VideoICL: Confidence-based Iterative In-context Learning for Out-of-Distribution Video Understanding

Supplementary Material

A. Discussion on Hyperparameter Choice

	Animal Kingdom	PitVQA	UCF-Crime	Avg.
Zero-shot	68.0	6.7	39.3	38.0
c_{th} = 0.1	69.4	53.6	50.3	57.7
c_{th} = 0.3	69.5	58.5	50.6	59.5
c_{th} = 0.5	70.7	61.3*	53.3*	61.8
c_{th} = 0.7	72.3*	**61.6**	52.7	62.2
c_{th} = 0.9	**72.6**	61.5	**53.6**	**62.6**

Table 5. **Ablation on confidence threshold c_{th}.** The values used for the main table are marked with *.

We compare the results on varying confidence threshold c_{th} in Tab. 5. While the accuracy generally increases with $c_{th} = 0.9$, it also increases the cost of the entire process by performing more iterations per query on average. Therefore, we choose $c_{th} = 0.5$ and 0.7 for the best trade-off between cost and accuracy.

B. Details on Datasets

Animal Kingdom We use the Animal Kingdom dataset [38] for our multiple choice question answering task. This dataset includes videos of animals with action labels such as *Yawning* and *Struggling*, covering 140 unique classes. While it was originally built for action recognition tasks, we modified its format to suit a multiple-choice QA task by pairing one true action label with four randomly chosen alternative labels. The dataset provides 24,004 labeled training examples and 6,096 test examples.

Sports-QA We employ the Sports-QA [28] dataset for open-ended question answering task, which is designed for sports video question answering. This dataset includes various sports, such as basketball, football, and gymnastics, and features diverse question types like descriptions, timelines, causalities, and hypothetical scenarios. The dataset includes 56,385 training examples and 18,718 test examples.

PitVQA We also use PitVQA [17], a dataset designed for VQA in endonasal pituitary surgery videos that requires specific medical knowledge, for the open-ended question answering task. PitVQA provides question-answer annotations at the frame level. For our experiments, we process a sequence of 10 consecutive frames as the video input, with question-answer pairs drawn from the middle, fifth frame. The dataset includes 75,010 training examples and 10,832 test examples.

UCF-Crime UCF-Crime [44], which classifies the type of crime in security camera footage into 13 categories, is used for video classification task. We include all crime categories in the prompt, guiding the model to select the appropriate crime class for the given video. The dataset also includes normal event videos as challenging negative examples. The official split of UCF-Crime provides four different train and test splits, with each split consisting of 532 training samples and 168 test samples. The result is reported as the average performance across the test sets of all four splits.

Drive&Act The Drive&Act dataset [37] is utilized for video classification tasks. This offers comprehensive labels for driver behaviors inside vehicles, including action segmentation information captured in Kinect-IR videos. We extract each segment from the video and ask the model to recognize the action. The official split of Drive&Act provides three different train and test splits. Each split consists of around 2,000 labeled training examples and around 600 test examples. The result is reported as the average performance across the test sets of all three splits.

CapERA For the video captioning task, we evaluate models on the CapERA dataset [5], which is specifically curated for describing scenes captured from an aerial perspective. CapERA provides concise captions for a range of scenarios viewed from above, including concerts, harvesting, and car racing, and consists of 1,473 labeled examples for training and 1,391 for testing.

C. Proof of Asymptotic Model Accuracy

Proposition (Asymptotic Model Accuracy)**.** *Let $a(n)$ be the expected accuracy of* VideoICL *with a maximum of n confidence-based iterations. Then,*

$$\lim_{n\to\infty} a(n) = \frac{1}{1 + \frac{\mathrm{FPR}}{\mathrm{TPR}} \cdot \frac{1-p_c}{p_c}},$$

where TPR *and* FPR *stand for the true positive rate (i.e., recall) and the false positive rate of the confidence estimation method, respectively.*

Proof. At each iteration, there are three possibilities:

- The model returns a correct response and is estimated to be confident, with probability $p_c \cdot \mathrm{TPR}$.
- The model returns an incorrect response, but is estimated to be confident, with a probability of $(1 - p_c) \cdot \mathrm{FPR}$.
- The model returns a response, and is estimated to be unconfident, occurring with probability $p_u := 1 - (p_c \cdot \mathrm{TPR} + (1 - p_c) \cdot \mathrm{FPR})$.

