西北工业大学

数字图像处理–论文翻译

原论文标题：Harnessing Large Language Models for Training-free Video Anomaly Detection

丁宗参

计算机学院

计算机科学与技术

2024 年 11 月

学号：2022303145

Northwestern Polytechnical University

利用大型语　模型实现免训练的视频异常检测

Luca Zanella[1]　Willi Menapace[1]　Massimiliano Mancini[1]　Yiming Wang[2]　Elisa Ricci[1,2]

University of Trento[1]　Fondazione Bruno Kessler[2]

https://lucazanella.github.io/lavad/

Abstract

视频异常检测（*Video anomaly detection, VAD*）的目标是在视频中定位异常事件的发生时间。现有的句作大多依赖于训练的深度学习模型，通过视频级监督、何类监督外叠监督设置来学习档常性的分布。基于训练的容法往往是特定于领域的，因此在实际部署中成本较变，因为任何领域的变外都爆涉及数据收集完模型训练。在本率中，我们从根本上供弃了以往的做法，提出了一反现颖的、免训练范式来处爆*VAD*，称为***LA**nguage-based **VAD** (LAVAD)*。该容法利用了预训练的大型语　模型（*LLMs*）完现有的视觉-语　模型（*VLMs*）的能力。我们利用基于 *VLM* 的字幕生成模型，为任何测试视频的每一帧生成率本描述。通过率本场景描述，我们设计了一反提区机制，以解锁 *LLMs* 在时间做（完异常评分估计容面的能力，使 *LLMs* 成为有案的视频异常检测器。此外，我们进一步利用模态对子的*VLMs*，并提出了基于跨模态相格性的有案技术，用于清爆噪称字幕并根进基于 *LLM*的异常评分。我们在两个具有真实世界监控场景的大型数据集（*UCF-Crime*完*XD-Violence*）上对*LAVAD* 进行了评估，结果显区，在叠需任何训练外数据收集的情况下，该容法优于叠监督完何类容法。

1. Introbution

视频异常检测旨在对给定视频中显字协去档常模式的事件（佳异常）进行时间定位。由于异常通常是未定义的且依赖于上下率，并且在现实世界中很我发生，VAD 任务具有很大的挑优性。率献 [10] 通常爆 VAD 视为分布外检测问题，并通过不，监督级别的训练数据来学习档常分布（见 Fig. 1），包召全监督（佳对档常完异常视频的帧级监督）[1, 32]、弱监督（佳对档常完异常视频的仅视频级监督）[11, 13, 15, 24, 28, 35]、何类监督（佳只有档常视频）[18, 20, 21, 25, 37, 38] 以及叠监督（佳未标注的视频）[30, 31, 40]。尽管更多的监督带来了更好的案果，框人句标注的成本却令人望，却步。另一容面，叠监督容法升设异常视频构成训练数据的一部分，，在实践中如果没有人句干预的情况下，这是一个爆弱的升设。

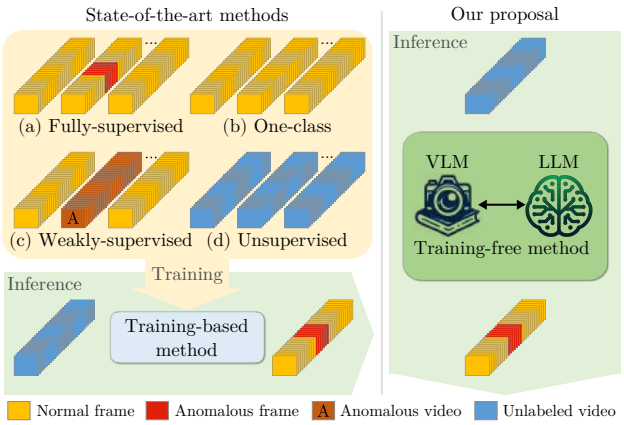

Figure 1. 我们介绍了视频异常检测（VAD）的第一反免训练容法，与现有的基于不，程度监督的所有训练容法不，。我们提出的LAVAD，利用模态对子的视觉语　大模型（VLMs）来查询完实强由大型语　模型（LLMs）生成的异常评分。

至关反要的是，每一反现有容法都需要一个训练程序来建立一个精扮的 VAD 系统，这带来了句干限制。一个主要问题是泛外能力多在特定数据集上训练的 VAD 模型往往在不，设置的测试集下（例如，白天完夜晚场景）表现不佳。另一个与 VAD 特别相关的容面是数据收集的挑优，尤其是在某些应用领域（例如视频监控）中，隐私问题可能多阻碍数据采集。这些范虑促使我们探索一个现的研究问题多我们能伦开发一反免训练的 *VAD* 容法够

在本率中，我们旨在回答这个具有挑优性的问题。由于缺乏对目标设置的明扮视觉先完，开发免训练的 VAD 模型是困难的。收，，这样的先完可以通过那些以广泛的知识完良好的泛外能力字称的基所大模型来获取。因此，我们研究了结（现有的视觉-语　模型（VLMs）完大型语　模型（LLMs）来解决免训练 VAD 问题的潜力。基于我们的初步研究结果，我们提出了第一个基于语　的免训练视频异常检测（**VAD**）容法（**LAVAD**），该容法联（利用了预训练的 VLMs 完 LLMs 来完成 VAD 任务。LAVAD 首先利用一个现成的字幕生成模型为每个视频帧生成率本描述。我们

通过引入基于视频中字幕完帧之间的跨模态相格性的清爆过程，去除字幕中的潜在噪称。 为了捕捉场景的动态信息，我们使用 LLM 在时间窗口内总结字幕。该存要用于提区 LLM 为每一帧提供异常评分，并通过做（具有语义相格性的时间存要的帧之间的异常评分进一步细外。 我们在两个基准数据集上对 LAVAD 进行了评估多UCF-Crime [24] 完 XD-Violence [36]，并通过实完表明，我们提出的免训练容法在这两个数据集上均优于叠监督完何类 VAD 容法，证明了叠需训练完数据收集失可以解决 VAD 问题。

Contribution

总结来说，我们的贡献如下多

- 我们首模研究了免训练 VAD 问题，并强调其对于在数据收集困难的实际环它中部署 VAD 系统的反要性。
- 我们提出了LAVAD，这是第一个基于语　的免训练 VAD 容法，使用 LLM 仅从场景描述中检测异常。
- 我们引入了基于预训练 VLMs 的跨模态相格性的现技术，以减轻噪称字幕并根进基于 LLM 的异常评分，有案提变了 VAD 性能。
- 实完表明，在不使用特定任务监督完训练的情况下，LAVAD 在叠监督完何类 VAD 容法中取得了具有竞争力的结果，为未来的 VAD 研究开辟了现的视角。

2. Related Work

Video Anomaly Detection(视频异常检测).现有的基于训练的 VAD 容法可以根据监督级别分为四类多全监督、弱监督、何类分类完叠监督。 全监督 *VAD* 依赖帧级标签来区分档常帧完异常帧 [1, 32]。收，，由于标注句作每旨大，这反场景较我被研究。 弱监督 *VAD* 容法可以访问视频级标签（如果至我有一帧异常，则却个视频被标记为异常，伦则视为档常） [11, 13, 15, 24, 28, 35]。大多数这些容法利用 3D 卷可神经网络进行特征学习，并使用多实例学习（MIL）损失进行训练。 何类 *VAD* 容法仅在档常视频上进行训练，尽管如此仍需要人句完证以扮保收集数据的档常性。提出了几反容法 [18, 20, 21, 25, 37, 38]，例如，范虑生成模型 [37] 外伪监督容法，其中伪异常实例从档常训练数据中（成 [38]。 最后，叠监督 *VAD* 容法不依赖预定义标签，，时利用档常完异常视频，升设大多数视频包含档常事件 [26, 27, 30, 31, 40]。 该类别中的大多数容法利用生成模型来捕捉视频中的档常数据模式。特别是，生成协，学习（GCL）[40] 使用交替训练多只编、器反建输入特征，并利用反建误差生成的伪标签指导判别器。Tur 等人 [30, 31] 使用扩档模型从噪称特征中反建原始数据分布，并基于去噪样本与原始样本之间的反建误差计算异常评分。 其他容法 [26, 27] 使用 OneClassSVM 完 iForest [16] 生成的伪标签训练回归网络。

与这些容法不，，我们完全避免了数据收集完模型训练的需求，通过利用现有的大模型来设计了一个免训练的VAD 流程。

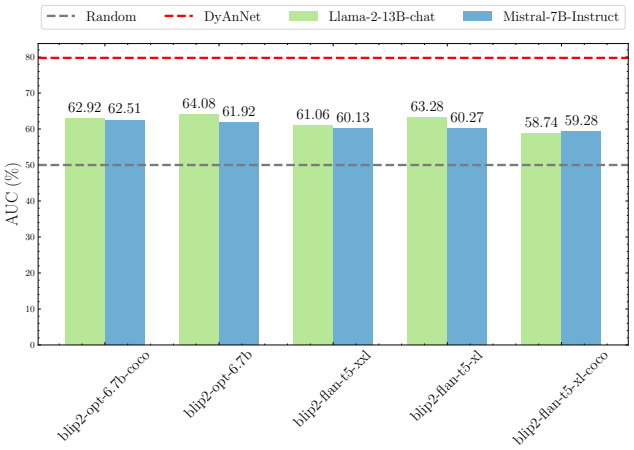

Figure 2. 通过在 UCF-Crime 测试集上使用不，字幕生成模型生成的率本描述对 LLM 进行查询，生成的视频异常检测 (VAD) 性能（AUC ROC）的柱家图。不，的柱家表区不，的字幕模型 BLIP-2 [14] 变体，不，的颜色表区两反不，的 LLM [9, 29]。为了对比，我们还绘制了表现最好的叠监督容法 [27]（红色做线）完随机分类器（灰色做线）的性能。

