VERA: Explainable Video Anomaly Detection via Verbalized Learning of Vision-Language Models

Muchao Ye[1] Weiyang Liu[2] Pan He[3]
[1]The University of Iowa [2]Max Planck Institute for Intelligent Systems, Tübingen [3]Auburn University
[1]muye@uiowa.edu [2]weiyang.liu@tuebingen.mpg.de [3]pan.he@auburn.edu

Supplementary Material

In this supplementary material, we first include more details on training in VERA (Sec. A) and additional experimental results (Sec. B). To specify:

- In Sec. A, we provide the pseudocodes and details on the initialization, the learner prompt template, and the optimizer prompt template for the training process in Sec. A.1. After that, we discuss the optimization process of the learned questions by the optimizer in Sec. A.2.
- In Sec. B, we first include comparison results with the state-of-the-art methods on XD-Violence measured by AP in Sec. B.1. We also discuss other good properties of VERA, including the good generalizability of the learned questions for different scenarios and the insensitivity of VERA regarding hyperparameters in Sec. B.2 and Sec. B.4, respectively. Finally, we include additional case studies with normal and abnormal videos in Sec. B.5.

We also include a further discussion on the limitations of VERA for future research exploration in Sec. C.

A. Training in VERA

A.1. Algorithm

We show the complete iterative training process of VERA in pseudocodes in Algorithm 1. It is an iterative process of using the learner to output binary prediction for each sample in a mini-batch and asking the optimizer to update the guiding questions after collecting the batched data. Meanwhile, we have a small validation set (10% samples randomly drawn from the original training set) for deciding the $\mathbf{Q}^*$ used for testing. We want to further detail on certain elements in Algorithm 1 as follows.

Initial $\mathbf{Q}_0$. The initial guiding questions $\mathbf{Q}_0$ are "*1. Is there any suspicious person or object that looks unusual in this scene? 2. Is there any behavior that looks unusual in this scene?*". These two questions are manually written and inspired by previous VAD methods, which assume anomaly as something or somebody with unusual appearance or motions [13, 43]. This set of questions is also the "**manually written questions by human**" in Table 5, which is suboptimal in guiding frozen VLMs to detect anomalies. The key idea of training is to use VL to iteratively update $\mathbf{Q}$ given a suboptimal $\mathbf{Q}_0$.

Learner Prompt Template θ. We detail the design of θ as follows. As shown in Fig. 2, the learner prompt template θ includes four sections, *i.e*., Model Description, Prompt Questions, Input, and Output Formatting. To specify:

- Model Description: This section introduces the learning task, providing the learner with the necessary background knowledge to understand the objective. It clarifies what the learner is expected to predict based on the given visual input data.
- Prompt Questions: This section presents a general prompt to guide the learner's reasoning process. Specific prompts, denoted as $\mathbf{Q}_t$, will be inserted here to facilitate reasoning within a frozen VLM.
- Input: This section simply stores the visual tokens. When the VLM reads this, it will correlate the read text with the visual inputs.
- Output Formatting: The last section in θ mainly provides information on output formats to ensure that VLMs think through the given questions $\mathbf{Q}_t$ and output a prediction in a format easy for post-processing in computers.

Optimizer Prompt Template ψ. As shown in Fig. 2, the optimizer prompt template includes seven sections, *i.e*., Instruction, Inputs, Model Description, Current Prompt Questions, Model Predictions & Targets, and Optimization Instruction:

- Instruction: The prompt template begins with an introduction outlining the responsibilities of the optimizer, clearly stating that its primary task is to optimize the guiding questions provided.
- Inputs: This section is used to attach the batched visual data for the reference of the optimizer.
- Model Description: The learning task of the learner is reiterated here for the information of the optimizer.
- Current Prompt Questions: The guiding questions used by the learner in the current iteration are shown here for the reference of the optimizer.

Algorithm 1: Optimizing Guiding Questions in VAD by VERA during Training

Inputs: Training data pairs $D_{\text{train}} = \{(\tilde{V}^{(j)}, Y^{(j)})\}_{j=1}^{N}$, iteration number P, initial guiding questions $\mathbf{Q}_0$, learner f_{learner}, optimizer f_{opt}, learner prompt template θ, optimizer prompt template ψ, validation set $D_{\text{val}} = \{(\tilde{V}_{\text{val}}^{(j)}, Y_{\text{val}}^{(j)})\}_{j=1}^{\eta}$, period for validation μ, batch size n.