For the first two cases, the loop terminates and returns a response, whereas in the third case, the loop continues with a new iteration. Let $c(n)$ represent the probability that the loop ends by correctly returning a response (first case) on the n-th iteration, and $l(n)$ represent the probability that the loop is still ongoing (third case) after n iterations.

The expected accuracy $a(n)$ after n iterations is the sum of the probabilities of ending with a correct response up to the n-th iteration, plus the probability of continuing after the $(n-1)$-th iteration, weighted by the probability of a correct response in the next iteration p_c:

$$a(n) = \sum_{i=1}^{n} c(i) + l(n-1) \cdot p_c. \tag{4}$$

The probability of continuing after the n-th iteration is $l(n) = l(n-1) \cdot p_u$, with $l(0) = 1$, leading to $l(n) = p_u^n$ by recursion. And the probability $c(n)$ of ending at the n-th iteration with a correct and confident response is:

$$c(n) = l(n-1) \cdot (p_c \cdot \mathrm{TPR}) = p_u^{n-1} \cdot (p_c \cdot \mathrm{TPR}), \tag{5}$$

where $p_c \cdot \mathrm{TPR}$ accounts for the likelihood that a response is classified as confident (p_c) and is also correct (TPR). Therefore, we have:

$$a(n) = p_c \cdot \mathrm{TPR} \cdot \frac{1 - p_u^n}{1 - p_u} + p_u^{n-1} \cdot p_c. \tag{6}$$

Since $0 < p_u < 1$, as $n \to \infty$,

$$\begin{aligned} \lim_{n \to \infty} a(n) &= \frac{p_c \cdot \mathrm{TPR}}{1 - p_u} \\ &= \frac{1}{1 + \frac{\mathrm{FPR}}{\mathrm{TPR}} \cdot \frac{1-p_c}{p_c}}. \end{aligned} \tag{7}$$

□

D. Additional Results

D.1. Additional Discussion on Main Results

For LoRA fine-tuning, we use a rank of 32 and train the model for 1 epoch on Animal Kingdom and PitVQA, 5 epochs on UCF-Crime, and 2 epochs on CapERA. In Tab. 1, VIDEOICL outperforms the LoRA fine-tuned model on all datasets except CapERA, showing that in-context examples are more effective than training in OOD video QA when domain knowledge requires extensive data and training.

N	Animal Kingdom	PitVQA	UCF-Crime	CapERA
0	68.0	6.7	39.3	0.027
25	69.8	28.7	43.9	0.127
50	71.3	28.0	42.4	0.125
100	70.6	33.3	47.8	0.154
Full	72.3	61.3	53.3	0.170

Table 6. **Results with N-size example pool.** BLEU-4 is reported for CapERA.

Interestingly, the LLaVA-Video-72B model underperforms compared to LLaVA-Video-7B model notably in video classification and captioning. For captioning, this is because 72B model often generates excessively long outputs filled with irrelevant details. In video classification, we suspect the limited capacity of 7B model may act as a form of regularization, helping it generalize better on OOD data, but this needs further investigation.

In addition, VIDEOICL outperforms SIMRANKVOTE in Tab. 1, highlighting the benefits of using confidence-based aggregation instead of majority voting. VIDEOICL also achieves better results than SIMRANKONCE, showing that using more examples leads to better performance. Lastly, SIMRANKVOTE outperforms RANDEXVOTE, demonstrating the effectiveness of selecting similar examples based on video and text features.

D.2. Additional Qualitative Results

In the following pages, we present qualitative results of VIDEOICL for each dataset. For each iteration, we use two examples with maximum of 4 iterations, and the outputs of the model are presented together with confidence scores.

