CoReVAD: A Contextual Reasoning Framework for Training-Free Video Anomaly Detection

Hyeongmuk Lim[1], Youngbum Hur[1*]

[1]Department of Industrial Engineering, Inha University, South Korea

*Corresponding author: youngbum.hur@inha.ac.kr

인하대학교 INHA UNIVERSITY

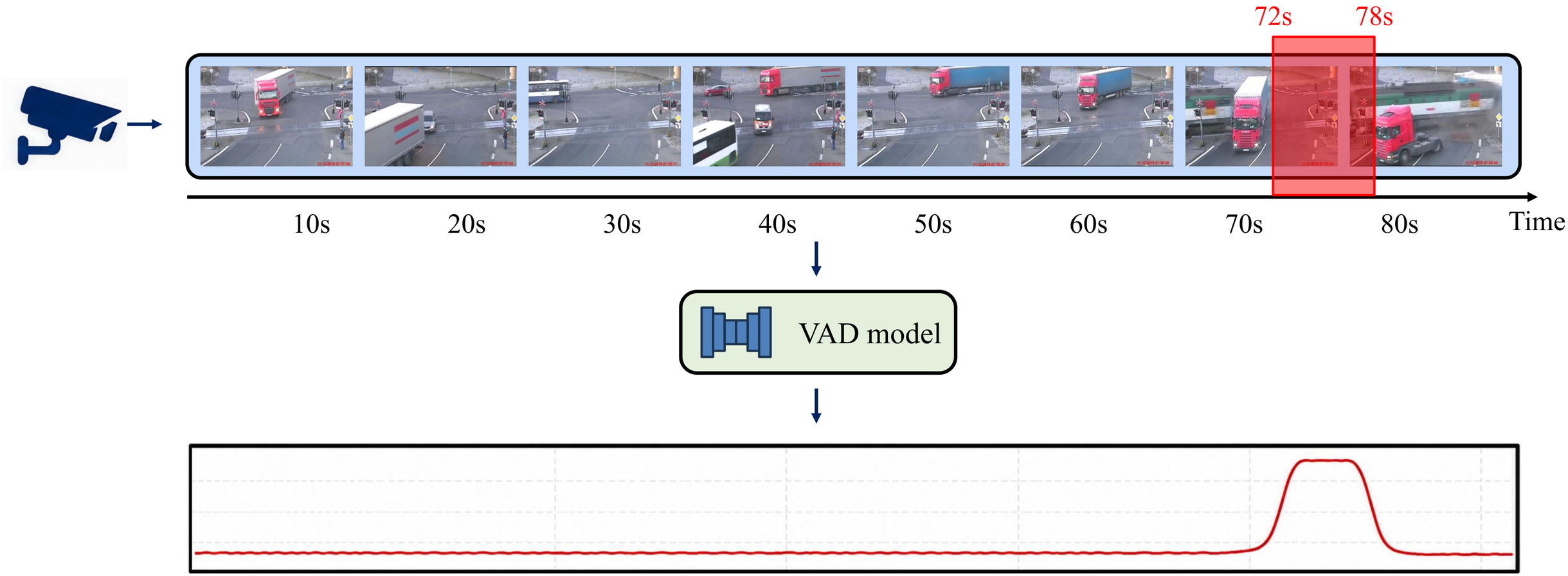
ORAIL Operations Research and Artificial Intelligence Lab

ICPR 2026 International Conference on Pattern Recognition LYON

Video Anomaly Detection (VAD)

Definition: Frame-Level Localization of Abnormal Events

- Video Anomaly Detection (VAD) is the task of identifying when abnormal events occur in surveillance video



Limitations of Conventional VAD Methods

1. Training Requirement & High Annotation Costs

- Large-scale annotations and task-specific training

2. Strong Domain Dependency

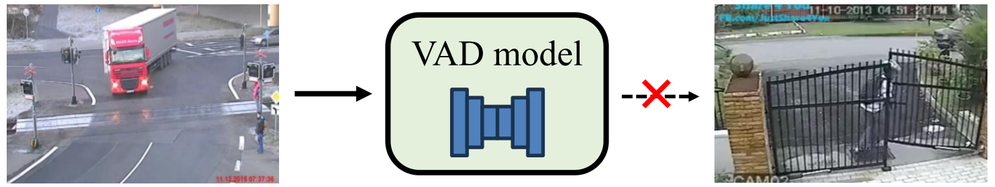
- Limited generalization to unseen environments

3. Limited Explainability

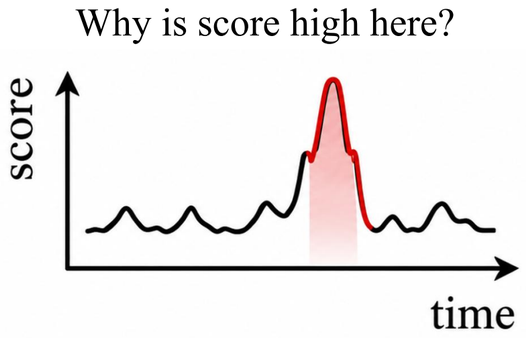
- Anomaly scores without interpretable reasoning

1) Need training data & annotation

2) Domain dependency



3) Lack of interpretability



Limitations of Conventional VAD Methods

Recent advances in VLMs and LLMs enable a new approach to prior VAD challenges

1. Training Requirement & High Annotation Costs

- Large-scale annotations and task-specific training

→

1. No Task-specific Training

- Use pre-trained VLMs/LLMs without VAD-specific training

2. Strong Domain Dependency

- Limited generalization to unseen environments

→

2. Cross-domain Applicability

- Better generalization to unseen scenes and environments

3. Limited Explainability

- Anomaly scores without interpretable reasoning

→

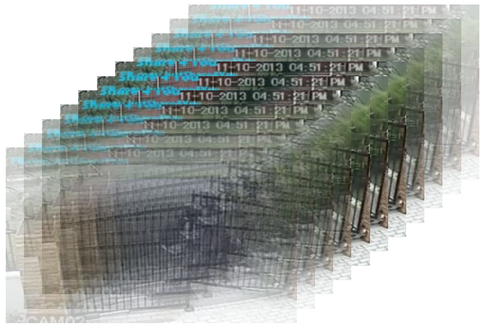
3. Interpretable Reasoning

- Provides human-readable explanations

ChatGPT deepseek Gemini InternVL3 Qwen3-VL

VLMs/LLMs Open a New Direction for VAD

1. Using External LLMs to Assist Frozen VLMs in Reasoning

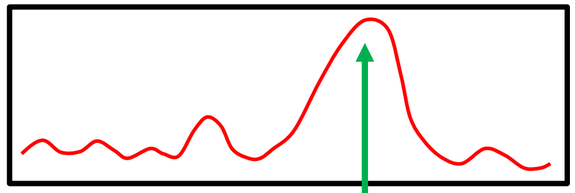


VLM

Describe

LLM

Reason



Reasoning

The person's forceful attempt to enter a residential area, ... heighten the suspicion.

2. Utilizing Instruction Tuning to Make VLMs Integrated VAD Systems

VLM

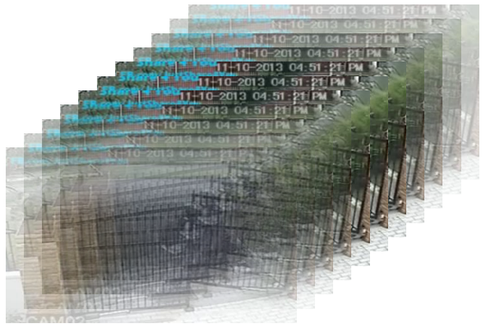
Describe & Reason

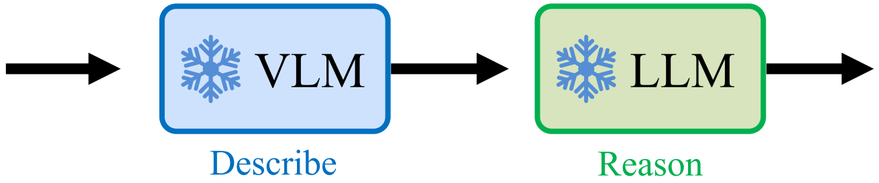
Reasoning

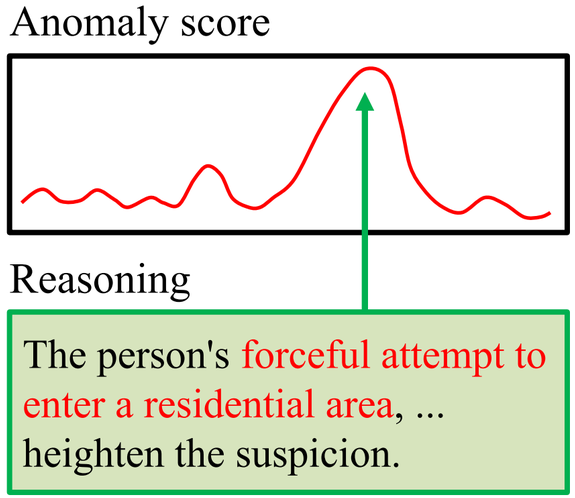
The person's forceful attempt to enter a residential area, ... heighten the suspicion.