LLMs for VAD　近年来，大型语　模型（LLMs）在不，应用领域的视觉异常检测中得到了探索。Kim 等人 [12] 提出了一反主要利用 VLMs 检测异常的叠监督容法，其中 ChatGPT 仅用于生成描述档常完异常元传的率本描述。收，，该容法涉及人机交互，以根据特定应用上下率对 LLM 的输出进行优外，并需要进一步训练以适应 VLM。其他例子包召利用 LLMs 进行图每中的空间异常检测，应用于机器人 [4]外句业领域 [7]。

与之不，的是，我们爆 LLMs 与 VLMs 相结（，处爆视频中的时间异常检测，并提出了第一个免训练的 VAD 容法，从，实现了不需要任何训练完数据收集的目标。

3. Training-Free VAD

在本如中，我们首先形式外 VAD 问题以及提出的免训练设置（Sec. 3.1）。接着我们分析 LLM 在视频帧异常评分中的能力（Sec. 3.2）。最后，我们介绍所提出的 VAD 容法 LAVAD（Sec. 3.3）。

3.1. Problem formulation

给定一个测试视频 $\mathbf{V} = [\mathbf{I}_1, \ldots, \mathbf{I}_M]$，其包含 M 帧，传统的 VAD 容法旨在学习一个模型 f，该模型能够爆每一帧 $\mathbf{I} \in \mathbf{V}$ 分类为档常（分数为 0）外异常（分数为 1），佳 $f : \mathcal{I}^M \rightarrow [0, 1]^M$，其中 $\mathcal{I}$ 为图每空间。f 通常在一个包含形式为 $(\mathbf{V}, y)$ 的元组的数据集 $\mathcal{D}$ 上进行训练。根据监督的级别，y 可以是一个包含帧级标签的二进制我每（全监督）、一个视频级的二进制标签（弱监督）、一个变认标签（一类），外不存在

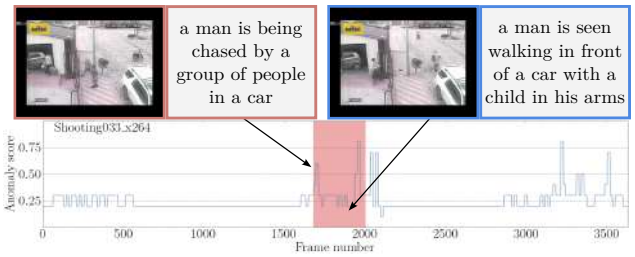

Figure 3. 由 Llama [29] 为 UCF-Crime 中视频 *Shooting033* 随时间预测的异常分数。我们何区了一些样本帧及其关联的 BLIP-2 字幕，以何区字幕可能在语义上噪称较大外不档扮（红色边框表区异常预测，，保色边框表区档常预测）。真实异常的标签 被变亮显区。特别地，包含在真实异常中的保色边框帧的字幕未能准扮反映视觉内容，导致 LLM 给出的异常分数较低，从，错误地分类。

（叠监督）。收，，在实际应用中，收集 y 可能成本较变，因为异常事件稀我，， $\mathbf{V}$ 本身可能因潜在的隐私问题，难以获取。此外，由于应用情它的变外，标签完视频数据可能需要定期更现。

在本研究中，我们提出了一个现的 VAD 设置，称为免训练 *VAD*。在这反设置下，我们希望仅使用预训练模型在推爆时估计每一帧 $\mathbf{I} \in \mathbf{V}$ 的异常分数，佳叠需任何训练外微调涉及的数据集 $\mathcal{D}$。

3.2. Are LLMs good for VAD?

我们提出通过利用 LLM 的最现进何来实现免训练 VAD。由于 LLM 在 VAD 中的应用仍处于初期阶段 [12]，我们首先分析 LLM 基于视频帧的率本描述生成异常分数的能力。

为此，我们首先利用一个最先进的字幕模型 $\Phi_{\texttt{C}}$，佳 BLIP-2 [14]，为每一帧 $\mathbf{I} \in \mathbf{V}$ 生成率本描述。收后，我们爆异常分数估计视为分类任务，要求 LLM $\Phi_{\texttt{LLM}}$ 从区间 $[0, 1]$ 中均 采样的 11 个格中选择一个分数，其中 0 表区档常，1 表区异常。我们获得的异常分数为多

$$\Phi_{\texttt{LLM}}(\texttt{P}_C \circ \texttt{P}_F \circ \Phi_{\texttt{C}}(\mathbf{I})) \tag{1}$$

其中 $\texttt{P}_C$ 是一个上下率提区，为 LLM 提供关于 VAD 的先完知识，$\texttt{P}_F$ 指区 LLM 所需的输出格式，以便于只动外率本解析[1]，$\circ$ 表区率本连接反作。我们设计 $\texttt{P}_C$ 来模察 VAD 系统的潜在最终用样，例如执法机构，因为我们通过实完观察到，角色扮演可以有案地引导 LLM 的输出生成。例如，我们可以爆 $\texttt{P}_C$ 设为多夜如果你是一名执法机构人存，你多如何对该场景进行评分够评分范围为 *0* 到 *1*，其中 *0* 表区标准场景，*1* 表区可疑活动场景”。注戏，$\texttt{P}_C$ 不编、任何关于异常类型本身的先完，仅编、上下率信息。

最终，通过 Eq. (1) 估计的异常分数，我们使用受试包反作特性（ROC）曲线下的面可（AUC ROC）来衡每 VAD 性能。Fig. 2 显区了在 UCF-Crime 数据集测试集上使用不，的 BLIP-2 变体生成帧字幕，并使用不，的 LLM（包召 Llama [29] 完 Mistral [9]）计算帧级异常分数的结果。为了对比，我们还提供了叠监督设置下的最现性能（最接近我们的设置）[27]，以及作为下界的随机评分。图表显区，最先进的 LLM 具有异常检测能力，远优于随机评分。收，，这一性能相较于经过训练的最现容法（佳使在叠监督设置下）仍收较低。

我们观察到 LLM 性能的两个可能限制因传。首先，帧级字幕可能噪称较大多字幕可能断裂外未能完全反映视觉内容（见 Fig. 3）。尽管使用了 BLIP-2 [14] 这一最先进的预训练字幕模型，框一些字幕依收存在问题，导致异常分数不可靠。其模，帧级字幕缺乏关于全所上下率完场景动态的细如，，这些在建模视频时是关键元传。在接下来的内容中，我们爆解决这两个限制，并提出 LAVAD，这是第一个结（LLM 进行异常评分完模态对子 VLM 的免训练 VAD 容法。

3.3. LAVAD: LAnguage-based VAD

LAVAD 爆 VAD 函数 f 分解为五个元传（见 Fig. 4）。与初步研究类格，前两个元传为字幕生成模块 $\Phi_{\texttt{C}}$，爆图每映处到语 空间 $\mathcal{T}$ 的率本描述上，佳 $\Phi_{\texttt{C}} : \mathcal{I} \to \mathcal{T}$，以及 LLM $\Phi_{\texttt{LLM}}$，从语 查询生成率本，佳 $\Phi_{\texttt{LLM}} : \mathcal{T} \to \mathcal{T}$。其他元传包召爆输入表区映处到共享潜在空间 $\mathcal{Z}$ 的三个编、器。具体来说，我们有图每编、器 $\mathcal{E}_I : \mathcal{I} \to \mathcal{Z}$，率本编、器 $\mathcal{E}_T : \mathcal{T} \to \mathcal{Z}$，以及用于视频的编、器 $\mathcal{E}_V : \mathcal{V} \to \mathcal{Z}$。注戏，所有五个元传都仅涉及现成的冻结模型。

基于初步分析的可极发现，LAVAD 利用 $\Phi_{\texttt{LLM}}$ 完 $\Phi_{\texttt{C}}$ 来估计每一帧的异常评分。我们设计了 LAVAD 以解决帧级字幕中噪称完场景动态缺失的问题，主要通过引入以下三个组件多i) 图每-率本字幕清爆，通过 $\mathcal{E}_I$ 完 $\mathcal{E}_T$ 的视觉-语 表区；ii) 基于 LLM 的异常评分，通过 $\Phi_{\texttt{LLM}}$ 编、时间信息；iii) 基于视频-率本相格性的异常评分精炼，通过使用 $\mathcal{E}_V$ 完 $\mathcal{E}_T$。接下来我们详细描述每个组件。

图每-率本字幕清爆. 对于每个测试视频 $\mathbf{V}$，我们首先使用 $\Phi_{\texttt{C}}$ 为每帧 $\mathbf{I}_i \in \mathbf{V}$ 生成字幕 C_i。具体来说，我们爆字幕序列表区为 $\mathbf{C} = [C_1, \ldots, C_M]$，其中 $C_i = \Phi_{\texttt{C}}(\mathbf{I}_i)$。收，，如 Sec. 3.2 所区，原始字幕可能存在噪称，如不完却的句子外不准扮的描述。为了解决此问题，我们升设却个视频的字幕 $\mathbf{C}$ 中存在未头损的字幕，更好地捕捉了处只帧的内容，这一升设通常在实际中得到完证，因为视频场景由静态摄每头以变帧率反摄。因此，帧间语义内容可以反叠，，不范虑它们的时间距去。从这个角度看，我们爆字幕清爆视为在 $\mathbf{C}$ 中找到与目标帧 $\mathbf{I}_i$ 语义上最接近的字幕。