D.3. Effect of Example Pool Size

To demonstrate that VIDEOICL does not require a large pool, we evaluate the performance with smaller pools, as shown in Tab. 6. Notably, even with a small pool of 25 examples, VIDEOICL achieves up to 22% improvement. This highlights that VIDEOICL remains effective even with limited and less related examples. Thus, users applying VIDEOICL to OOD data can take with only a handful of examples while still achieving significant performance gains.

E. Limitation

While VIDEOICL delivers remarkable performance, it does have some limitations. First, VIDEOICL requires more time compared to single-step in-context learning because it performs multiple rounds of inference. This additional computation may make it less suitable for applications that demand low latency, such as real-time video analysis. However, VIDEOICL mitigates this issue by using early termination when the model confidence in its output is sufficiently high, which significantly reduces computation time. It is also much faster than training a model from scratch.

Second, VIDEOICL relies on having an example pool to select reference examples from. We have not tested its performance with extremely small datasets. Considering the challenges of generating out-of-distribution video data, exploring the effectiveness of VIDEOICL with very limited examples is an important direction for future research.

Multiple Choice QA: Animal Kingdom

▶ **Iteration 1**

Example 1: What action is the animal doing in the video? Answer with the option's letter from the given choices directly. Options: (A) Dancing On Water (B) Urinating (C) Eating (D) Sleeping in its nest (E) Sharing Food
The answer is (C)

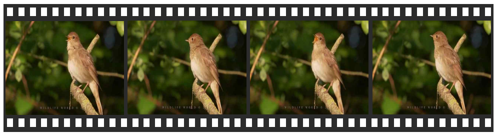
Example 2: What action is the animal doing in the video? Answer with the option's letter from the given choices directly. Options: (A) Biting (B) Climbing (C) Dancing (D) Drinking (E) Being Dragged
The answer is (D)

User: What action is the animal doing in the video? Answer with the option's letter from the given choices directly. Options: (A) Trapped (B) Attending (C) Walking On Water (D) Sharing Food (E) Showing Affection
LLaVA-Video: The answer is (D) (Wrong, Confidence 0.409)