Output: Optimal guiding questions $\mathbf{Q}^*$.

Set iteration counter $t \leftarrow 1$;

Set $\mathbf{Q}^* \leftarrow \mathbf{Q}_0$, test $\mathbf{Q}_0$ on validation set D_{val} and compute its validation accuracy as Acc^*;

while $t \leq P$ **do**

 # Conduct the learning task with a mini-batch by the learner

Randomly sample a batch without repetition from D_{train} with a visual input batch $V_{\text{batch}} = [\tilde{V}_{\text{batch}}^{(1)}, \cdots, \tilde{V}_{\text{batch}}^{(n)}]$ and ground truths $Y_{\text{batch}} = [Y_{\text{batch}}^{(1)}, \cdots, Y_{\text{batch}}^{(n)}]$;

for $1 \leq j \leq n$ **do**

Obtain a prediction $\hat{Y}_{\text{batch}}^{(j)}$ for $\tilde{V}_{\text{batch}}^{(j)}$ from f_{learner} with prompt $(\theta, \mathbf{Q}_t)$ by Eq. (1) as $\hat{Y}_{\text{batch}}^{(j)} = f_{\text{learner}}^{(t)}(\tilde{V}_{\text{batch}}^{(j)})$;

end

 # Update the guiding questions with the batched data by the optimizer

Input the batched prediction $\hat{Y}_{\text{batch}} = [\hat{Y}_{\text{batch}}^{(1)}, \cdots, \hat{Y}_{\text{batch}}^{(n)}]$ with V_{batch} and Y_{batch} into the optimizer for obtaining a new set of guiding questions by Eq. (2) as: $\mathbf{Q}_{t+1} = f_{\text{opt}}^{(t)}(V_{\text{batch}}, \hat{Y}_{\text{batch}}, Y_{\text{batch}})$;

 # Compute the validation accuracy with the learned guiding questions periodically

$t \leftarrow t + 1$;

if $t \bmod \mu = 0$ **then**

Test $\mathbf{Q}_t$ on the validation set D_{val} and compute the validation accuracy Acc_t;

if $\text{Acc}_t > \text{Acc}^*$ **then**

Update $\mathbf{Q}^* \leftarrow \mathbf{Q}_t$;

Update $\text{ACC}^* \leftarrow \text{ACC}_t$;

end

end

end

Return $\mathbf{Q}^*$;

- Model Predictions & Targets: The batched numerical predictions and the ground truths are shown here for f_{opt}. These two inputs can tell the optimizer how well the learner does in the learning task on the mini-batch data.
- Optimization Instruction: The final section includes the instruction to ask the optimizer to think step by step with all the information above and output a new set of prompt questions with the required format.

A.2. Details for Iterative Update by the Optimizer

In training, we assess the quality of the learned guiding questions by the accuracy of the validation set. We show the validation accuracy from different questions $\mathbf{Q}_t$ obtained every 100 iterations (mini-batches) in Fig. 9. In the duration of up to 5000 iterations in training, the observed plot in Fig. 9 contains three oscillations, each consisting of an increase in validation accuracy followed by a decrease. The increase represents that the optimizer VLM gradually finds better questions for the binary classification learning task when it sees more batched data, which shows the optimizer can understand its responsibility well and find better questions effectively. Meanwhile, we note that verbal optimization may not always lead to an increase. This is probably because the optimization is completely verbalized, and the VLM will have an inertial thinking behavior like humans, which gets the optimizer stuck in the wrong direction and makes it continue the optimization in a direction that is not beneficial. As a result, this causes the validation accuracy to decrease sometimes. Despite that, because of the guidance provided by the optimizer prompt template ψ, the optimizer can overcome its pitfalls in thinking and find good guiding questions in a new direction, which leads to an increase in validation accuracy afterward. This is an interesting phenomenon due to the distinction between verbal learning and traditional numerical optimization algorithms, and it will be a promising future direction to reduce the time in overcoming pitfalls in thinking for VLMs during VL.