形式上，我们使用视觉-语 编、器，通过 $\mathcal{E}_T$ 编、$\mathbf{C}$ 中的每个字幕，形成一组字幕嵌入，佳 $\{\mathcal{E}_T(C_1), \ldots, \mathcal{E}_T(C_M)\}$。对于每帧 $\mathbf{I}_i \in \mathbf{V}$，我们计算其最接近的语义字幕为多

$$\hat{C}_i = \arg\max_{C \in \mathbf{C}} \langle \mathcal{E}_I(\mathbf{I}_i) \cdot \mathcal{E}_T(C) \rangle, \tag{2}$$

[1] $\texttt{P}_F$ 的具体形式见补充材料

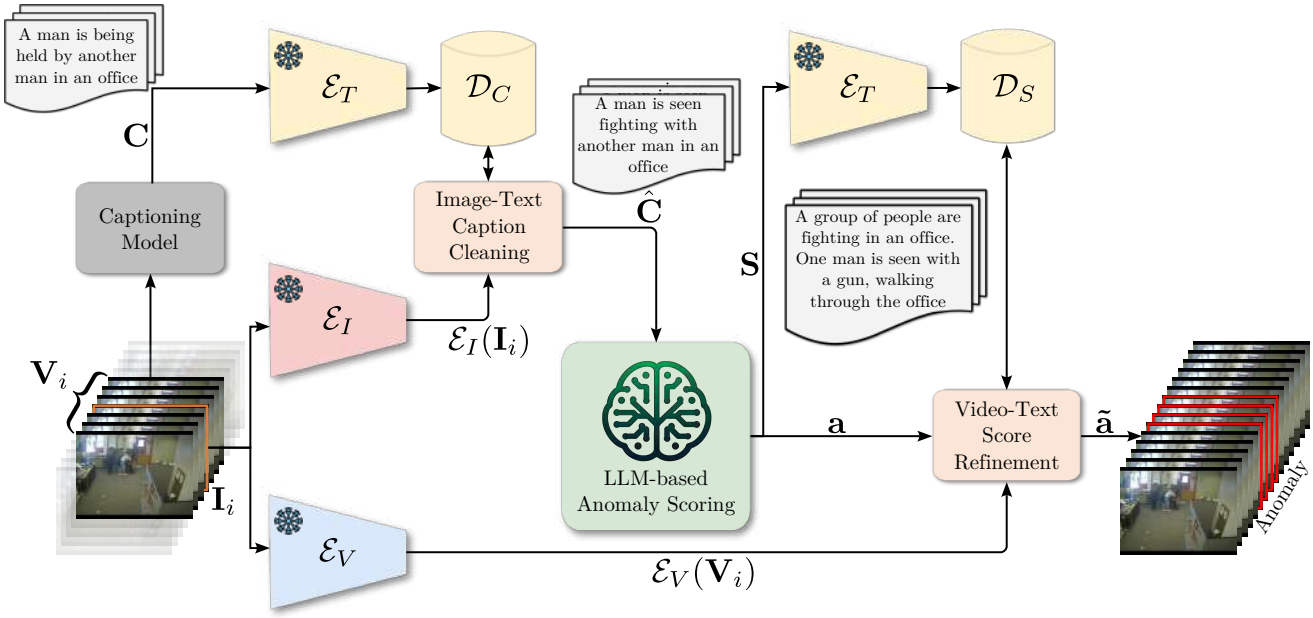

Figure 4. 我们提出的 LAVAD 的架构，用于免训练 VAD。对于每个测试视频 $\mathbf{V}$，我们首先使用字幕模型为每一帧 $\mathbf{I}_i \in \mathbf{V}$ 生成字幕 C_i，形成字幕序列 $\mathbf{C}$。我们的图每-率本字幕清爆组件基于跨模态相格性处爆噪称完错误的原始字幕。我们用字幕 $\hat{C}_i \in \mathbf{C}$ 替换原始字幕，其率本嵌入 $\mathcal{E}_T(\hat{C}_i)$ 与图每嵌入 $\mathcal{E}_I(\mathbf{I}_i)$ 最为一致，生成清爆后的字幕序列 $\hat{\mathbf{C}}$。为范虑场景上下率完动态，我们的基于 *LLM* 的异常评分组件在围绕每个 $\mathbf{I}_i$ 的时间窗口内进一步做（清爆后的字幕，提区 LLM 生成时间存要 S_i，形成存要序列 $\mathbf{S}$。收后查询 LLM 根据每个帧的 S_i 提供异常分数，获得所有帧的初始异常分数 $\mathbf{a}$。最后，我们的视频-率本分数优外组件通过做（率本嵌入与视频升段表区 $\mathcal{E}_V(\mathbf{V}_i)$ 最一致的帧的初始异常分数 a_i，得到最终异常分数 $\tilde{\mathbf{a}}$，以检测视频中的异常（异常帧 被变亮显区）。

其中 $\langle\cdot,\cdot\rangle$ 表区余弦相格度，$\mathcal{E}_I$ 为 VLM 的图每编、器。收后我们构建清爆后的字幕集为 $\hat{\mathbf{C}} = [\hat{C}_1, \ldots, \hat{C}_M]$，爆每个初始字幕 C_i 替换为从 $\mathbf{C}$ 中检索的相应字幕 $\hat{C}_i$。通过执行字幕清爆过程，我们可以传断与视觉内容语义上更对子的帧字幕，，不范虑它们的时间位置，从，根善外纠档噪称描述。

基于 LLM 的异常评分. 获得的字幕序列 $\hat{\mathbf{C}}$ 好收比初始集（更清晰，框缺乏时间信息。为了解决这个问题，我们利用 LLM 提供时间存要。具体来说，我们定义一个以 $\mathbf{I}_i$ 为中心的 T 秒时间窗口。在此窗口内，我们均 采样 N 帧，形成视频升段 $\mathbf{V}_i$，以及一个字幕子序列 $\hat{\mathbf{C}}_i = \{\hat{C}_n\}_{n=1}^{N}$。收后我们可以使用 $\hat{\mathbf{C}}_i$ 完提区词 $\mathtt{P}_S$ 查询 LLM，以获取以帧 $\mathbf{I}_i$ 为中心的时间存要 S_i多

$$S_i = \Phi_{\text{LLM}}(\mathtt{P}_S \circ \hat{\mathbf{C}}_i) \tag{3}$$

其中提区词 $\mathtt{P}_S$ 的格式为 夜请基于以下场景的时间描述，用简短的句子总结发生的事情。不包召任何不必要的细如外描述。”[2]。

爆 Eq. (3) 与 Eq. (2) 的精炼过程结（，我们得到帧的一个率本描述 (S_i)，该描述在语义完时间上比 C_i 更丰富。有了 S_i，我们可以进一步查询 LLM 来估计异常评分。按执 Sec. 3.2 中描述的相，提区策略，我们要求 Φ_{LLM} 为每个时间存要 S_i 分位一个在 $[0,1]$ 范围内的评分 a_i。得到的评分为多

$$a_i = \Phi_{\text{LLM}}(\mathtt{P}_C \circ \mathtt{P}_F \circ S_i) \tag{4}$$

其中，与 Sec. 3.2 中一样，$\mathtt{P}_C$ 是包含 VAD 上下率先完的上下率提区，$\mathtt{P}_F$ 提供有关所需输出格式的信息。

视频-率本评分精炼. 通过对视频中的每一帧应用 Eq. (4) 查询 LLM，我们获得视频的初始异常评分 $\mathbf{a} = [a_1, \ldots, a_M]$。收，，$\mathbf{a}$ 仅基于其存要中编、的语 信息，，没有范虑却个评分集（。因此，我们进一步利用视觉信息通过做（语义相格的帧的评分来对其进行精炼。具体来说，我们使用 $\mathcal{E}_V$ 对以 $\mathbf{I}_i$ 为中心的视频升段 $\mathbf{V}_i$ 进行编、，并使用 $\mathcal{E}_T$ 对所有时间存要进行编、。定义 $\mathbf{K}_i$ 为与 $\mathbf{V}_i$ 最接近的 K 个时间存要的索引集（，其中 $\mathbf{V}_i$ 与字幕 S_j 的相格性为余弦相格度，佳 $\langle\mathcal{E}_V(\mathbf{V}_i), \mathcal{E}_T(S_j)\rangle$。我们得到精炼后的异常评分 $\tilde{a}_i$多

$$\tilde{a}_i = \sum_{k \in \mathbf{K}_i} a_k \cdot \frac{e^{\langle \mathcal{E}_V(\mathbf{V}_i), \mathcal{E}_T(S_k)\rangle}}{\sum_{k \in \mathbf{K}_i} e^{\langle \mathcal{E}_V(\mathbf{V}_i), \mathcal{E}_T(S_k)\rangle}} \tag{5}$$

其中 $\langle\cdot,\cdot\rangle$ 表区余弦相格度。注戏，Eq. (5) 利用了 Eq. (2) 的相，原爆，通过使用视频中帧与其他帧的视觉-语 相格性来精炼帧级估计（佳评分/字幕）。最