▶ **Iteration 2**

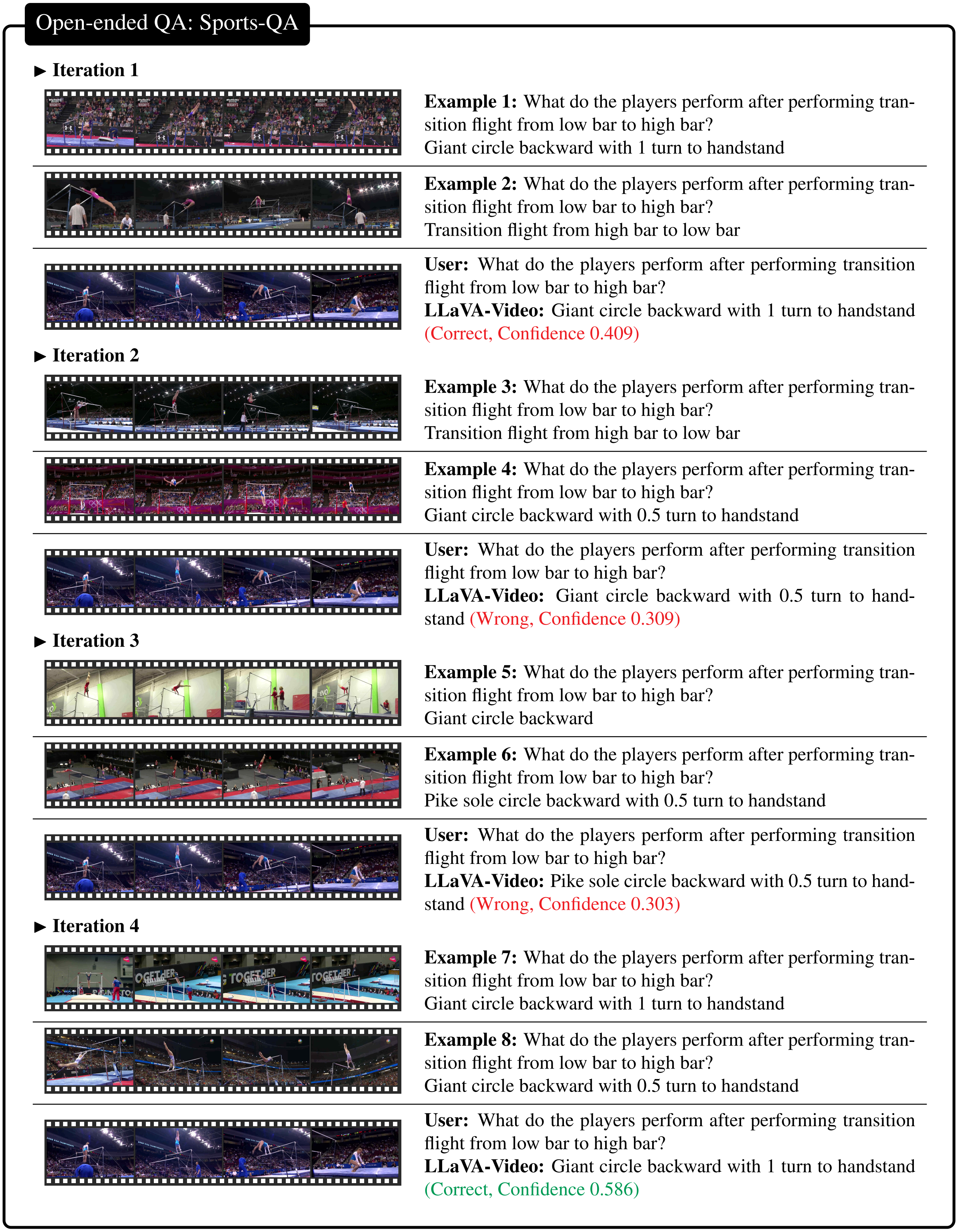
Example 3: What action is the animal doing in the video? Answer with the option's letter from the given choices directly. Options: (A) Chasing (B) Rubbing its head (C) Moving (D) Dying (E) Doing somersault
The answer is (C)

Example 4: What action is the animal doing in the video? Answer with the option's letter from the given choices directly. Options: (A) Puffing its throat (B) Hatching (C) Doing A Back Kick (D) Attending (E) Lying Down
The answer is (D)



User: What action is the animal doing in the video? Answer with the option's letter from the given choices directly. Options: (A) Trapped (B) Attending (C) Walking On Water (D) Sharing Food (E) Showing Affection
LLaVA-Video: The answer is (B) (Correct, Confidence 0.886)

Figure 5. Qualitative result on the Animal Kingdom dataset.



Open-ended QA: Sports-QA

► **Iteration 1**

Example 1: What do the players perform after performing transition flight from low bar to high bar?
Giant circle backward with 1 turn to handstand

Example 2: What do the players perform after performing transition flight from low bar to high bar?
Transition flight from high bar to low bar

User: What do the players perform after performing transition flight from low bar to high bar?
LLaVA-Video: Giant circle backward with 1 turn to handstand (Correct, Confidence 0.409)

► **Iteration 2**

Example 3: What do the players perform after performing transition flight from low bar to high bar?
Transition flight from high bar to low bar

Example 4: What do the players perform after performing transition flight from low bar to high bar?
Giant circle backward with 0.5 turn to handstand

User: What do the players perform after performing transition flight from low bar to high bar?
LLaVA-Video: Giant circle backward with 0.5 turn to handstand (Wrong, Confidence 0.309)

► **Iteration 3**

Example 5: What do the players perform after performing transition flight from low bar to high bar?
Giant circle backward

Example 6: What do the players perform after performing transition flight from low bar to high bar?
Pike sole circle backward with 0.5 turn to handstand