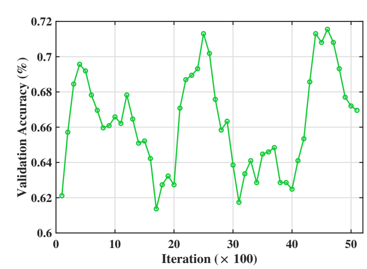

Figure 9. The validation accuracy given different learned guiding questions from each iteration. The graph is smoothed with moving average (window size 5) for better readability.

In addition, w.l.og., we take learned questions from the 100th iteration to the 700th iteration (which are within the first

Learned Questions after 100 Iterations	Learned Questions after 300 Iterations	Learned Questions after 500 Iterations	Learned Questions after 700 Iterations
1. Are there any vehicles in the video frames that are not typically found on a street, such as a motorcycle in a store or a backpack in a parking lot?	1. Are there any people in the video frame who are not engaging in typical activities, such as loitering or not interacting with the surroundings?	1. Are there any individuals in the video frames who are carrying unusual or unexpected items?	1. Are there any frames where the presence of a specific object (e.g., unusual item, unexpected tool) is not typical for the scene?
2. Are there any people in the video frames who are interacting with objects in a manner that is not typical for a street setting, such as handling a large object in a store or carrying a bag in a parking lot?	2. Are there any objects in the video frame that are not consistent with the typical environment, such as unusual equipment or structures?	2. Are there any objects in the video frames that are not typically found in that environment and are being used in an unusual or unexpected way?	2. Are there any frames where the arrangement of people or objects is unusual or unsafe?
3. Are there any people in the video frames who are standing or sitting in a manner that seems suspicious or out of place, such as sitting on the ground in a store or standing in a parking lot?	3. Are there any sudden or unusual movements in the video frame that are not typical for the environment, such as equipment malfunctions or unexpected actions?	3. Are there any individuals in the video frames who are interacting with objects in an unusual or unexpected way?	3. Are there any frames where an individual is carrying an object that is not typical for the scene?
4. Are there any people in the video frames who are wearing clothing that is not typical for a street setting, such as wearing a uniform in a parking lot or wearing protective gear in a store?	4. Are there any people in the video frame who are not following the normal flow of activities, such as standing in unusual places or engaging in suspicious behaviors?	4. Are there any individuals in the video frames who are performing actions that are not typical for that environment?	4. Are there any frames where an individual is interacting with an object in an unusual manner?
5. Are there any people in the video frames who are in unusual positions or postures, such as sitting on the ground or standing in a way that is not typical for a store setting?	5. Are there any objects in the video frame that are not interacting with the environment in a typical manner, such as being placed in unusual locations or not being used for their intended purposes?	5. Are there any unusual or unexpected activities occurring in the video frames, such as interactions between objects or individuals that are not typical for that environment?	5. Are there any frames where the overall environment or setting is not consistent with normal conditions?

Figure 10. We take the guiding questions $\mathbf{Q}$ learned from the 100th iteration to the 700th iteration for illustration purpose. During the updating process, the optimizer gradually concretizes anomaly patterns that can be applied to different scenarios in a concise expression.

epoch) for illustration to show the process of updating $\mathbf{Q}$ by the optimizer in Fig. 10. First, as the optimizer sees more videos, it tries to make the questions focus on a more general setting. For example, the questions in the 100th iteration focus on "street" and "store" scenes. After more iterations, the questions become more generalizable for a general environment and focus on the elements that cause anomalies. Additionally, the anomalous pattern descriptions become more diverse as the optimization continues. To illustrate, in the beginning, the questions mostly pay attention to the humans, objects, and their interaction. In later iterations, the optimizers gradually summarize some previous questions into one and raise questions considering the overall environment (Q5 from the 700th iteration). Therefore, the VL framework proposed in this paper is effective in finding a diverse set of guiding questions for VAD that apply to general cases, which can elicit the reasoning of a frozen VLM in VAD.