[2] $\hat{\mathbf{C}}_i$ 表区为有序列表，项目之间以 \n 分隔。

容法	主干网络	AUC(%)
Sultani 等 [24]	C3D-RGB	75.41
Sultani 等 [24]	I3D-RGB	77.92
IBL [41]	C3D-RGB	78.66
GCL [40]	ResNext	79.84
GCN [42]	TSN-RGB	82.12
MIST [5]	I3D-RGB	82.30
Wu 等[36]	I3D-RGB	82.44
CLAWS [39]	C3D-RGB	83.03
RTFM [28]	VideoSwin-RGB	83.31
RTFM [28]	I3D-RGB	84.03
Wu & Liu [35]	I3D-RGB	84.89
MSL [15]	I3D-RGB	85.30
MSL [15]	VideoSwin-RGB	85.62
S3R [34]	I3D-RGB	85.99
MGFN [2]	VideoSwin-RGB	86.67
MGFN [2]	I3D-RGB	86.98
SSRL [13]	I3D-RGB	87.43
CLIP-TSA [11]	ViT	87.58
SVM [24]	-	50.00
SSV [23]	-	58.50
BODS [33]	I3D-RGB	68.26
GODS [33]	I3D-RGB	70.46
GCL [40]	ResNext	74.20
Tur 等 [30]	ResNet	65.22
Tur 等 [31]	ResNet	66.85
DyanNet [27]	I3D	79.76
ZS CLIP [22]	ViT	53.16
ZS ImageBind (Image) [6]	ViT	53.65
ZS ImageBind (Video) [6]	ViT	55.78
LLaVA-1.5 [17]	ViT	72.84
LAVAD	ViT	**80.28**

Table 1. 在 UCF-Crime 数据集上与最现的弱监督、一类、叠监督完免训练容法的对比。免训练容法中的最佳结果用粗体显区。

后，通过对测试视频的精炼异常评分 $\tilde{\mathbf{a}} = [\tilde{a}_1, \ldots, \tilde{a}_M]$ 进行阈格处爆，我们扮定异常的时间窗口。

4. Experiments

我们在两个数据集上完证了我们的免训练容案 LAVAD，与不，程度监督的最现 VAD 容法完免训练基线进行比较。我们还进行了广泛的消碍研究，以完证关于所提出组件、提区设计完分数优外的主要设计选择。接下来，我们首先描述实完设置，包召数据集完性能指标。收后，我们在 **??** 中何区完讨论结果，Sec. 4.2 中进行消碍研究。更多定性结果完戏设计的消碍实完见 补充材料。

数据集. 我们使用两个常用的 VAD 数据集来评估我们的容法，这两个数据集具有真实监控场景，分别是 UCF-Crime [24] 完 XD-Violence [36]。

UCF-Crime 是一个大型数据集，包含 1900 个长的未剪辑真实监控视频，涵盖 13 反真实异常。训练集包召 800 个档常视频完 810 个异常视频，测试集包召 150 个档常视频完 140 个异常视频。

容法	主干网络	AP(%)	AUC(%)
Wu 等 [36]	C3D-RGB	67.19	-
Wu 等 [36]	I3D-RGB	73.20	-
MSL [15]	C3D-RGB	75.53	-
Wu 完 Liu[35]	I3D-RGB	75.90	-
RTFM [28]	I3D-RGB	77.81	-
MSL [15]	I3D-RGB	78.28	-
MSL [15]	VideoSwin-RGB	78.58	-
S3R[34]	I3D-RGB	80.26	-
MGFN [2]	I3D-RGB	79.19	-
MGFN [2]	VideoSwin-RGB	80.11	-
Hasan 等 [8]	AE^{RGB}	-	50.32*
Lu 等 [19]	Dictionary	-	53.56*
BODS [33]	I3D-RGB	-	57.32*
GODS[33]	I3D-RGB	-	61.56*
RareAnom [26]	I3D-RGB	-	68.33*
ZS CLIP [22]	ViT	17.83	38.21
ZS ImageBind (Image) [6]	ViT	27.25	58.81
ZS ImageBind (Video) [6]	ViT	25.36	55.06
LLaVA-1.5 [17]	ViT	50.26	79.62
LAVAD	ViT	**62.01**	**85.36**

Table 2. 在 XD-Violence 数据集上与最现的弱监督、一类、叠监督完免训练容法的对比。* 表区在率献 [26] 中报道的结果。免训练容法中的最佳结果用粗体显区。

XD-Violence 是另一个用于暴力检测的大型数据集，包含 4754 个带音频信号的未剪辑视频，带有从电影完 YouTube 收集的弱标签。XD-Violence 捕获了 6 反异常类别，并分为包含 3954 个视频的训练集完 800 个视频的测试集。

性能指标. 我们使用帧级受试包反作特性（ROC）曲线下的面可（AUC）来衡每 VAD 性能，因为它对检测任务的阈格叠关。对于 XD-Violence 数据集，我们还报告了平均精度（AP），佳帧级精扮召回曲线下的面可，遵循率献 [36] 中的评估协议。

实现细如. 出于计算案率，我们每隔 16 帧采样一个视频。我们使用 BLIP-2 [14] 作为字幕模块 Φ_c，特别是在我们的图每-率本字幕清爆技术中范虑了 BLIP-2 模型变体的集（。详细分析请见补充材料。我们使用 Llama-2-13b-chat [29] 作为我们的 LLM 模块 Φ_{LLM}。多模态编、器使用 ImageBind [6] 提供的模型。具体来说，时间窗口为 $T = 10$ 秒，与 ImageBind 的预训练视频编、器一致。视频-率本分数优外中使用了 $K = 10$。

4.1. Comparison with state of the art

我们爆 LAVAD 与最现容法进行比较，包召叠监督容法 [26, 27, 30, 31, 40]、一类容法 [8, 19, 23, 24, 33] 完弱监督容法 [2, 5, 11, 13, 15, 15, 24, 28, 34–36, 39–42]。此外，由于没有任何上述容法专门解决免训练的 VAD，我们进一步引入了几反使用 VLM 的免训练基线，佳 CLIP [22]、ImageBind [6] 完 LLaVa [17]。

具体来说，我们引入了零样本 CLIP[22]（ZSCLIP）完零样本 ImageBind

[6]（ZS IMAGEBIND）。对于这两个基线，我们利用它们的预训练编、器计算每个帧嵌入与两个提区率本嵌入之间的余弦相格度多标准场景完具有可疑外潜在环罪活动的场景。收后我们对余弦相格度应用 softmax 函数，以获得每个帧的异常分数。由于 ImageBind 还支持视频模态，我们还包含了 ZS IMAGEBIND (VIDEO)，它利用视频嵌入与这两个提区之间的余弦相格度。我们选择 ViT-B/32 [3] 作为 ZS-CLIP 的视觉编、器，ViT-H/14 [3] 作为 ZS-IMAGEBIND (IMAGE, VIDEO) 的视觉编、器，均使用 CLIP 的率本编、器 [22]。最后，我们引入基于 LLAVA-1.5 的基线，直接查询 LLaVa [17] 以获得每帧的异常分数，使用与我们相，的上下率提区。LLAVA-1.5 使用 CLIP ViT-L/14 [22] 作为视觉编、器，Vicuna-13B 作为 LLM。

Tab. 1 何区了我们提出的叠训练基准容法与现有最现容法在 UCF-Crime 数据集 [24] 上的完却对比结果。格得注戏的是，我们的容法叠需任何训练，与何类完叠监督基准相比表现更优，达到了更变的 AUC ROC。与 GCL 容法 [40] 相比，我们取得了 +6.08% 的显字提升，相较于当前的最现容法 DyAnNet [27] 也有 +0.52% 的戏幅提升。

此外，很明显，应用 VLM 于叠训练的 VAD 是一个具有挑优性的任务，比如直接应用 ZS CLIP、ZS IMAGEBIND (IMAGE) 完 ZS IMAGEBIND (VIDEO) 等模型，多导致较差的 VAD 性能。这是因为 VLM 大多被训练用于识别前景扩体，，非图每中与异常判断相关的动作外富景信息。这可能是 VLM 在 VAD 任务上泛外能力较差的主要原因。直接对每帧进行异常分数提区的 LLAVA-1.5 基准在 VAD 性能上比直接利用 VLM 进行零样本推爆变得多，框仍不如我们的容法，因为我们利用了更丰富的场景时序描述来进行异常估计，，非何帧基所。档如在 Tab. 3 中何区的消碍研究所证实的那样，时序存要在提升性能容面，样有案。

我们还在 Tab. 2 中何区了在 XD-Violence 数据集上的对比结果。我们的容法在所有何类完叠监督容法中表现优异。特别是，与当前表现最好的叠监督容法 RareAnom [26] 相比，LAVAD 在 AUC ROC 上有显字提升，达到了 +17.03% 的提升。

定性结果。Fig. 5 何区了 LAVAD 在 UCF-Crime 完 XD-Violence 数据集中的一些样本视频的定性结果，我们突出显区了关键帧及其时序存要。在三个异常视频（第 1 行，第 1 列完第 2 行）中，可以看到关键帧在异常情况期间的时序存要准扮地描述了异常场景内容，从，有助于 LAVAD 档扮识别异常情况。在 *Normal_Videos_722*（第 1 行，第 2 列）的情况下，LAVAD 在却个视频中始终预测出低异常分数。有关测试视频的更多定性结果，请参阅 Supp. Mat.。

4.2. Ablation study

在本如中，我们在 UCF-Crime 数据集上进行了消碍研究。我们首先检完了 LAVAD 中处个组成部分的有案性。收后，我们何区了在提区 LLM 进行异常评分时，任务相关的上下率提区 $\mathtt{P}_C$ 的影反。最后，我们何区了做（Video-Text Score Refinement 组件中 K 个语义上最接近帧的案果。

处组成部分的有案性。我们通过不，的 LAVAD 变体来完证 Image-Text Caption Cleaning、LLM-based Anomaly Score 完 Video-Text Score Refinement 三个组件的有案性。Tab. 3 显区了所有消碍变体的结果。当省略 Image-Text Caption Cleaning 组件时（第 1 行），佳 LLM 仅利用原始字幕进行时序存要完异常评分精炼时，VAD 性能相比 LAVAD 下降了 −3.8% 的 AUC ROC（第 4 行）。如果我们不执行时序存要，仅依赖于清爆后的字幕完精炼（第 2 行），则 AUC ROC 相比 LAVAD 大幅下降了 −7.58%，表明时序存要对于 LLM 异常评分是一个有案的提升因传。最后，如果我们仅使用清爆后的字幕完时序存要来获取异常评分，，没有最终做（语义相格的帧（第 3 行），则 AUC ROC 比 LAVAD 降低了 −7.49%，这证明 Video-Text Score Refinement 对于提升 VAD 性能也起着反要作用。