User: What do the players perform after performing transition flight from low bar to high bar?
LLaVA-Video: Pike sole circle backward with 0.5 turn to handstand (Wrong, Confidence 0.303)

► **Iteration 4**

Example 7: What do the players perform after performing transition flight from low bar to high bar?
Giant circle backward with 1 turn to handstand

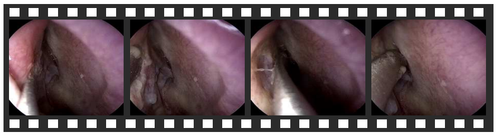
Example 8: What do the players perform after performing transition flight from low bar to high bar?
Giant circle backward with 0.5 turn to handstand

User: What do the players perform after performing transition flight from low bar to high bar?
LLaVA-Video: Giant circle backward with 1 turn to handstand (Correct, Confidence 0.586)

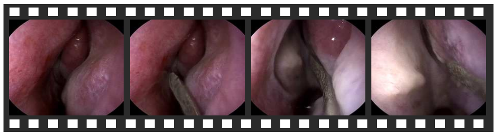
Figure 6. Qualitative result on the Sports-QA dataset.

Open-ended QA: PitVQA

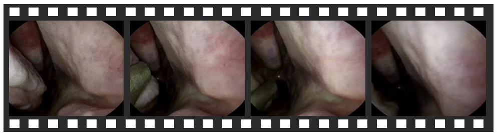
► **Iteration 1**



Example 1: Where is the surgical instrument freer elevator tip located in the middle of the video?
Top-left

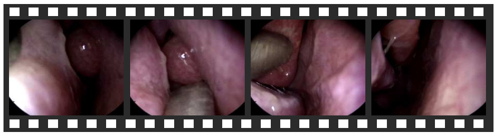


Example 2: Where is the surgical instrument freer elevator tip located in the middle of the video?
Centre

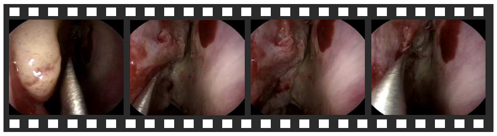


User: Where is the surgical instrument freer elevator tip located in the middle of the video?
LLaVA-Video: Bottom-right (Wrong, Confidence 0.167)

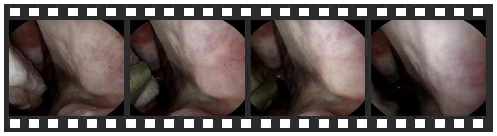
► **Iteration 2**



Example 3: Where is the surgical instrument freer elevator tip located in the middle of the video?
Top-left

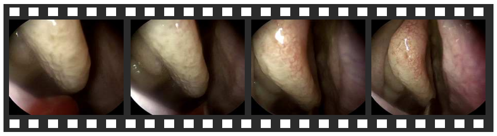


Example 4: Where is the surgical instrument freer elevator tip located in the middle of the video?
Centre

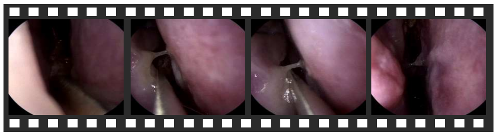


User: Where is the surgical instrument freer elevator tip located in the middle of the video?
LLaVA-Video: Top (Wrong, Confidence 0.204)

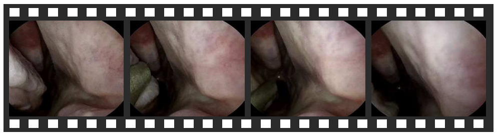
► **Iteration 3**



Example 5: Where is the surgical instrument freer elevator tip located in the middle of the video?
Bottom-left



Example 6: Where is the surgical instrument freer elevator tip located in the middle of the video?
Bottom-left



User: Where is the surgical instrument freer elevator tip located in the middle of the video?
LLaVA-Video: Bottom-left (Correct, Confidence 0.809)

Figure 7. Qualitative result on the PitVQA dataset.