B. Additional Experiments and Results

B.1. Comparison to the State-of-the-art Methods on XD-Violence Measured by AP

The comparison results regrading average precision (AP), *i.e.*, the area under the frame-level precision-recall curve, on XD-Violence are shown in Table 10. Compared to AUC, AP focuses on measuring the ability to identify the positive class (anomaly), while AUC measures how well a method separates anomaly and normalcy in general. We provide the analysis of the results as follows.

Method	AP
Non-Explainable VAD Methods	
Wu et al.† [42]	78.64
OVVAD [43]	66.53
S3R† [41]	80.26
RTFM† [36]	77.81
MSL† [19]	78.58
MGFN† [6]	80.11
CLIP-TSA† [15]	**82.19**
Explainable VAD Methods	
Holmes-VAD† [55]	**84.96**
LAVAD [52]	62.01
ZS CLIP [52]	17.83
ZS IMAGEBIND-I [52]	27.25
ZS IMAGEBIND-V [52]	25.36
LLAVA-1.5 [20]	50.26
VERA	70.54

Table 10. AP (%) on XD-Violence. † indicates VAD methods are trained on entire training frames. No IT is used for Holmes-VAD.

Firstly, under such a distinct property of AP, as pointed out by [43], methods trained on the whole training set and utilizing all frames will enjoy advantages when measuring VAD performance by AP. As a result, CLIP-TSA and Holmes-VAD, two methods using the whole training frames, attain the highest AP in the category of non-explainable and explainable VAD, respectively. We acknowledge there is a gap between VERA and these two methods under AP on XD-Violence, which is understandable because they use the whole training frames to improve the ability to find anomalies of classifiers. To illustrate, in training VERA only samples 8 frames for each video and only uses 0.19% total frames (31,632 out of 16,378,527) for training on XD-Violence. Thus, our training is dramatically light compared to the methods like CLIP-TSA and Holmes-VAD in Table 10. With fewer frames used for training, VERA unavoidably achieve lower AP (which only considers positive cases) compared to those that have more, for it relies on fewer training data. In addition, we want to point out that judging the VAD performance solely by AP on XD-Violence can be biased. This is because the ratio of positive frames in XD-Violence (23.07%) in test videos is overly higher than other datasets like UCF-Crime (7.92%), which is unrealistic because the anomaly is sparse in the real world [32]. Given that, only focusing on the comparison in AP on XD-Violence would amplify the bias in VAD performance evaluation, and we recommend taking into consideration other factors like training costs and the comprehensive ability of distinguishing anomaly and normality by the methods in evaluation.

Secondly, among the methods (OVVAD, LAVAD, ZS CLIP, ZS IMAGEBIND, and LLAVA-1.5) that does not use full frames for training, VERA achieves the best AP in this fair comparison, surpassing the second best method in the Explainable VAD category (LAVAD) over 8.53%, which showcases the effectiveness of using learned guiding question to prompt frozen VLMs for VAD.

To conclude, it is unfair to only judge VAD performance by AP on XD-Violence without considering the training costs and the relatively imbalanced frame distribution in test videos. Considering all factors into consideration, VERA is a favorable method used for VAD in detecting anomalies.

B.2. Discussion on Generalizability of Used Questions

During the optimization of $\mathbf{Q}$, because of the randomness involved in this process, the optimizer may output certain guiding questions that only focus on one specific surrounding. We find an interesting phenomenon on VLMs in VAD that guiding questions related to a specific scenario yield inferior VAD performance compared to the general questions in both general cases and specific cases.

To illustrate, we take two sets of specific questions obtained on UCF-Crime for analysis. The first example is a set of guiding questions $\mathbf{Q}_{\text{traffic}}$ that only ask the VLM to consider anomalies related to the traffic as follows:

1. *Are there any vehicles or people violating traffic rules?*
2. *Are there any accidents or near-accidents occurring?*
3. *Are there any objects or people obstructing the normal flow of traffic?*
4. *Are there any unusual or unexpected behaviors from pedestrians or drivers?*
5. *Are there any emergency vehicles or personnel present?*

The second example is another set of guiding questions $\mathbf{Q}_{\text{store}}$ that only ask the VLM to identify anomalies in a store setting, which includes questions like:

1. *Are there any individuals loitering or behaving suspiciously inside the store?*
2. *Is there any unusual activity inside the store, such as tampering with items or attempting to enter restricted areas?*
3. *Are there any signs of forced entry or damage to the store's entrance?*
4. *Are there any individuals present who seem to be watching or waiting for something specific inside the store?*
5. *Are there any interactions between individuals inside the store that appear suspicious or out of the ordinary?*

Thus, $\mathbf{Q}_{\text{traffic}}$ and $\mathbf{Q}_{\text{store}}$ focuses on the specific anomalies of traffic accidents and shoplifting, respectively, while the $\mathbf{Q}^*$ that we find focuses on general cases and includes the following questions:

1. *Are there any people in the video who are not in their typical positions or engaging in activities that are not consistent with their usual behavior?*
2. *Are there any vehicles in the video that are not in their typical positions or being used in a way that is not consistent with their usual function?*
3. *Are there any objects in the video that are not in their typical positions or being used in a way that is not consistent with their usual function?*
4. *Is there any visible damage or unusual movement in the video that indicates an anomaly?*
5. *Are there any unusual sounds or noises in the video that suggest an anomaly?*

The comparison results of $\mathbf{Q}^*$, $\mathbf{Q}_{\text{traffic}}$, and $\mathbf{Q}_{\text{store}}$ in detecting anomalies in general cases (all testing videos on UCF-Crime), traffic scenes (testing videos from the Traffic Accident category on UCF-Crime), and the store scenes (testing videos from the Shoplifting category on UCF-Crime) are shown in Table 11. It indicates that $\mathbf{Q}^*$ performs the best in both general cases and two specific cases like in traffic and store scenes. This is because the overly specific definition of anomalies like $\mathbf{Q}_{\text{traffic}}$ and $\mathbf{Q}_{\text{store}}$ makes it harder for a VLM to classify one clip into an anomaly and leads to more false negatives in its prediction given those specific questions, which degrades the performance. Therefore, we recommend using general questions like the ones shown in $\mathbf{Q}^*$ in frozen VLMs for VAD.

Questions	Scenario		
	All	Traffic	Store
$\mathbf{Q}^*$	**86.55**	**70.43**	**72.58**
$\mathbf{Q}_{\text{traffic}}$	82.59	67.53	/
$\mathbf{Q}_{\text{store}}$	76.67	/	44.84

Table 11. General guiding questions outperform specific ones measured by AUC (%) on UCF-Crime. Specific questions are not tested on other specific scenarios, which is indicated by a slash (/).

B.3. Hyperparameters in Training

Batch Size and Sampled Frame Number. Key hyperparameters that need to be set in training are the batch size n and the number of sampled frames S for each video $V^{(j)}$ in the VL framework. They are correlated because they determine the total number of frames for the optimizer to skim and provide feedback as $S \cdot n$. Considering memory constraints when implementing VLMs on GPUs, we set $S \cdot n = 16$ in training. We further explore the trade-off between S and n given the constraints for input frames to decide S and n. The results are shown in Table 12. If the batch size n is 1 with $S = 16$, the learned questions cannot be generalized due to the limited video sample in the batch which leads to a suboptimal AUC, and it takes longer to train for VERA. Meanwhile, if we set n as large numbers like 4 or 8 (with $S = 4$ or $S = 2$), the learned questions are suboptimal too because relatively few sampled frames generally lack the temporality for the optimizer to look into the details and conceive good questions. Thus, setting n to 2 and S to 8 is in default in this paper, which strikes the balance between training efficiency and effectiveness.

Batch Size	Sampled Frames	AUC (%)
n = 1	S = 16	81.53
n = 2	S = 8	**86.55**
n = 4	S = 4	83.19
n = 8	S = 2	79.91

Table 12. The choice of batch size and sampling frames affects the effectiveness of the learned guiding questions in VAD. The results are obtained by InternVL2-8B as VERA's backbone.

