questions are related to the interaction between objects.

After Person 2 pushes the door open at the entrance of the store, a humanoid doorbell appears in the center of the frame, hanging right in front of the glass door.

3.2 O-H

O-H type questions are related to the interaction between objects and human.

The man at the computer is smoking, with a cigarette held between his fingers.

3.3 H-O

H-O type questions are related to the interaction between human and objects.

The man counting money is indeed Person 5.

3.4 H-H

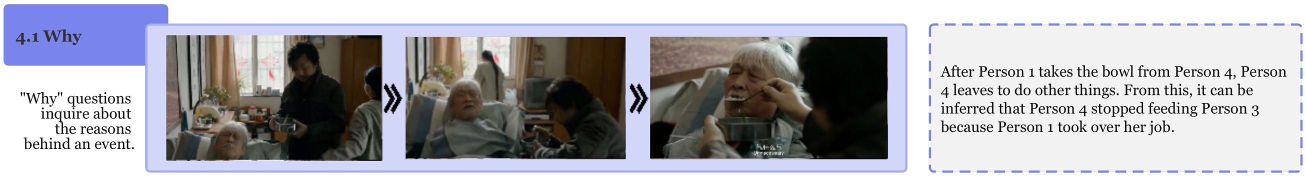
H-H type questions are related to the interaction between humans.

Three characters appear, and the person near the window alongside Person 1 and Person 3 is Person 4.

Figure 14. **CineClipQA**, the detailed description for the dataset.

4.0 Intention:

The questions under the "Temporality" category primarily involve understanding temporal information.

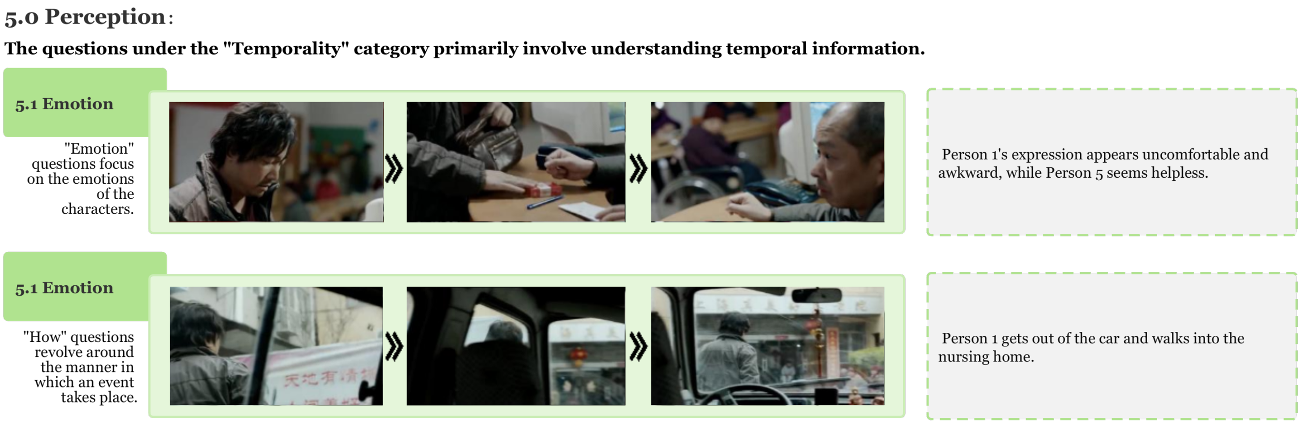


4.1 Why

"Why" questions inquire about the reasons behind an event.

After Person 1 takes the bowl from Person 4, Person 4 leaves to do other things. From this, it can be inferred that Person 4 stopped feeding Person 3 because Person 1 took over her job.

Figure 15. **CineClipQA**, the detailed description for the dataset.



5.0 Perception:

The questions under the "Temporality" category primarily involve understanding temporal information.

5.1 Emotion

"Emotion" questions focus on the emotions of the characters.

Person 1's expression appears uncomfortable and awkward, while Person 5 seems helpless.

5.1 Emotion

"How" questions revolve around the manner in which an event takes place.

Person 1 gets out of the car and walks into the nursing home.

Figure 16. **CineClipQA**, the detailed description for the dataset.