Video Classification: UCF-Crime

► **Iteration 1**

Example 1: Classify the following video into one of the following categories: {14 categories}
Stealing

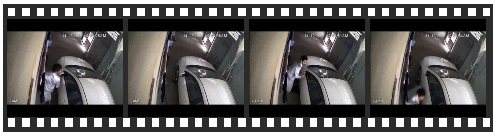
Example 2: Classify the following video into one of the following categories: {14 categories}
Shooting

User: Classify the following video into one of the following categories: {14 categories}
LLaVA-Video: Stealing (Correct, Confidence 0.484)

► **Iteration 2**

Example 3: Classify the following video into one of the following categories: {14 categories}
Stealing

Example 4: Classify the following video into one of the following categories: {14 categories}
Vandalism



User: Classify the following video into one of the following categories: {14 categories}
LLaVA-Video: Stealing (Correct, Confidence 0.591)

Figure 8. Qualitative result on the UCF-Crime dataset.

Video Classification: Drive&Act

▶ Iteration 1

Example 1: Classify the following video into one of the following categories: {34 categories}
Putting on sunglasses

Example 2: Classify the following video into one of the following categories: {34 categories}
Taking off sunglasses

User: Classify the following video into one of the following categories: {34 categories}
LLaVA-Video: Taking off sunglasses
(Wrong, Confidence 0.395)

▶ Iteration 2

Example 3: Classify the following video into one of the following categories: {34 categories}
Eating

Example 4: Classify the following video into one of the following categories: {34 categories}
Fetching an object

User: Classify the following video into one of the following categories: {34 categories}
LLaVA-Video: Eating (Correct, Confidence 0.354)

▶ Iteration 3

Example 5: Classify the following video into one of the following categories: {34 categories}
Eating

Example 6: Classify the following video into one of the following categories: {34 categories}
Eating

User: Classify the following video into one of the following categories: {34 categories}
LLaVA-Video: Eating (Correct, Confidence 0.818)

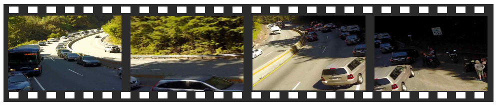
Figure 9. Qualitative result on the Drive&Act dataset.

Video Captioning: CapERA

► **Iteration 1**

Example 1: Provide a concise depiction of this video.
Two police cars were chasing a white car down a busy street while the white car was walking on the sidewalk.

Example 2: Provide a concise depiction of this video.
The winding mountain road is crowded with cars and surrounded by trees.



User: Provide a concise depiction of this video.
LLaVA-Video: The traffic is moving slowly on the road. (Confidence 0.114)

► **Iteration 2**

Example 3: Provide a concise depiction of this video.
Three cars collided on a treelined road and policemen explored the accident.

Example 4: Provide a concise depiction of this video.
The road is crowded with cars and surrounded by small buildings and trees.

User: Provide a concise depiction of this video.
LLaVA-Video: Cars are driving on a road surrounded by trees and greenery. (Confidence 0.090)

► **Iteration 3**

Example 5: Provide a concise depiction of this video.
The winding mountain road is crowded with cars and surrounded by trees.

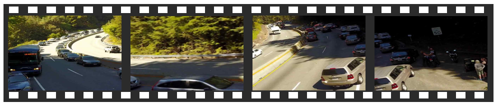
Example 6: Provide a concise depiction of this video.
The winding mountain road is crowded with cars and surrounded by trees.

User: Provide a concise depiction of this video.
LLaVA-Video: The winding mountain road is crowded with cars and surrounded by trees. (Confidence 0.271)

► **Iteration 4**

Example 7: Provide a concise depiction of this video.
Crowded road with cars.

Example 8: Provide a concise depiction of this video.
Winding mountain road crowded with cars and surrounded by trees.



User: Provide a concise depiction of this video.
LLaVA-Video: Traffic on a highway surrounded by trees. (Confidence 0.179)

Figure 10. Qualitative result on the CapERA dataset.