B.4. Hyperparameters in Inference and Sensitivity Test

Hyperparameters in Inference During inference, in Step 1, following [52], the interval between each segment center d is 16 frames. In Step 2, we use ImageBind [10] as the feature extractor in computing segment similarity as [52] does, and the number of retrieved segments K depends on the total number of segments h in each test video V. Setting K to $(0.1 \cdot h)$ to $(0.15 \cdot h)$ is generally good. We set K to $(0.1 \cdot h)$ for UCF-Crime and to $(0.15 \cdot h)$ for XD-Violence. The temperature τ in the Softmax function is set to 10 for both datasets in Eq. (4). In Step 3, due to the properties of datasets, we set the filter size ω of $G(p)$ to 15 and σ_1 to 10 for UCF-Crime, while setting ω to 30 and σ_1 to 30 for XD-Violence. For position weighting, we set $c = \text{floor}(F/2)$ and $\sigma_2 = \text{floor}(F/2)$ for both datasets to make sure the position weight covers the whole video sequence.

W.l.o.g, we test the sensitivity of the VAD performance of VERA regarding hyperparameters on UCF-Crime.

Sensitivity Test for K. As shown in Table 13, as the number of retrieved segments increases from 0 to $0.15 \cdot h$, the AUC gradually increases from to 85.21% to 86.61%. Meanwhile, if we randomly select $0.1 \cdot h$ segments for retrieval, the AUC is even lower than the performance without retrieval. Thus, using Eq. (4) for retrieval is necessary. Meanwhile, having a large K greater than $0.15 \cdot h$ will introduce some noise in Eq. (4) and downgrade the AUC slightly. Thus, selecting $0.1 \cdot h$ or $0.15 \cdot h$ for K is generally good choice.

Ratio (%)	0	5	10	15	20	25
AUC (%)	85.21	86.48	86.55	86.61	86.42	86.19

Table 13. Influence of the number of retrieved segments on AUC. The AUC of not using retrieval (Ratio = 0%) and randomly selecting 10% segments for Eq. (4) is 85.21% and 84.55%, respectively.

Sensitivity Test for ω. The filter size decides how many local segments are incorporated for the current segment for Gaussian smoothing. From Table 14, we find that AUC converges when the filter size increases to 15. Meanwhile, the VAD performance measured AUC is insensitive to ω and does not fluctuate much. Thus, we can set the filter size with a medium number like 15.

ω	5	10	15	20	25
AUC (%)	86.25	86.43	86.55	86.61	86.60

Table 14. Influence of filter size ω in Gaussian Smoothing on AUC.

Sensitivity Test for σ_1. The AUC performance is also robust on the choice of σ_1. As, shown in Fig. 15, when we set σ_1 greater than 1, the AUC generally remains around 86.50%, which again shows the robustness of the design of anomaly scoring in VERA. We can set σ_1 as 10 for VERA.

σ_1	1	5	10	15	20
AUC (%)	86.17	86.49	86.55	86.49	86.54

Table 15. Influence of σ_1 in Gaussian Smoothing on AUC.

Sensitivity Test for τ. The temperature hyperparameter τ in Eq. (4) controls the entropy of the distribution obtained from the Softmax function while preserving the rank of each element. As demonstrated in Table 16, when τ is a small number like 10e-8 that is close to 0, the distributions tend to become a trivial distribution with all mass concentrated on the highest-probability class (corresponding to the segment itself), and the result is the same as the one by not using retrieval. As we gradually increase τ to a reasonably large number (from 0.01 to 1), the AUC value converges around 86.55% with no obvious fluctuation, again proving the robustness of anomaly scoring in VERA regarding hyperparameter selection. Note that when τ approaches $+\infty$, the distribution tends to become a uniform distribution, which yields an AUC of 86.59%. From the discussion above, we can generally choose τ to be an number in [0.01, 1] in implementation.

τ	10e-8	0.01	0.1	1	$+\infty$
AUC (%)	85.21	86.31	86.55	86.58	86.59

Table 16. Influence of τ in Eq. (4) on AUC.

Sensitivity Test for σ_2. From Table 17, we find that setting $\sigma_2 = 0.5F$ encodes the position information best in the anomaly score. A drop is noticeable if we choose σ_2 less than $0.5F$ for it will not cover the whole sequence, which is reasonable, while choosing a σ_2 great than $0.5F$ does not change much. Thus, based on the physical meaning of σ_2, which controls the width of the distribution, we should make σ_2 equal to $0.5F$ in anomaly scoring.