VLMs/LLMs Open a New Direction for VAD

1. Using External LLMs to Assist Frozen VLMs in Reasoning







2. Utilizing Instruction Tuning to Make VLMs Integrated VAD Systems

Anomaly score
enter a residential area, ...
heighten the suspicion.

✓ **No Task-specific Training**

✓ **Cross-domain Applicability**

✓ **Interpretable Reasoning**

⚠ **High Computational Overhead**

VLMs/LLMs Open a New Direction for VAD

1. Using External LLMs to Assist Frozen VLMs in Reasoning

VLM

Describe

LLM

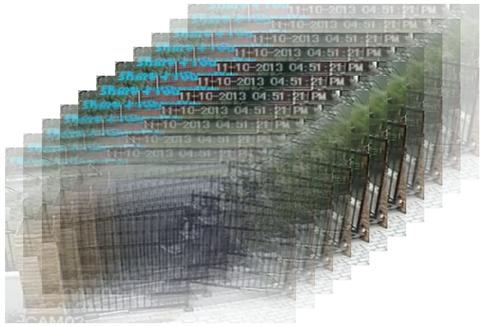
Reason

Anomaly score

Reasoning

The person's forceful attempt to enter a residential area, ... heighten the suspicion.

2. Utilizing Instruction Tuning to Make VLMs Integrated VAD Systems



VLM

Describe & Reason

Anomaly score

Reasoning

The person's forceful attempt to enter a residential area, ... heighten the suspicion.

VLMs/LLMs Open a New Direction for VAD

1. Using External LLMs to Assist Frozen VLMs in Reasoning

Anomaly score

heighten the suspicion.

✓ **Reduced Inference Overhead**

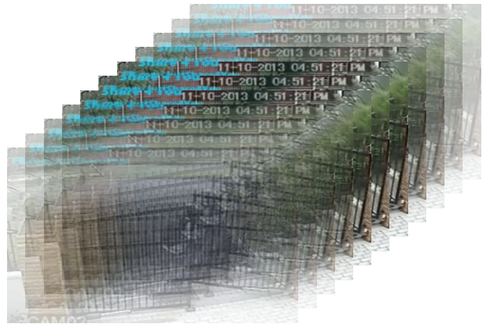
✓ **No External LLM**

✓ **Interpretable Reasoning**

⚠ **Additional Training & Data Required**

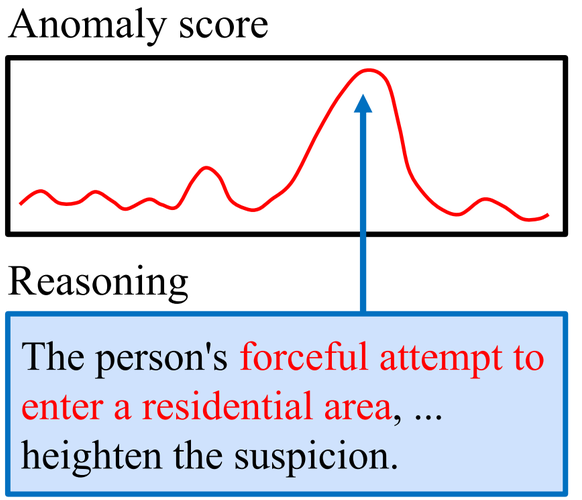
⚠ **Strong Domain Dependency**

2. Utilizing Instruction Tuning to Make VLMs Integrated VAD Systems



VLM

Describe & Reason



VLMs/LLMs Open a New Direction for VAD

1. Using External LLMs to Assist Frozen VLMs in Reasoning

VLM → LLM →

Anomaly score

Reasoning

Our Approach: CoReVAD

✓ No Task-specific Training ✓ Cross-domain Applicability ✓ Interpretable Reasoning

✓ No External LLM Resolves: high inference cost + additional training/data requirement

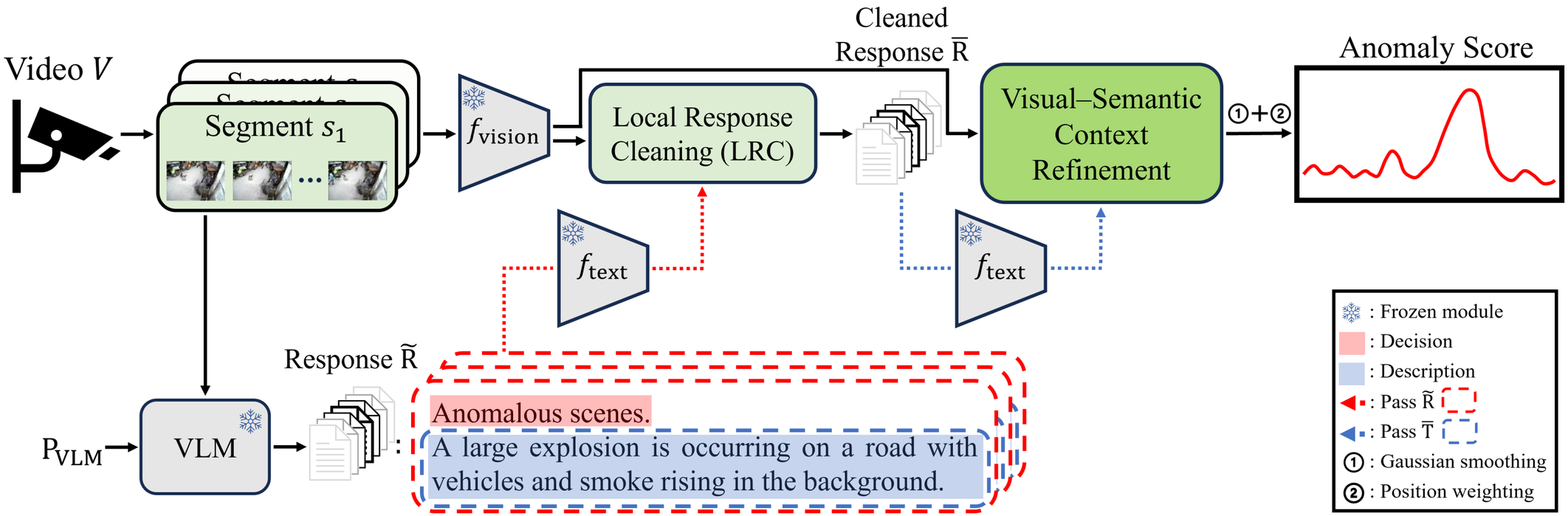
VLM

Describe & Reason

Reasoning

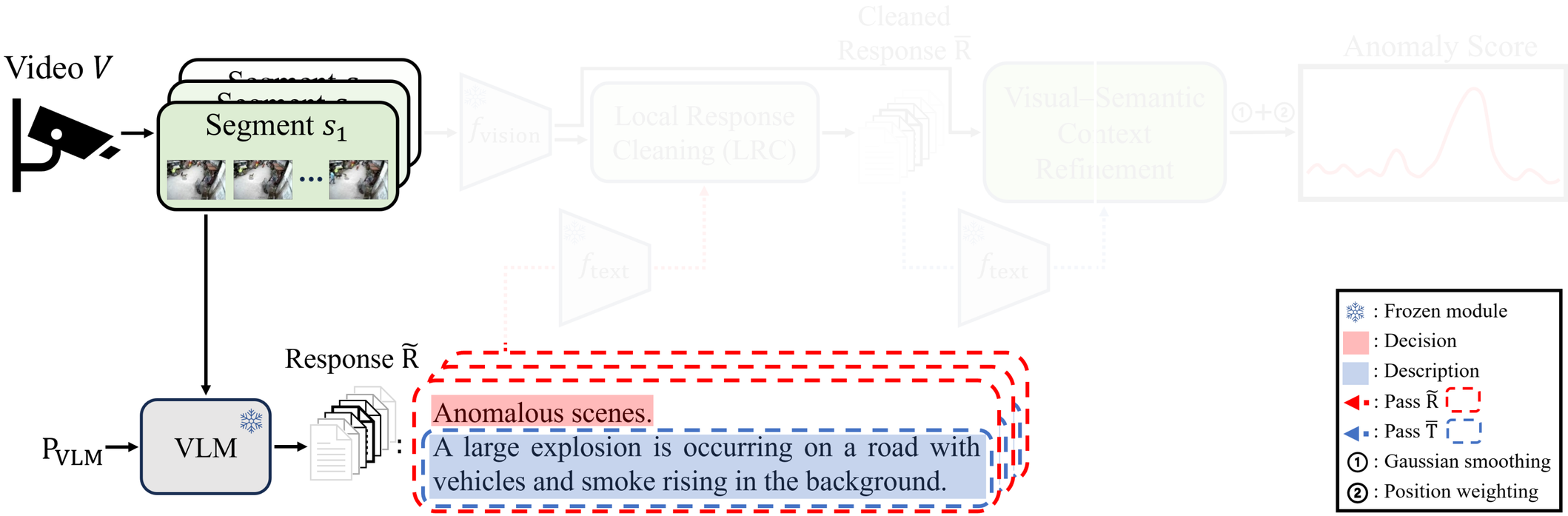
The person's forceful attempt to enter a residential area, ... heighten the suspicion.