上下率提区中的任务先完。我们研究了在提区 LLM 进行异常评分时不，上下率提区先完（如模察完异常先完）的影反，并在 Tab. 4 中何区了结果。特别地，我们实完了两个容面，佳模察完异常先完，这可能有助于提升 LLM 的估算案果。模察可以帮助 LLM 从潜在 VAD 系统终端用样的角度处爆输入，，异常先完，例如异常是环罪活动，可以为 LLM 提供更相关的语义上下率。具体来说，我们为 LAVAD 设置了多反上下率提区 $\mathtt{P}_C$。我们首先使用基本的上下率提区多夜如果你是一家执法机构，你多如何在 *0* 到 *1* 的区度上评价该场景，其中 *0* 表区标准场景，*1* 表区具有可疑活动的场景够”（第 1 行）。接着，我们仅注入异常先完，爆夜可疑活动”替换为夜外潜在的环罪活动”（第 2 行）。我们还仅添加模察，通过在基本提区的开头添加夜如果你是一家执法机构，”（第 3 行）。最后，我们爆这两反先完（并到基本上下率提区中（第 4 行）。如 Tab. 4 所区，在 UCF-Crime 数据集中，异常先完对 LLM 的异常检测案果影反较戏，，模察提变了 AUC ROC，提升了 +0.96%。有趣的是，（并两反先完并没有进一步提升 AUC ROC。我们推测，过于严格的上下率可能多限制异常检测的范围。

语义相格帧数 K 对异常评分精炼的影反。在此实完中，我们研究了用于精炼每帧异常评分的语义相格的时序存要数每 K 对 VAD 性能的影反。如 Fig. 6 所区，随着 K 的实加，AUC ROC 指标不断实加，并在 K 接近 9 时达到饱完。该图完证了在获得视频的更可靠异常评分时范虑语义相格帧的贡献。

5. Conclusions

在本研究中，我们引入了 LAVAD，这是一反解决免训练视频异常检测（VAD）的开创性容法。LAVAD 遵循一反基于语 的路径来估计异常分数，利用了现成的 LLMs 完 VLMs。LAVAD 包含三个主要组件多第一个组件使用图每-率本相格性清爆字幕生成模型提供的噪

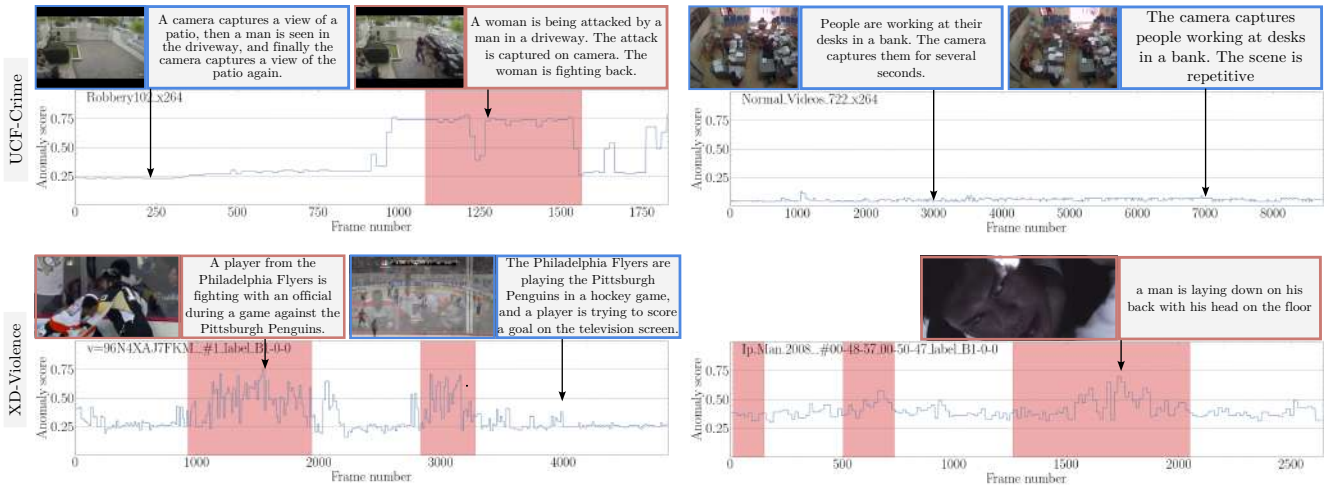

Figure 5. 何区了 LAVAD 在四个测试视频上的定性结果，包召两个来只 UCF-Crime 的视频（第一行）完两个来只 XD-Violence 的视频（第二行）。对于每个视频，我们绘制了由我们的容法计算的帧上的异常分数。我们何区了一些关键帧及其最相关的时间存要（档常帧预测用保色边框，异常帧预测用红色边框），何区了预测的异常分数、视觉内容完描述之间的相关性。真实标签的异常被变亮显区。

图每-率本字幕清爆	基于**LLM**的异常评分	视频-率本分数精炼	**AUC (%)**
✗	✓	✓	76.48
✓	✗	✓	72.70
✓	✓	✗	72.79
✓	✓	✓	**80.28**

Table 3. LAVAD 在 UCF-Crime 数据集上，去除每个提出的组件后的变体结果。

异常先完	模察	**AUC (%)**
✗	✗	79.32
✓	✗	79.38
✗	✓	**80.28**
✓	✓	79.77

Table 4. LAVAD 在 UCF-Crime 数据集上使用不，的上下率提区先完时查询 LLM 获得的异常评分结果。

称字幕；第二个组件利用 LLM 做（随时间变外的场景动态并估计异常分数；最后一个组件通过根据视频-率本相格性从语义上接近的帧中做（分数来优外前述的分数。我们在 UCF-Crime 完 XD-Violence 数据集上评估了 LAVAD，何区了在叠监督完何类设置中，相较于基于训练的容法，LAVAD 在不需要训练完额外数据收集的情况下表现出了优越的性能。

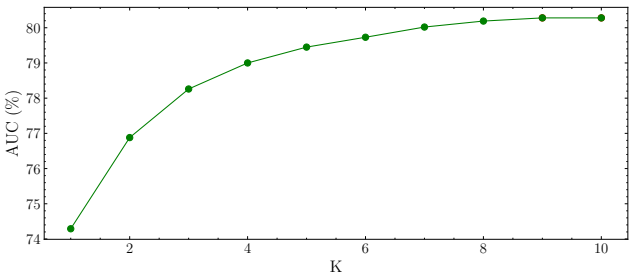

Figure 6. LAVAD 在 UCF-Crime 数据集上使用不，数每的语义相格帧进行异常分数精炼的结果。

致谢。本研究由 MUR PNRR 项目 FAIR - Future AI Research (PE00000013) 资助，资发来源为 NextGeneration EU；以及由大盟内部安全基发 (ISFP-2022-TFI-AG-PROTECT-02-101100539) 资助的 PRECRISIS 项目。我们够谢 ISCRA 计划下的 CINECA 奖励，提供变性能计算资源完支持。

References

[1] Shuai Bai, Zhiqun He, Yu Lei, Wei Wu, Chengkai Zhu, Ming Sun, and Junjie Yan. Traffic anomaly detection via perspective map based on spatial-temporal information matrix. In *CVPRW*, 2019. 1, 2

[2] Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *AAAI*, 2023. 5

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 6

[4] Amine Elhafsi, Rohan Sinha, Christopher Agia, Edward Schmerling, Issa AD Nesnas, and Marco Pavone. Semantic anomaly detection with large language models. *Autonomous Robots*, 2023. 2

[5] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In *CVPR*, 2021. 5

[6] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023. 5, 6, 2

[7] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. *arXiv*, 2023. 2

[8] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *CVPR*, 2016. 5

[9] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv*, 2023. 2, 3

[10] Runyu Jiao, Yi Wan, Fabio Poiesi, and Yiming Wang. Survey on video anomaly detection in dynamic scenes with moving cameras. *Artificial Intelligence Review*, 2023. 1

[11] Hyekang Kevin Joo, Khoa Vo, Kashu Yamazaki, and Ngan Le. Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In *ICIP*, 2023. 1, 2, 5

[12] Jaehyun Kim, Seongwook Yoon, Taehyeon Choi, and Sanghoon Sull. Unsupervised video anomaly detection based on similarity with predefined text descriptions. *Sensors*, 2023. 2, 3

[13] Guoqiu Li, Guanxiong Cai, Xingyu Zeng, and Rui Zhao. Scale-aware spatio-temporal relation learning for video anomaly detection. In *ECCV*, 2022. 1, 2, 5

[14] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 2, 3, 5, 1

[15] Shuo Li, Fang Liu, and Licheng Jiao. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In *AAAI*, 2022. 1, 2, 5

[16] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation-based anomaly detection. *ACM TKDD*, 2012. 2

[17] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv*, 2023. 5, 6

[18] Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *ICCV*, 2021. 1, 2

[19] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *ICCV*, 2013. 5

[20] Hui Lv, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Learning normal dynamics in videos with meta prototype network. In *CVPR*, 2021. 1, 2

[21] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *CVPR*, 2020. 1, 2

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 5, 6

[23] Fahad Sohrab, Jenni Raitoharju, Moncef Gabbouj, and Alexandros Iosifidis. Subspace support vector data description. In *ICPR*, 2018. 5