σ_2	w/o Weighting	0.25	0.5	0.75
AUC (%)	85.48	85.43	86.55	86.27

Table 17. Influence of σ_2 in Position Weighting on AUC.

B.5. Additional Qualitative Results & Case Studies

W.l.o.g., we take one normal video ("Normal_Videos_018_x264") and another abnormal video ("RoadAccidents127_x264") from the UCF-Crime dataset to demonstrate the explanations provided by a frozen VLM (InternVL2-8B) achieved by using the learned guiding questions $\mathbf{Q}^*$.

First, in Fig. 11 we showcase the explanation of anomaly scoring by VERA regarding a normal video "Normal_Videos_018_x264" in UCF-Crime, which is taken in an airport hallway where no anomaly happens. For this video, VERA assigns a 0 score to each frame. As shown in Fig. 11, for the selected scenes in this video, VERA explains that this is because there are no events that conform to the anomaly descriptions in $\mathbf{Q}^*$. Such explanations are consistent with the recording and again manifest the effectiveness of eliciting the reasoning ability in a frozen VLM for VAD by

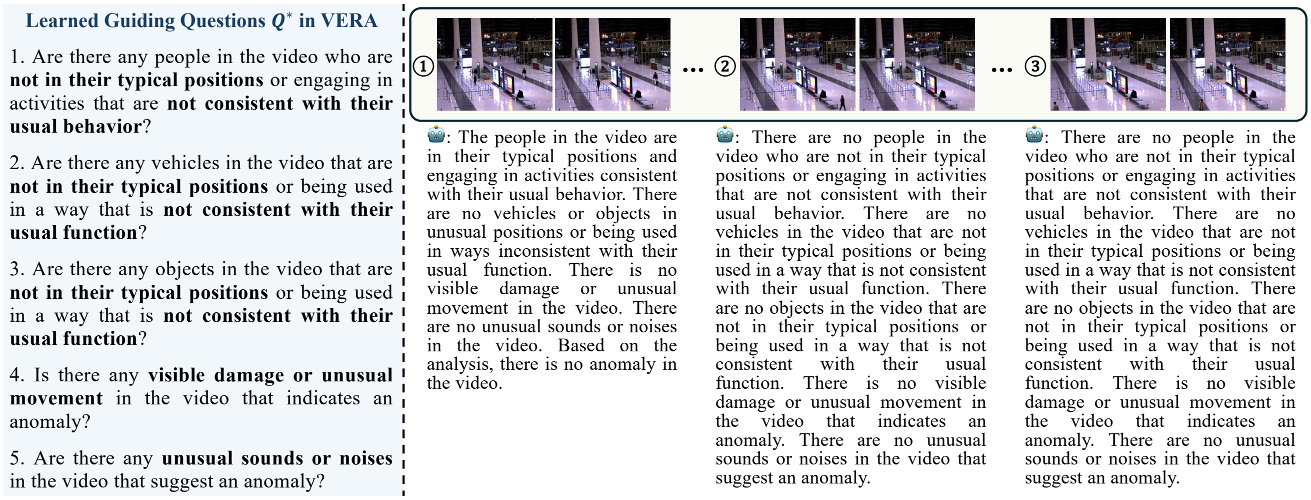

Figure 11. Given the normal video "Normal_Videos_018_x264", the frozen VLM (InternVL2-8B) can conclude that no anomaly happens in the video under the guidance of $\mathbf{Q}^*$, which is aligned with the ground truth. Since the anomaly scores for all scenes are zeros by VERA, we do not show the complete anomaly scores with an additional figure.