Overall Process



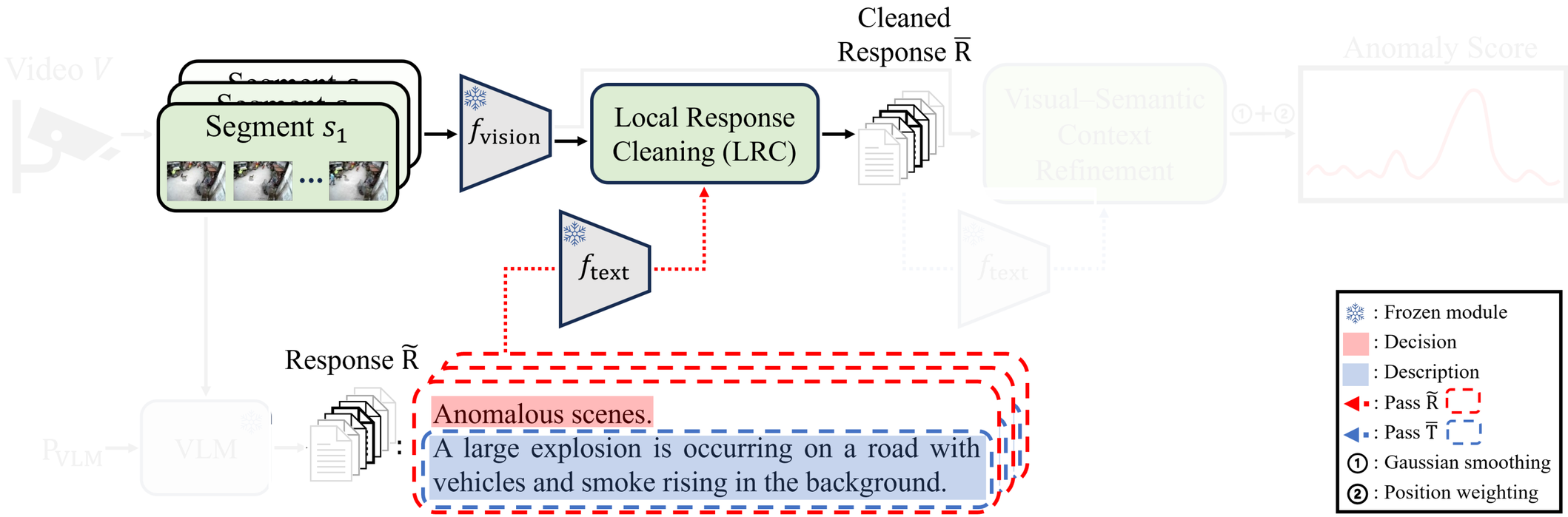
1. **VLM-Based Anomaly Scoring & Reasoning Generation:** Generate anomaly decisions and descriptions for each segment
2. **Local Response Cleaning (LRC):** Clean noisy VLM responses using local vision–text alignment
3. **Visual & Temporal Context Refinement:** Refine anomaly scores using global context and temporal progression

Overall Process



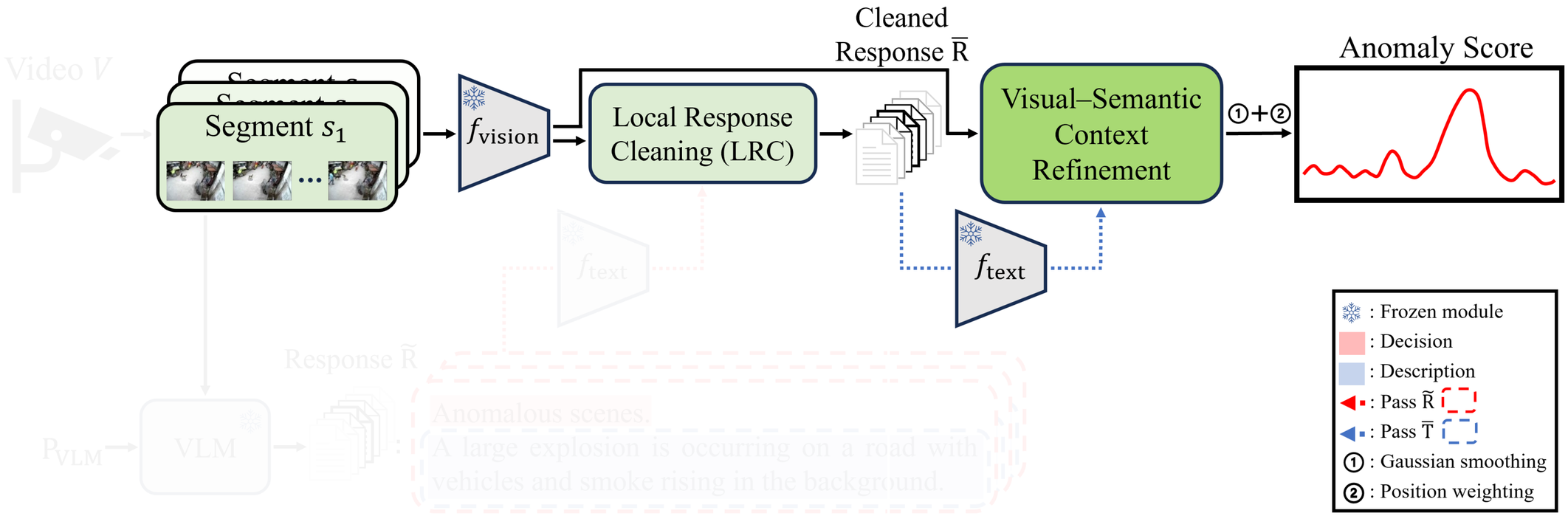
1. VLM-Based Anomaly Scoring & Reasoning Generation: Generate an anomaly decision and a description for each segment