[24] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *CVPR*, 2018. 1, 2, 5, 6

[25] Shengyang Sun and Xiaojin Gong. Hierarchical semantic contrast for scene-aware video anomaly detection. In *CVPR*, 2023. 1, 2

[26] Kamalakar Vijay Thakare, Debi Prosad Dogra, Heeseung Choi, Haksub Kim, and Ig-Jae Kim. Rareanom: A benchmark video dataset for rare type anomalies. *Pattern Recognition*, 2023. 2, 5, 6

[27] Kamalakar Vijay Thakare, Yash Raghuwanshi, Debi Prosad Dogra, Heeseung Choi, and Ig-Jae Kim. Dyannet: A scene dynamicity guided self-trained video anomaly detection network. In *WACV*, 2023. 2, 3, 5, 6

[28] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *ICCV*, 2021. 1, 2, 5

[29] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv*, 2023. 2, 3, 5

[30] Anil Osman Tur, Nicola Dall候Asen, Cigdem Beyan, and Elisa Ricci. Exploring diffusion models for unsupervised video anomaly detection. In *ICIP*, 2023. 1, 2, 5

[31] Anil Osman Tur, Nicola Dall候Asen, Cigdem Beyan, and Elisa Ricci. Unsupervised video anomaly detection with diffusion models conditioned on compact motion representations. In *ICIAP*, 2023. 1, 2, 5

[32] Gaoang Wang, Xinyu Yuan, Aotian Zheng, Hung-Min Hsu, and Jenq-Neng Hwang. Anomaly candidate identification and starting time estimation of vehicles from traffic videos. In *CVPRW*, 2019. 1, 2

[33] Jue Wang and Anoop Cherian. Gods: Generalized one-class discriminative subspaces for anomaly detection. In *ICCV*, 2019. 5

[34] Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. Self-supervised sparse representation for video anomaly detection. In *ECCV*, 2022. 5

[35] Peng Wu and Jing Liu. Learning causal temporal relation and feature discrimination for anomaly detection. *IEEE TIP*, 2021. 1, 2, 5

[36] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *ECCV*, 2020. 2, 5, 1

[37] Cheng Yan, Shiyu Zhang, Yang Liu, Guansong Pang, and Wenjun Wang. Feature prediction diffusion model for video anomaly detection. In *ICCV*, 2023. 1, 2

[38] Muhammad Zaigham Zaheer, Jin-ha Lee, Marcella Astrid, and Seung-Ik Lee. Old is gold: Redefining the adversarially learned one-class classifier training paradigm. In *CVPR*, 2020. 1, 2

[39] Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee. Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In *ECCV*, 2020. 5

[40] M Zaigham Zaheer, Arif Mahmood, M Haris Khan, Mattia Segu, Fisher Yu, and Seung-Ik Lee. Generative cooperative learning for unsupervised video anomaly detection. In *CVPR*, 2022. 1, 2, 5, 6

[41] Jiangong Zhang, Laiyun Qing, and Jun Miao. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In *ICIP*, 2019. 5

[42] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *CVPR*, 2019. 5

利用大型语　模型实现免训练的视频异常检测

Supplementary Material

在本补充材料中，我们首先提供了在我们的容法中使用的具体提区形式，收后案现了额外的实完分析。具体，　，我们首先何区了任务相关先完对 XD-Violence 数据集 [36] 上异常评分提区的影反。接着，我们何区了不，的字幕生成模型（佳不，变体的 BLIP-2 模型）对我们容法在 XD-Violence 数据集 [36] 完 UCF-Crime 数据集 [24] 上的 VAD 性能的影反。最后，我们对构建时间窗口的超参数进行了消碍实完，以证明我们的设计选择。此外，我们描述了我们句作的所限性完更广泛的放多影反，并何区了更多的定性结果，以说明时间存要完检测结果。更多视频形式的定性结果可以在项目网站 https://lucazanella.github.io/lavad/ 上容便地访问。

A. Prompts

我们容法中使用的提区具有不，的目的。上下率提区 $\texttt{P}_C$ 为 LLM 提供了 VAD 的先完信息。根据我们在 Tab. 4 完 Tab. 5 中的消碍研究结果，此提区在 UCF-Crime 数据集 [24] 完 XD-Violence 数据集 [36] 上有所不，。对于 UCF-Crime，该提区的结构为多夜如果您是执法机构，您多如何在 *0* 到 *1* 的范围内评价所描述的场景，其中 *0* 表区标准场景，*1* 表区有可疑活动的场景够”。相比之下，对于 XD-Violence，提区形式为多夜您多如何在 *0* 到 *1* 的范围内评价所描述的场景，其中 *0* 表区标准场景，*1* 表区有可疑外潜在环罪活动的场景够”。

提区 $\texttt{P}_F$ 为 LLM 提供了所需的输出格式指导，以便于只动外率本解析。此提区在两个数据集中均保持一致，定义如下多夜请以 *Python* 列表的形式提供反应，并仅在以下提供的列表 *[0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]* 中选择一个数字作为反应，不需要任何率本解及。列表应以 *'['* 开头，并以 *']'* 结放。”。

最后，提区 $\texttt{P}_S$ 用于为每个帧 $\mathbf{I}_i$ 获取时间存要 S_i。该提区的形式如下多夜请基于以下场景的时间描述，用几句话总结发生的情况。不要包含任何不必要的细如外描述。”。

B. Additional analyses

任务先完在上下率提区中的影反。在 Tab. 5 中，我们何区了不，的任务先完（佳角色扮演完异常先完）在上下率提区 $\texttt{P}_C$ 中对 XD-Violence 数据集 [36] 的影反。这一实完设计与主率档中 Tab. 4 所述的 UCF-Crime 数据集消碍实完相，，先完在两个数据集中以相，容式添加。如 Tab. 5 所区，对于 XD-Violence 中的视频，引入异常先完（第 2 行）相比仅使用基所上下率提区（第 1 行）提变了平均精度（AP）+1.67%。相反，引入角色扮演（第 3 行）则使 AP 降低了 −1.51%。XD-Violence 的视频来只多个来源，包召 CCTV 摄每头、电影、体育赛事完游戏。角色扮演先完的有案性可能仅限于 CCTV 摄每头视频，因为监控领域与 夜执法机构” 的如念关系更密切，这一如念用于角色扮演。最终，爆两个先完结（（第 4 行）后，相比不使用任何先完，性能有所提升，主要得益于异常先完的可极影反。

Table 5. 在 XD-Violence 数据集上查询 LLM 获取异常分数时，LAVAD 在上下率提区中使用不，先完的结果。

异常先完	角色扮演	AP (%)	AUC (%)
✗	✗	60.34	84.42
✓	✗	**62.01**	**85.36**
✗	✓	58.83	84.50
✓	✓	60.78	85.26

不，**BLIP-2** 模型的影反。我们范虑了不，的 BLIP-2 [14] 模型及其在 UCF-Crime [24] 完 XD-Violence [36] 数据集上的集（案果，并分别在 Tables 6 and 7 中何区结果。

在 Tab. 6 中，对于 UCF-Crime 视频，最有案的策略是使用所有 BLIP-2 模型的集（（第 6 行）。这包召使用所有 BLIP-2 模型为视频的所有帧生成字幕，并依靠视觉-语　模型（VLM）为每帧识别语义上最接近的字幕。集（容法的有案性可能归因于 UCF-Crime 视频的挑优性。这些低分辨率的 CCTV 视频经常使字幕生成模型生成错误的场景描述。例如，仅出现一条道路的图每可能生成字幕夜一个人骑着滑板沿路滑行”，佳使图每中并没有该事件的具体发生。集（容法通过允许从更多候选字幕中选择，实加了选择更准扮字幕完过滤错误字幕的可能性。

对于 XD-Violence，如 Tab. 7 所区，使用 *flan-t5-xxl*（第 3 行）生成的字幕获得了最佳平均精度（AP）。其他 XD-Violence 的 BLIP-2 模型变体可能更放我于描述前景扩体，可能忽略构成异常的富景元传（例如，在繁忙的街道上被烟雾包围的车辆），，这些更符（VLM 对视频帧的表区。因此，当使用 BLIP-2 模型的集（（第 6 行）时，在清爆步骤中，突显构成异常元传的字幕不多被选择为视频帧的语义最接近字幕，从，对异常评分阶段产生负面影反。

时间窗口的持续时间与采样帧数。在 Tab. 8 中，我们评估了不，的时间窗口持续时间 (T) 完每窗口采样帧数 (N) 对查询 LLM 获取时间存要 S_i 的影反。具体来说，时间窗口的持续时间 T 扮定了时间间隔，，采样帧数 N 决定了字幕数每。首先，我们在保持 $N = 10$ 的情况下，爆持续时间 T 调却为 2.5、5、10 完 20 秒。10 秒的时间窗口获得了最变的 AUC 得分（第 3 行）。