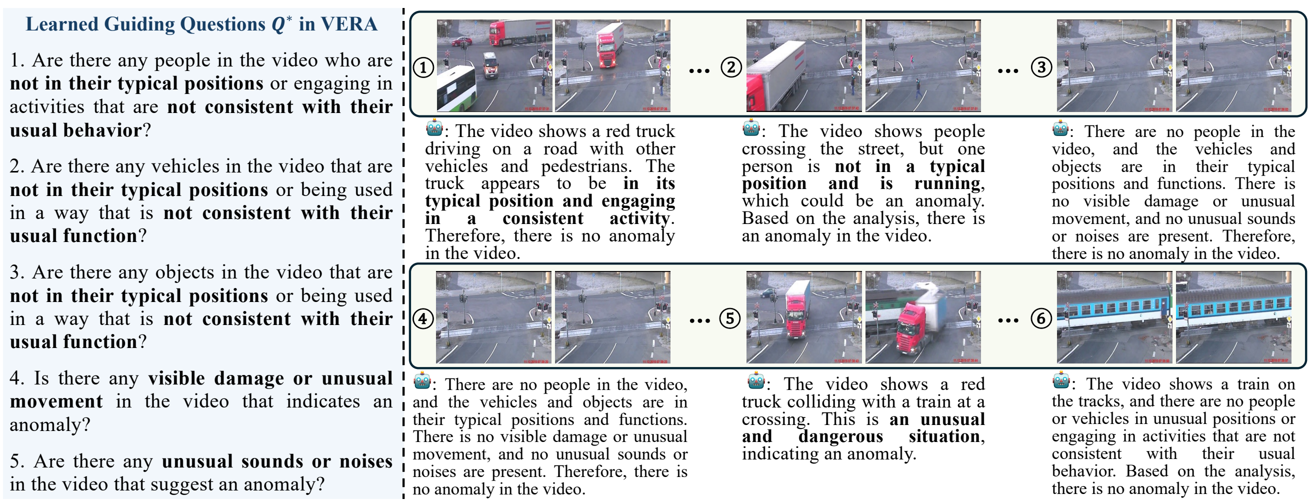

Figure 12. Given the abnormal video "RoadAccidents127_x264", the frozen VLM (InternVL2-8B) can generate reasonable explanations aligned with the semantic change observed in each scene under the guidance of $\mathbf{Q}^*$. The complete anomaly scores are shown in Fig. 13.

using learned guiding questions. Note that we do not have an additional figure illustrating the anomaly score dynamic for this video because all scenes are assigned 0 scores by VERA. Next, we select 6 representative scenes in the abnormal video ("RoadAccidents127_x264") and show the corresponding explanation provided by the frozen VLM in Fig. 12. The main anomaly that happens in this video is a traffic accident where a truck crashes into a train from Frame 2160 to Frame 2299, which corresponds to the 5th scene in Fig. 12. In particular, the figure shows that the learned question "Is there any visible damage or unusual movement in the video that indicates an anomaly?" in $\mathbf{Q}^*$ makes the frozen VLM find a good way to express what it sees in the 5th scene and understand this is an anomaly because the crash is unusual and dangerous. The other scenes are also well explained by the frozen VLM under $\mathbf{Q}^*$. Thus, this again verifies that the learned guiding questions can successfully trigger reasonable explanations in the adopted frozen VLM for VAD.

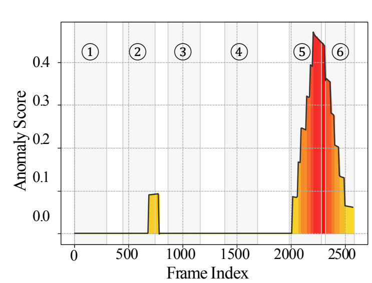

Figure 13. Anomaly scores generated by VERA (with InternVL2-8B) in "RoadAccidents127_x264" from UCF-Crime.

Meanwhile, we also include the anomaly scores generated by VERA for the abnormal video in Fig. 13. Most frames are assigned to zero except the scenes when someone crosses the road at an unusual speed (the 2nd scene in Fig. 12) and the truck-train crash happens (the 5th scene in Fig. 12). This fluctuation is aligned with the ground truth annotation and common sense about an anomaly, which shows that the anomaly scoring proposed in VERA is reasonable.

C. Further Discussion on Limitations

Like existing VLM-based VAD methods, VERA's performance relies heavily on the visual perception capabilities of VLMs. Most VLMs employ the CLIP vision encoder [30], which has limitations in capturing fine-grained visual details. This limitation can impair precise anomaly detection. If important visual features are missing during the visual encoding process, then it is unlikely for VERA to perform meaningful VL. Therefore, a fundamental challenge for VLM-based VAD is to ensure sufficient visual and temporal features are encoded. Having verified this capability, VERA can perform VL to extract crucial cues that guide video anomaly reasoning.