Overall Process



2. Local Response Cleaning (LRC): Clean noisy VLM responses using local vision–text alignment

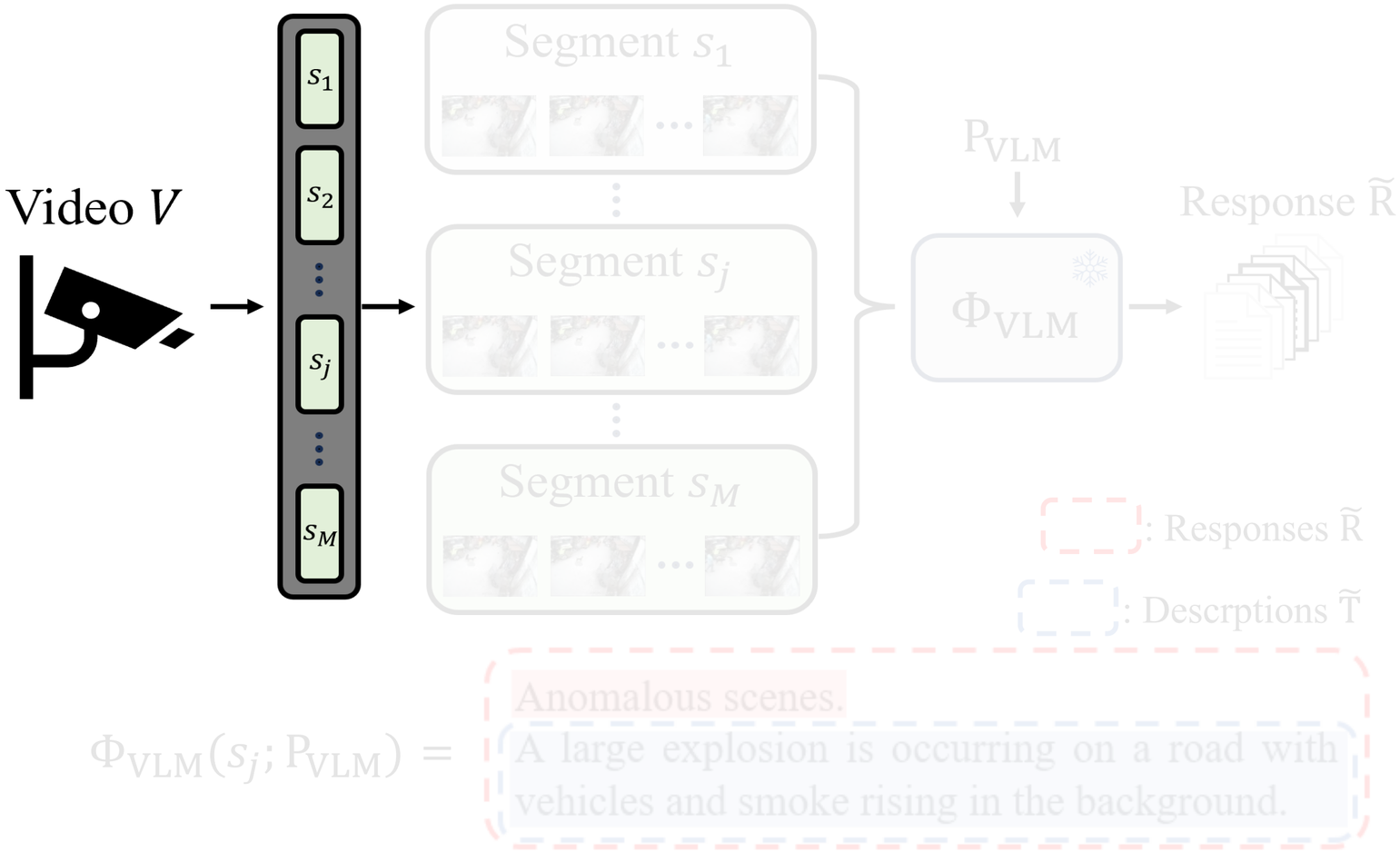
Overall Process



3. Visual & Temporal Context Refinement: Refine anomaly scores using global context and temporal progression

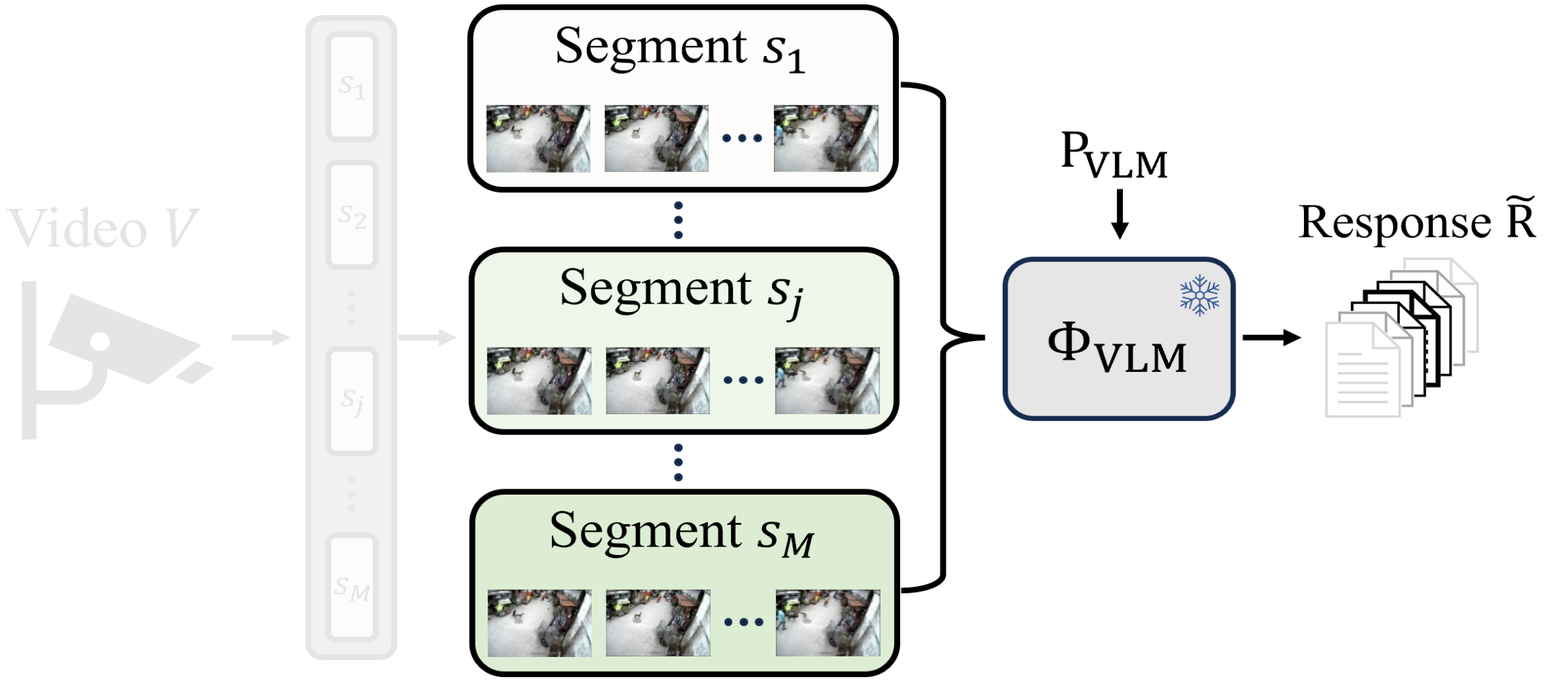
VLM-Based Anomaly Scoring & Reasoning Generation

- The input video is divided into multiple fixed-length segments
- Each segment serves as an independent unit for VLM-based analysis



VLM-Based Anomaly Scoring & Reasoning Generation

- Frames are uniformly sampled from each segment and fed to the frozen VLM with a predefined prompt
- Each VLM response consists of a binary anomaly decision and a temporal description



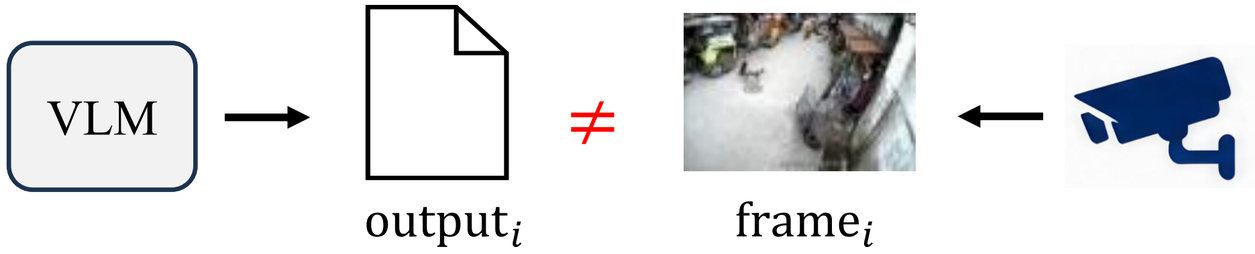
$$\Phi_{\mathrm{VLM}}(s_j; \mathrm{P_{VLM}}) =$$

Anomalous scenes.
A large explosion is occurring on a road with vehicles and smoke rising in the background.

: Responses $\tilde{\mathrm{R}}$
: Anomaly Decision
: Descriptions

Local Response Cleaning (LRC): Motivation

VLM outputs may contain noise



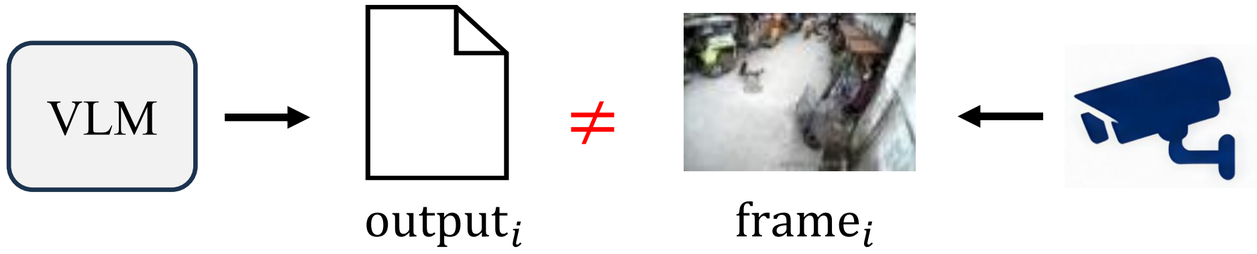
Conventional Global Caption Cleaning

- For each frame, the most relevant response is selected from all captions across the entire video.