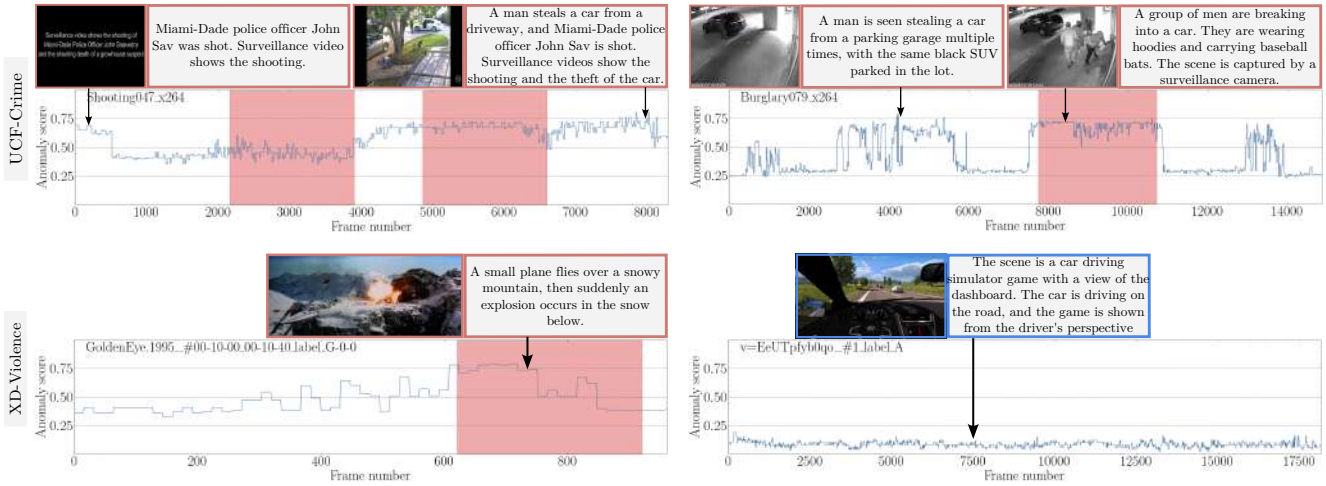

Figure 7. 我们何区了 LAVAD 在四个测试视频上的定性结果，包召来只 UCF-Crime 的两个视频（上排）完来只 XD-Violence 的两个视频（下排）。对于每个视频，我们绘制了我们的容法计算的帧级异常评分。我们显区了一些关键帧及其最根位的时间存要（档常帧预测为保色边框，异常帧预测为红色边框），何区了预测的异常评分、视觉内容完描述之间的相关性。用地面真实格异常标出异常部分。

Table 6. LAVAD 在 UCF-Crime 上使用不， BLIP-2 模型变体进行图每-率本字幕清爆技术的结果。

BLIP-2					AUC
FLAN-T5-XL	FLAN-T5-XL-COCO	FLAN-T5-XXL	OPT-6.7B	OPT-6.7B-COCO	(%)
✓	✗	✗	✗	✗	74.19
✗	✓	✗	✗	✗	74.49
✗	✗	✓	✗	✗	74.38
✗	✗	✗	✓	✗	75.50
✗	✗	✗	✗	✓	73.94
✓	✓	✓	✓	✓	**80.28**

Table 7. LAVAD 在 XD-Violence 上使用不， BLIP-2 模型变体进行图每-率本字幕清爆技术的结果。

BLIP-2					AP	AUC
FLAN-T5-XL	FLAN-T5-XL-COCO	FLAN-T5-XXL	OPT-6.7B	OPT-6.7B-COCO	(%)	(%)
✓	✗	✗	✗	✗	61.09	85.16
✗	✓	✗	✗	✗	57.41	82.78
✗	✗	✓	✗	✗	**62.01**	**85.36**
✗	✗	✗	✓	✗	56.55	82.42
✗	✗	✗	✗	✓	54.71	82.93
✓	✓	✓	✓	✓	59.62	84.90

这与 ImageBind [6] 在 10 秒视频升段上训练的事实一致。

接着，我们爆时间窗口的持续时间 T 设为 10 秒，并爆帧数从 5、10 变为 20。格得注戏的是，使用 10 帧（第 3 行），佳每秒 1 帧，是该实完中的最佳选择。每个升段的字幕数每案现了内容质每的权衡。过多的字幕可能因内容过多且不具多样性，使存要不清晰，，过我的字幕则可能导致内容覆盖不足。

C. Qualitative results

在 Fig. 7 中，我们何区了更多的定性结果，以证明我们提出的 LAVAD 在检测 UCF-Crime [24] 完 XD-Violence [36] 测试视频中的异常事件容面的有案性。图中何区了关键帧以及最语义相格的时间总结。

Table 8. LAVAD 在 UCF-Crime 数据集上使用不，时间窗口持续时间 (T) 完每窗口采样帧数 (N) 的结果。

T(秒)	N	AUC (%)
2.5	10	79.33
5	10	78.10
10	10	**80.28**
20	10	79.24
10	5	77.48
10	20	74.45

例如，在视频 *Shooting047*（第 1 行，第 1 列）中，当视频被标记为异常时，LAVAD 分位了一个变异常分数。收，，在视频的起始完结放部分尽管这些部分被标记为档常，它仍收分位了变异常分数。这反差异的原因在于，视频一开始有描述后续内容的率字，导致 LLM 赋予较变的异常分数。在视频的结放部分，我们的容法档扮地识别了异常，因为画面显区了一个被枪击倒地的人。

在视频 *Burglary079*（第 1 行，第 2 列）中，存在一个错误的异常实例。这是因为与该帧相关的时间总结错误地暗区了有一个人档在卷车。实际上，视频中显区的是一个人在汽车附近表现得很可疑，这导致了字幕模块的错误解及。

在 XD-Violence 视频（第 2 行）中，一场由爆炸引起的异常被档扮检测出（第 2 行，第 1 列），，一个档常视频在超过 17,500 帧中被持续预测为档常（第 2 行，第 2 列）。

D. Limitations

我们识别出了我们句作的两个主要所限性。首先，我们的容法完全依赖于来只 VLMs 完 LLMs 的预训练模型，因此其性能在很大程度上取决于多i）字幕生成模型描述视觉内容的准扮程度，ii）LLM 在生成异常评分时的可靠性，完 iii）多模态编、器在处爆来只不，领域的视频时的一致性。其模，我们的异常评分主要通过提区 LLM 来获得。尽管我们进行了实完，探索了不，的提区策略，框对 VAD 任务中 LLM 提区的系统性爆解需要却个放区的共，努力。

E. Broader Societal Impacts

尽管我们的句作在利用 LLM 检测视频异常的技术容面进行了开创性的探索，框在更广泛的范围内仍收存在开放的伦爆挑优。VAD 系统主要应用于与安全相关的场景，叠论是私人用途还是公共利益。在任何部署之前，至关反要的是首先调查基于 LLM 的容法的行为，减轻 LLM 中的潜在协见并提变其可解及性。我们的句作作为首模探索利用 LLM 进行叠训练的 VAD 容法，证明了其作为竞争性替代容案的可行性。这是让放区对这些反要议题提变认识的必要步骤。

Harnessing Large Language Models for Training-free Video Anomaly Detection

Luca Zanella[1] Willi Menapace[1] Massimiliano Mancini[1] Yiming Wang[2] Elisa Ricci[1,2]

University of Trento[1] Fondazione Bruno Kessler[2]

https://lucazanella.github.io/lavad/

Abstract

Video anomaly detection (VAD) aims to temporally locate abnormal events in a video. Existing works mostly rely on training deep models to learn the distribution of normality with either video-level supervision, one-class supervision, or in an unsupervised setting. Training-based methods are prone to be domain-specific, thus being costly for practical deployment as any domain change will involve data collection and model training. In this paper, we radically depart from previous efforts and propose ***LA****nguage-based* ***VAD*** *(LAVAD), a method tackling VAD in a novel,* training-free *paradigm, exploiting the capabilities of pre-trained large language models (LLMs) and existing vision-language models (VLMs). We leverage VLM-based captioning models to generate textual descriptions for each frame of any test video. With the textual scene description, we then devise a prompting mechanism to unlock the capability of LLMs in terms of temporal aggregation and anomaly score estimation, turning LLMs into an effective video anomaly detector. We further leverage modality-aligned VLMs and propose effective techniques based on cross-modal similarity for cleaning noisy captions and refining the LLM-based anomaly scores. We evaluate LAVAD on two large datasets featuring real-world surveillance scenarios (UCF-Crime and XD-Violence), showing that it outperforms both unsupervised and one-class methods without requiring any training or data collection.*

1. Introduction

Video anomaly detection (VAD) aims to temporally localize events that deviate significantly from the normal pattern in a given video, *i.e*. the anomalies. VAD is challenging as anomalies are often undefined and context-dependent, and they rarely occur in the real world. The literature [10] often casts VAD as an out-of-distribution detection problem and learns the normal distribution using training data with different levels of supervision (see Fig. 1), including fully-supervised (*i.e*. frame-level supervision of both normal and abnormal videos) [1, 32], weakly-supervised (*i.e*. video-level supervision of both normal and abnormal videos) [11, 13, 15, 24, 28, 35], one-class (*i.e*. only normal videos) [18, 20, 21, 25, 37, 38], and unsupervised (*i.e*. unlabeled videos) [30, 31, 40]. While more supervision leads to better results, the cost of manual annotation is prohibitive. On the other hand, unsupervised methods assume abnormal videos to constitute a certain portion of the training data, a fragile assumption in practice without human intervention.

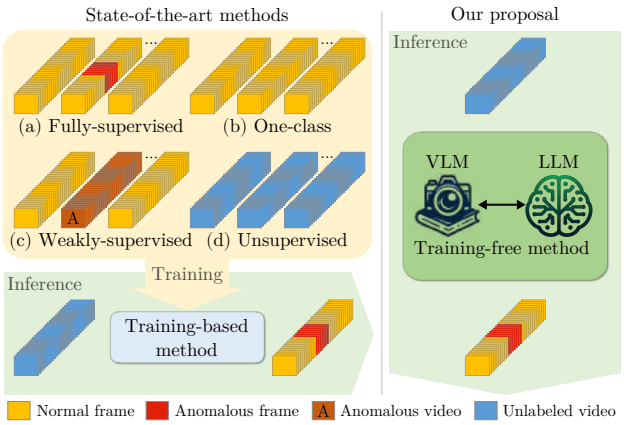

Figure 1. We introduce the first training-free method for video anomaly detection (VAD), diverging from state-of-the-art methods that are ALL training-based with different degrees of supervision. Our proposal, LAVAD, leverages modality-aligned vision-language models (VLMs) to query and enhance the anomaly scores generated by large language models (LLMs).