[1] Harnessing Large Language Models for Training-free Video Anomaly Detection (Zanella et al. 2024 CVPR)

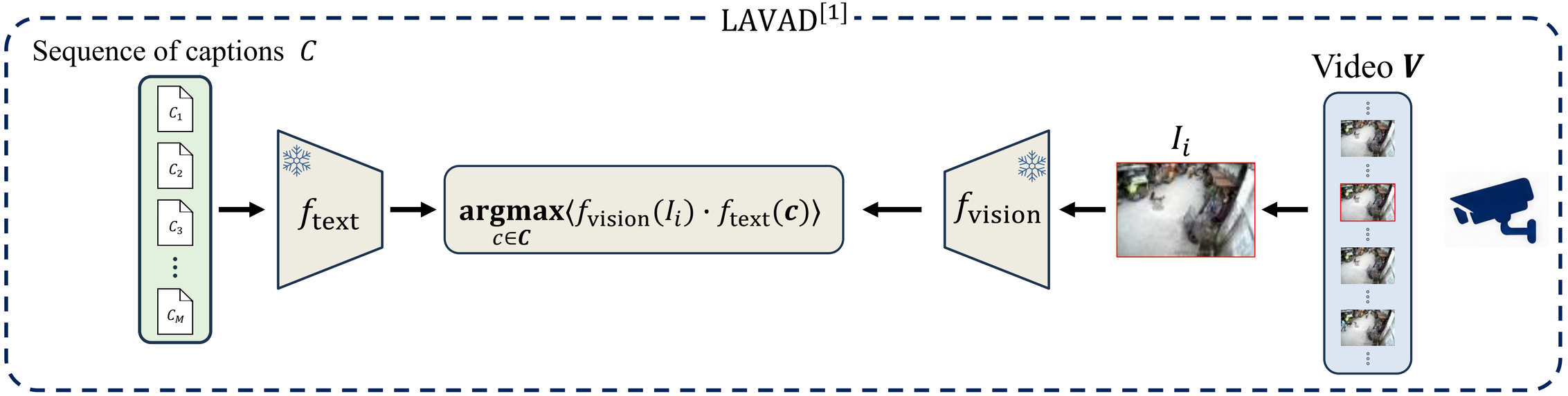
Local Response Cleaning (LRC): Motivation

VLM outputs may contain noise



Conventional Global Caption Cleaning

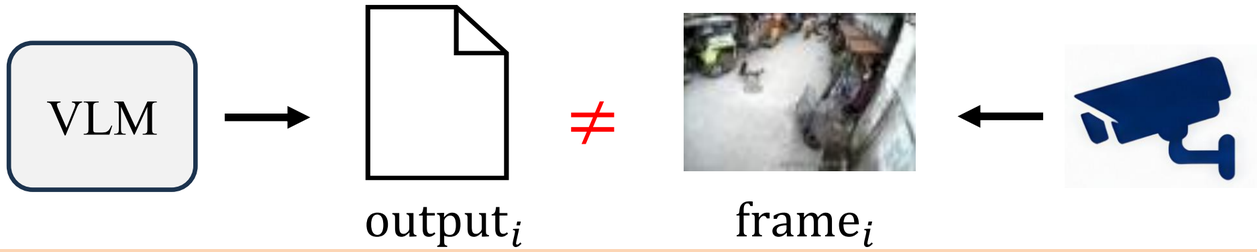
- For each frame, the most relevant response is selected from all captions across the entire video.



[1] Harnessing Large Language Models for Training-free Video Anomaly Detection (Zanella et al. 2024 CVPR)

Local Response Cleaning (LRC): Motivation

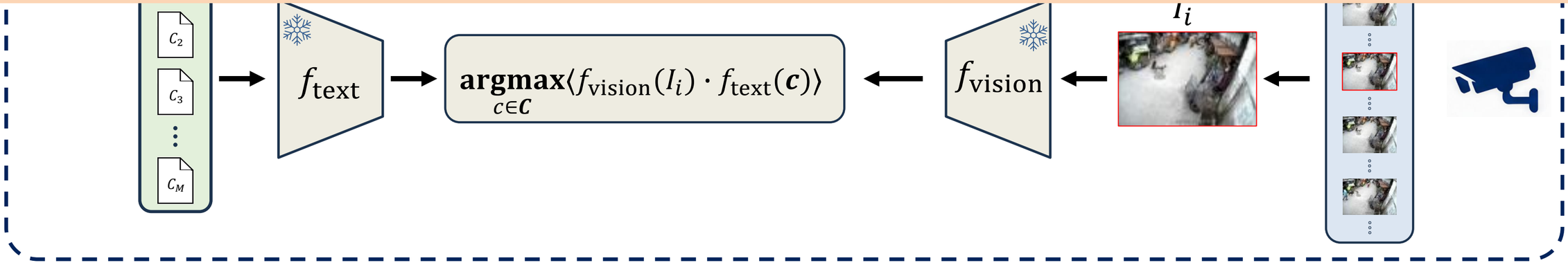
VLM outputs may contain noise



Problem

High Computational Cost: Requires comparisons with all frame-level video captions

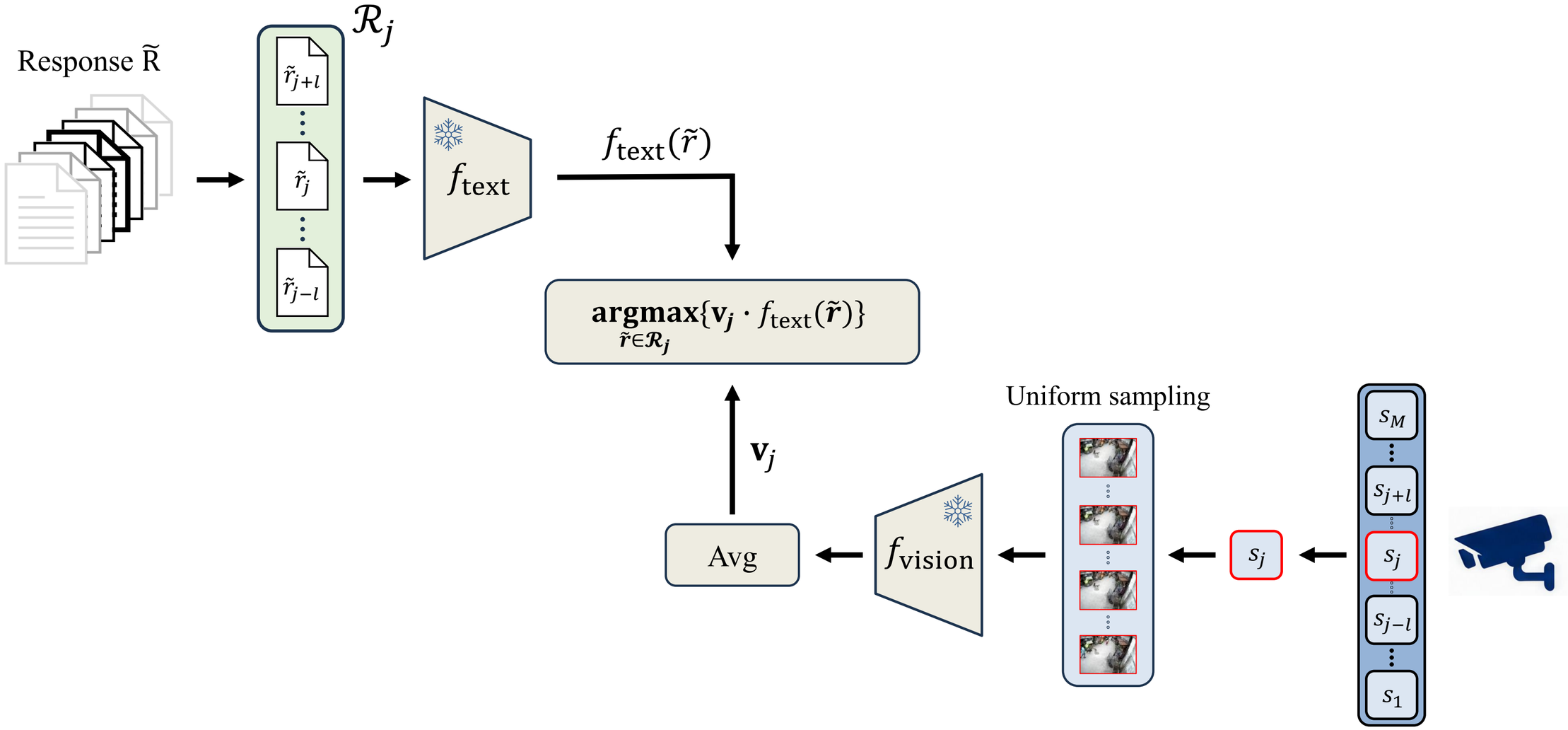
Irrelevant Global Context: May be replaced by a caption from an unrelated scene



[1] Harnessing Large Language Models for Training-free Video Anomaly Detection (Zanella et al. 2024 CVPR)

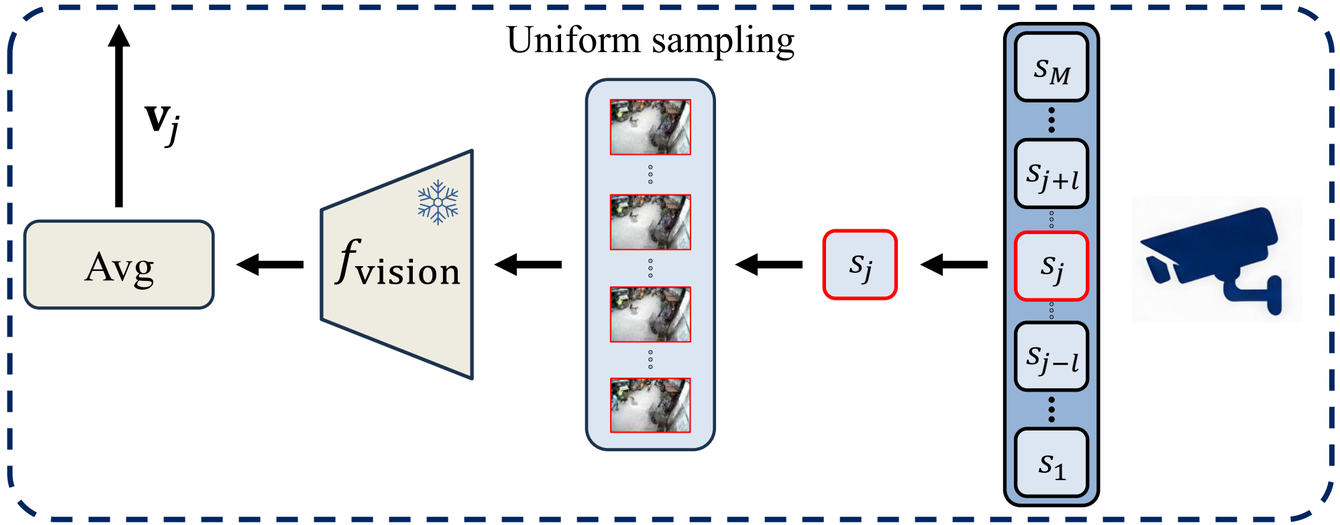
Local Response Cleaning (LRC)

Detailed process of LRC

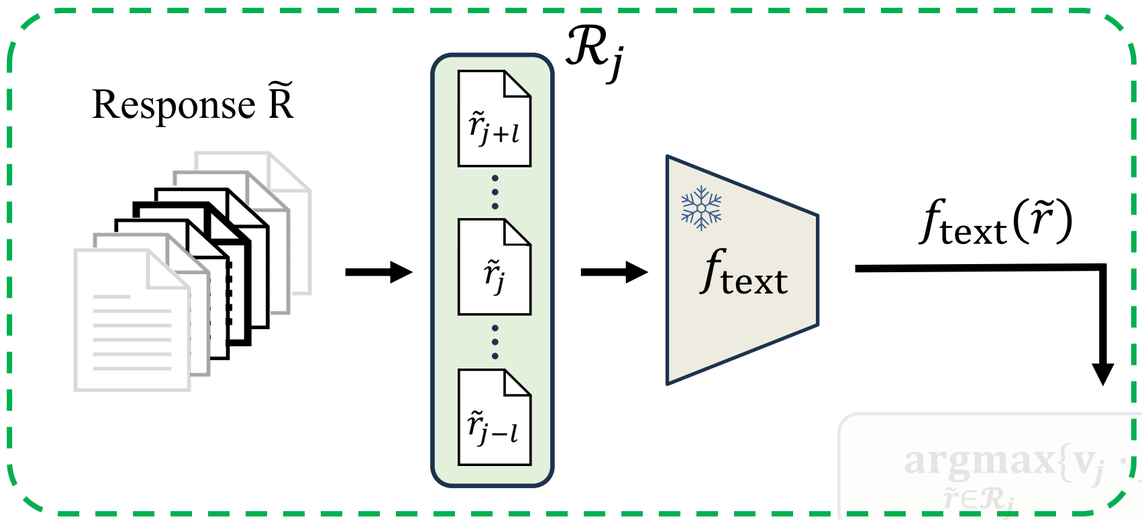


Local Response Cleaning (LRC)

- Uniformly sample frames from the target segment
- Extract and average the visual features in the target segment frames

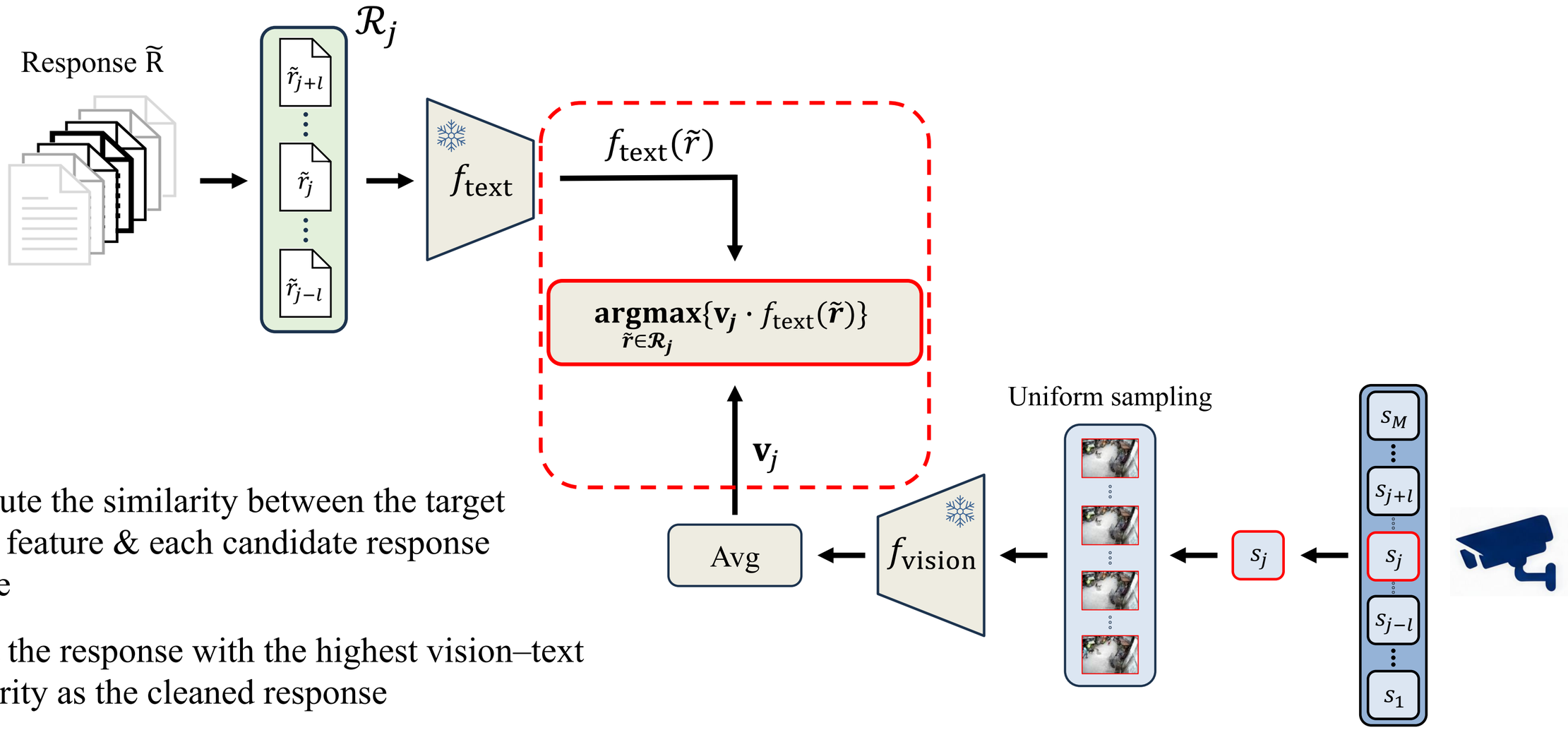


Local Response Cleaning (LRC)



- Extract text features from the responses of the target and neighboring segments

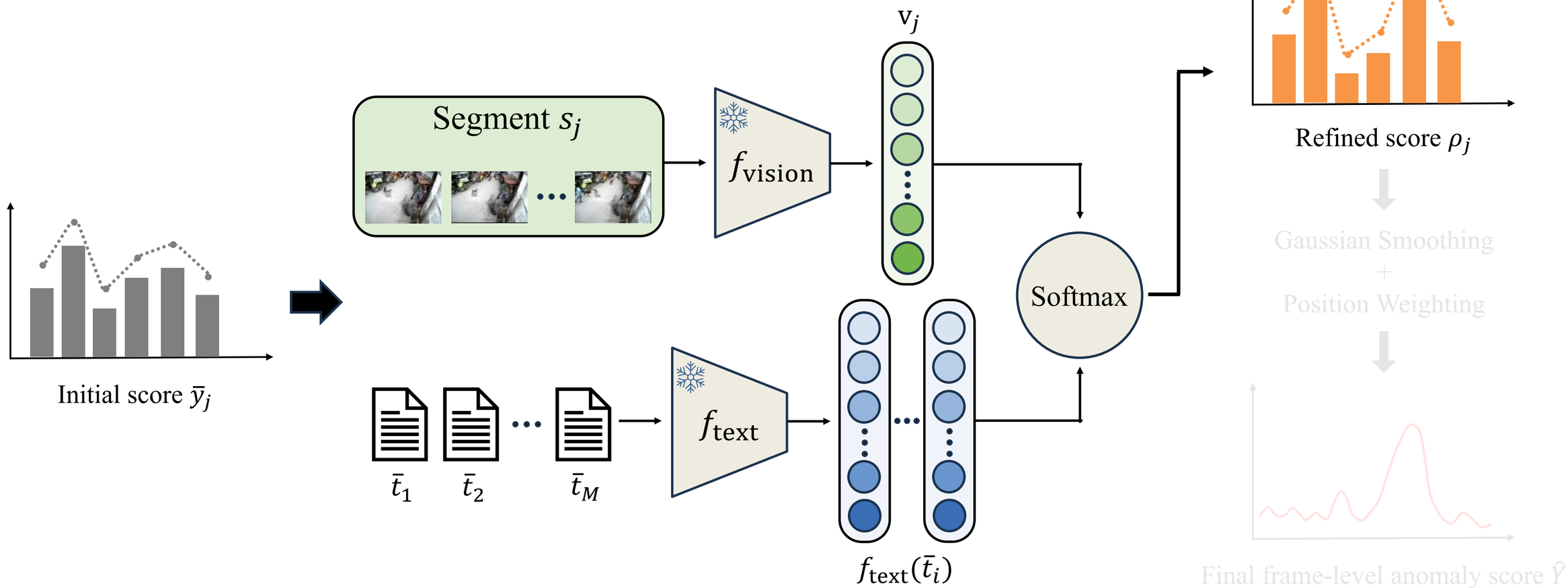
Local Response Cleaning (LRC)



- Compute the similarity between the target visual feature & each candidate response feature
- Select the response with the highest vision–text similarity as the cleaned response

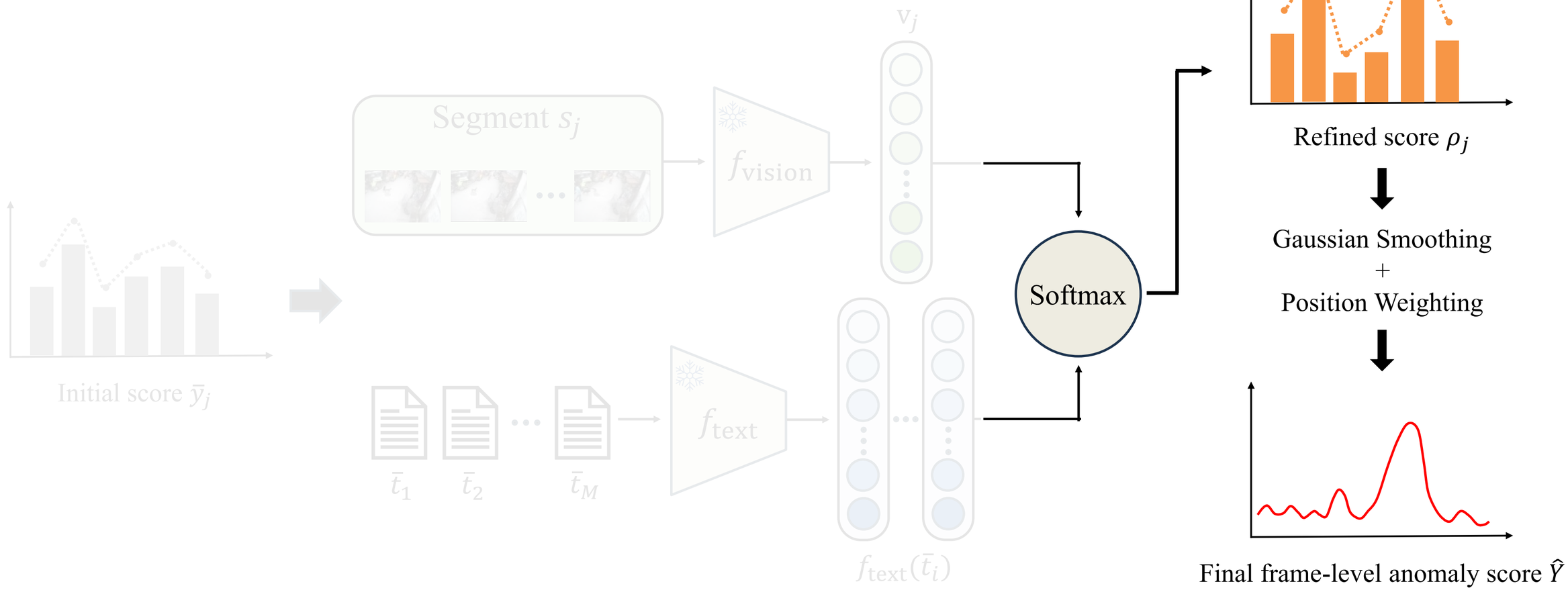
Visual and Temporal Context Refinement

- LRC-refined outputs capture local temporal context
- Refinement incorporates global temporal context



Visual and Temporal Context Refinement

- Reflect temporal continuity and anomaly progression
- Segment-level refinement anomaly score → stable frame-level anomaly score



Experimental Setup

Dataset

UCF-Crime

- 1,900 videos
- 290 test video

XD-Violence

- 4,754 videos
- 800 test video

Implementation Details

Frozen VLM: InternVL2-8B

↓

30-frame segments

↓

8 sampled frames

↓

Vision-text alignment: CLIP

UCF-Crime

XD-Violence

Comparison with State-of-the-Art Methods

UCF-Crime

Explainable	Category	Method	AUC(%)
Non-explainable VAD Methods	Weakly supervised	Sultani [23]	77.92
		GCL [36]	79.84
		Wu et al. [31]	82.44
		RTFM [26]	84.30
		MSL [13]	85.62
		S3R [29]	85.99
		MGFN [3]	86.98
		CLIP-TSA [10]	87.58
		VADCLIP [32]	88.02
	One-class	SVM [23]	50.00
		Hasan et al. [9]	51.20
		SSV [22]	58.50
		BODS [28]	68.26
		GODS [28]	70.46
	Unsupervised	GCL [36]	71.04
		DyAnNet [25]	79.76
Explainable VAD Methods	Fine-tuning	VADor [18]	88.10
		Holmes-VAU [38]	88.96
	Verbalized	VERA [35]	86.55
	Training-free	ZS CLIP [37]	53.16
		ZS IMAGEBIND (Image) [37]	53.65
		ZS IMAGEBIND (Video) [37]	55.78
		LLAVA-1.5 [8]	72.84
		LAVAD [37]	80.28
		EventVAD [21]	82.03
		MCANet [5]	82.47
		Ours	**82.51**

XD-Violence

Explainable	Category	Method	AP(%)	AUC(%)
Non-explainable VAD Methods	Weakly supervised	Wu et al. [31]	73.20	-
		Wu & Liu [30]	75.90	-
		RTFM [26]	77.81	-
		MSL [13]	78.58	-
		S3R [29]	80.26	-
		MGFN [3]	80.11	-
		CLIP-TSA [10]	82.17	-
		VADCLIP[32]	84.51	-
	One-class	Hasan et al. [9]	-	50.32
		Lu et al. [17]	-	53.56
		BODS [28]	-	57.32
		GODS [28]	-	61.56
	Unsupervised	RareAnom [24]	-	68.33
Explainable VAD Methods	Fine-tuning	Holmes-VAU[38]	87.68	-
	Verbalized	VERA [35]	70.54	88.26
	Training-free	ZS CLIP [37]	17.83	38.21
		ZS IMAGEBIND (Image) [37]	27.25	58.81
		ZS IMAGEBIND (Video) [37]	25.36	55.06
		LLAVA-1.5 [8]	50.26	79.62
		LAVAD [37]	62.01	85.36
		EventVAD [21]	64.04	87.51
		MCANet [5]	69.72	87.43
		Ours	**70.94**	**91.44**

Comparison with State-of-the-Art Methods

UCF-Crime

Explainable	Category	Method	AUC(%)
Non-explainable VAD Methods	Weakly supervised	Sultani [23]	77.92
		GCL [36]	79.84
		Wu et al. [31]	82.44
		RTFM [26]	84.30
		MSL [13]	85.62
		S3R [29]	85.99
		MGFN [3]	86.98
		CLIP-TSA [10]	87.58
		VADCLIP [32]	88.02
	One-class	SVM [23]	50.00
		Hasan et al. [9]	51.20
		SSV [22]	58.50
		BODS [28]	68.26
		GODS [28]	70.46
	Unsupervised	GCL [36]	71.04
		DyAnNet [25]	79.76
Explainable VAD Methods	Fine-tuning	VADor [18]	88.10
		Holmes-VAU [38]	88.96
	Verbalized	VERA [35]	86.55
	Training-free	ZS CLIP [37]	53.16
		ZS IMAGEBIND (Image) [37]	53.65
		ZS IMAGEBIND (Video) [37]	55.78
		LLAVA-1.5 [8]	72.84
		LAVAD [37]	80.28
		EventVAD [21]	82.03
		MCANet [5]	82.47
		Ours	**82.51**

XD-Violence

Explainable	Category	Method	AP(%)	AUC(%)
Non-explainable VAD Methods	Weakly supervised	Wu et al. [31]	73.20	-
		Wu & Liu [30]	75.90	-
		RTFM [26]	77.81	-
		MSL [13]	78.58	-
		S3R [29]	80.26	-
		MGFN [3]	80.11	-
		CLIP-TSA [10]	82.17	-
		VADCLIP[32]	84.51	-
	One-class	Hasan et al. [9]	-	50.32
		Lu et al. [17]	-	53.56
		BODS [28]	-	57.32
		GODS [28]	-	61.56
	Unsupervised	RareAnom [24]	-	68.33
Explainable VAD Methods	Fine-tuning	Holmes-VAU[38]	87.68	-
	Verbalized	VERA [35]	70.54	88.26
	Training-free	ZS CLIP [37]	17.83	38.21
		ZS IMAGEBIND (Image) [37]	27.25	58.81
		ZS IMAGEBIND (Video) [37]	25.36	55.06
		LLAVA-1.5 [8]	50.26	79.62
		LAVAD [37]	62.01	85.36
		EventVAD [21]	64.04	87.51
		MCANet [5]	69.72	87.43
		Ours	**70.94**	**91.44**

Ablation Studies & Qualitative Analysis

Cleaning	AUC(%)
No cleaning (baseline)	81.51
Caption cleaning [37]	80.47
LRC	**82.51**

🔍 Effect of Response Cleaning

- Prior cleaning may select temporally unrelated responses
- LRC preserves local context and improves response reliability

Vision-Text Score Refinement	Gaussian Smoothing	Position Weighting	AUC (%)
-	-	-	73.76
✓	-	-	81.46
✓	✓	-	81.75
✓	✓	✓	**82.51**

🔍 Detection–Explanation Alignment

- Anomaly scores (—) closely align with ground-truth intervals (■).
- Generated descriptions correctly explain the detected events.

🔍 Contribution of Score Refinement

- Visual–semantic context provides the largest performance improvement.
- Temporal refinement produces more stable frame-level anomaly scores.

Ablation Studies & Qualitative Analysis

Cleaning	AUC(%)
No cleaning (baseline)	81.51
Caption cleaning [37]	80.47
LRC	**82.51**

Effect of Response Cleaning

- Prior cleaning may select temporally unrelated responses
- LRC preserves local context and improves response reliability

Detection–Explanation Alignment

- Anomaly scores (—) closely align with ground-truth intervals ( ).
- Generated descriptions correctly explain the detected events.

Vision-Text Score Refinement	Gaussian Smoothing	Position Weighting	AUC (%)
-	-	-	73.76
✓	-	-	81.46
✓	✓	-	81.75
✓	✓	✓	**82.51**

Contribution of Score Refinement

- Visual–text refinement provides the largest performance improvement.
- Temporal refinement produces more stable frame-level anomaly scores.

Ablation Studies & Qualitative Analysis

Cleaning	AUC(%)
No cleaning (baseline)	81.51
Caption cleaning [37]	80.47
LRC	**82.51**

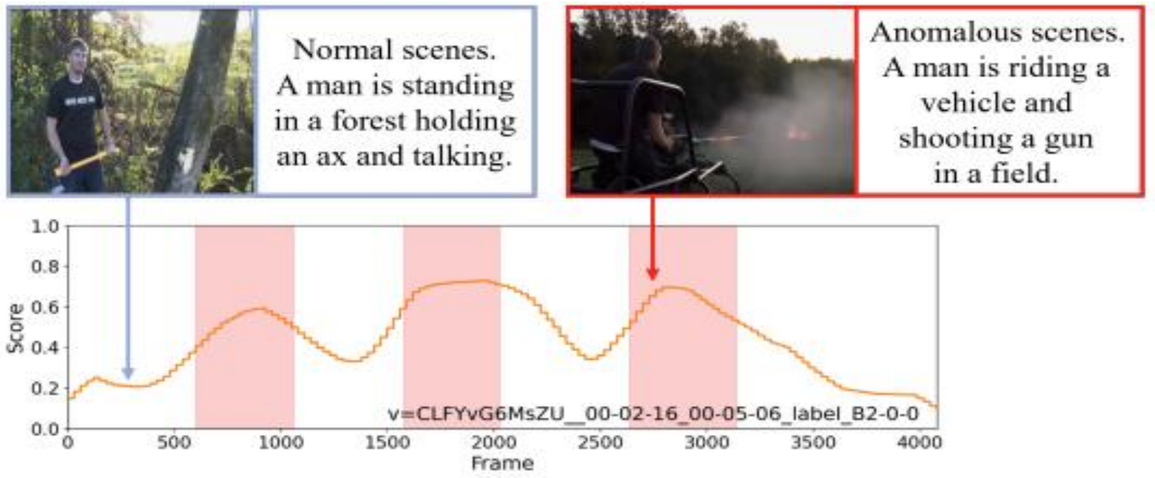
Effect of Response Cleaning

- Prior cleaning may select temporally unrelated responses
- LRC preserves local context and improves response reliability

Vision-Text Score Refinement	Gaussian Smoothing	Position Weighting	AUC (%)
-	-	-	73.76
✓	-	-	81.46
✓	✓	-	81.75
✓	✓	✓	**82.51**

Contribution of Score Refinement

- Visual–semantic context provides the largest performance improvement.
- Temporal refinement produces more stable frame-level anomaly scores.



Detection–Explanation Alignment

- Anomaly scores (—) closely align with ground-truth intervals (■).
- Generated descriptions correctly explain the detected events.

Conclusion

1. **Training-free VAD with a single frozen VLM**, without additional training or external LLMs

2. **Reliable anomaly scoring** and **interpretable descriptions** through LRC and contextual refinement

3. **Competitive performance** with an efficient and explainable video anomaly detection framework

Thank you

Operations Research and
Artificial Intelligence Lab