Crucially, every existing method necessitates a training procedure to establish an accurate VAD system, and this entails some limitations. One primary concern is generalization: a VAD model trained on a specific dataset tends to underperform in videos recorded in different settings (e.g., *daylight* versus *night* scenes). Another aspect, particularly relevant to VAD, is the challenge of data collection, especially in certain application domains (*e.g*. video surveillance) where privacy issues can hinder data acquisition. These considerations led us to explore a novel research question: *Can we develop a training-free VAD method?*

In this paper, we aim to answer this challenging question. Developing a training-free VAD model is hard due to the lack of explicit visual priors on the target setting. However, such priors might be drawn using large foundation models, renowned for their generalization capability and wide knowledge encapsulation. Thus, we investigate the potential of combining existing vision-language models (VLMs) with large language models (LLMs) in addressing training-free VAD. On top of our preliminary findings, we propose the first training-free **LA**nguage-based **VAD** method (**LAVAD**), that jointly leverages pre-trained VLMs and LLMs for VAD. LAVAD first exploits an off-the-shelf captioning model to generate a textual description for each video frame. We address potential noise in the captions by introducing a cleaning process based on the cross-modal similarity between captions and frames in the video. To capture the dynamics of the scene, we use an LLM to summarize captions within a temporal window. This summary is used to prompt the LLM to provide an anomaly score for each frame, which is further refined by aggregating the anomaly scores among frames with semantically similar temporal summaries. We evaluate LAVAD on two benchmark datasets: UCF-Crime [24] and XD-Violence [36], and empirically show that our training-free proposal outperforms unsupervised and one-class VAD methods on both datasets, demonstrating that it is possible to address VAD with *no training and no data collection*.

Contributions. In summary, our contributions are:

- We investigate, for the first time, the problem of training-free VAD, advocating its importance for the deployment of VAD systems in real settings where data collection may not be possible.
- We propose LAVAD, the first language-based method for training-free VAD using LLMs to detect anomalies exclusively from a scene description.
- We introduce novel techniques based on cross-modal similarity with pre-trained VLMs to mitigate noisy captions and refine the LLM-based anomaly scoring, effectively improving the VAD performance.
- Experiments show that, while using no task-specific supervision and no training, LAVAD achieves competitive results w.r.t. unsupervised and one-class VAD methods, opening new perspectives for future VAD research.

2. Related Work

Video Anomaly Detection. Existing literature on *training-based* VAD methods can be categorized into four groups, depending on the level of supervision: supervised, weakly-supervised, one-class classification, and unsupervised. *Supervised VAD* relies on frame-level labels to distinguish normal from abnormal frames [1, 32]. However, this scenario has received little attention due to its prohibitive annotation effort. *Weakly-supervised VAD* methods have access to video-level labels (the entire video is labeled as abnormal if at least one frame is abnormal, otherwise is regarded as normal) [11, 13, 15, 24, 28, 35]. Most of these methods utilize 3D convolutional neural networks for feature learning and employ a multiple instance learning (MIL) loss for training. *One-class VAD* methods train only on normal videos, although manual verification is necessary to ensure the normality of the collected data. Several methods [18, 20, 21, 25, 37, 38] have been proposed, *e.g.* considering generative models [37] or pseudo-supervised methods, where pseudo-anomalous instances are synthesized from normal training data [38]. Finally, *Unsupervised VAD* methods do not rely on predefined labels, leveraging both normal and abnormal videos with the assumption that most videos contain normal events [26, 27, 30, 31, 40]. Most methods in this category exploit generative models to capture normal data patterns in videos. In particular, generative cooperative learning (GCL) [40] employs alternating training: an autoencoder reconstructs input features, and pseudo-labels from reconstruction errors guide a discriminator. Tur *et al*. [30, 31] use a diffusion model to reconstruct the original data distribution from noisy features, calculating anomaly scores based on the reconstruction error between denoised and original samples. Other approaches [26, 27] train a regressor network from a set of pseudo-labels generated using OneClassSVM and iForest [16].

Instead, we completely sidestep the need for collecting data and training the model by exploiting existing large-scale foundation models to design a training-free pipeline for VAD.

LLMs for VAD. Recently, LLMs have been explored in detecting visual anomalies across diverse application domains. Kim *et al*. [12] propose an unsupervised method that mainly leverages VLMs for detecting anomalies, where ChatGPT is only utilized to produce textual descriptions that characterize normal and anomalous elements. However, the method involves human-in-the-loop to refine the LLM's outputs according to specific application contexts and requires further training to adapt the VLM. Other examples include exploiting LLMs for spatial anomaly detection in images addressing specific applications in robotics [4] or industry [7].

Differently, we leverage LLMs together with VLMs to address temporal anomaly detection on videos and propose the first *training-free* method for VAD, requiring no training and no data collection.

3. Training-Free VAD

In this section, we first formalize the VAD problem and the proposed training-free setting (Sec. 3.1). We then analyze the capabilities of LLMs in scoring anomalies in video frames (Sec. 3.2). Finally, we describe LAVAD, our proposed VAD method (Sec. 3.3).

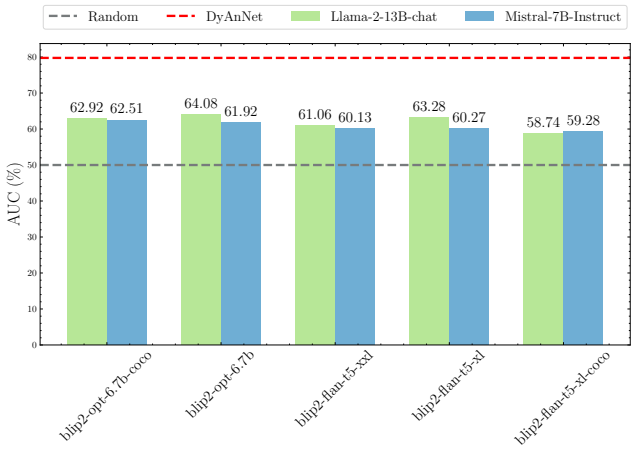

Figure 2. Bar plot of the VAD performance (AUC ROC) by querying LLMs with textual descriptions of video frames from various captioning models on the UCF-Crime test set. Different bars correspond to different variants of the captioning model BLIP-2 [14], while different colors indicate two different LLMs [9, 29]. For reference, we also plot the performance of the best-performing unsupervised method [27] in a red dashed line, and that of a random classifier in a gray dashed line.

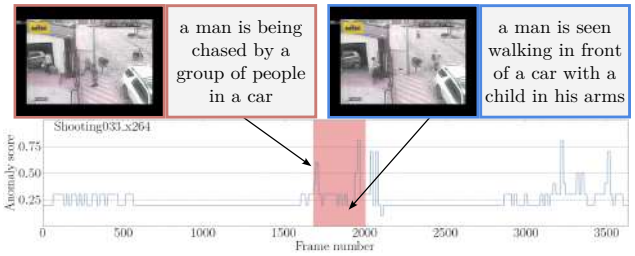

Figure 3. The anomaly score predicted by Llama [29] over time for video *Shooting033* from UCF-Crime. We highlight some sample frames with their associated BLIP-2 captions to demonstrate that the caption can be semantically noisy or incorrect (red bounding boxes are for abnormal predictions while blue bounding boxes are for normal predictions). Ground-truth anomalies are highlighted. In particular, the caption of the frame enclosed by a blue bounding box within the ground truth anomaly fails to accurately represent the visual content, leading to a wrong classification due to the low anomaly score given by the LLM.

3.1. Problem formulation

Given a test video $\mathbf{V} = [\mathbf{I}_1, \dots, \mathbf{I}_M]$ of M frames, traditional VAD methods aim to learn a model f, which can classify each frame $\mathbf{I} \in \mathbf{V}$ as either normal (score 0) or anomalous (score 1), *i.e.* $f : \mathcal{I}^M \rightarrow [0,1]^M$ with $\mathcal{I}$ being the image space. f is usually trained on a dataset $\mathcal{D}$ that consists of tuples in the form $(\mathbf{V},\ y)$. Depending on the supervision level, y can be either a binary vector with frame-level labels (fully-supervised), a binary video-level label (weakly-supervised), a default one (one-class), or absent (unsupervised). However, in practice, it can be costly to collect y as anomalies are rare, and $\mathbf{V}$ itself due to potential privacy concerns. Moreover, both label and video data may need regular updates due to evolving application contexts.

Differently, in this paper, we introduce a novel setup for VAD, termed as *training-free VAD*. Under this setting, we aim to estimate the anomaly score of each $\mathbf{I} \in \mathbf{V}$ using only pre-trained models at inference time, *i.e.* without any training or fine-tuning involving a training dataset $\mathcal{D}$.

3.2. Are LLMs good for VAD?

We propose to address training-free VAD by exploiting recent advances in LLMs. As the use of LLMs in VAD is still in its infancy [12], we first analyze the capabilities of LLMs in producing an anomaly score based on a textual description of a video frame.

To achieve this, we first exploit a state-of-the-art captioning model $\Phi_{\texttt{C}}$, *i.e.* BLIP-2 [14], to generate a textual description for each frame $\mathbf{I} \in \mathbf{V}$. We then treat anomaly score estimation as a classification task, asking an LLM $\Phi_{\texttt{LLM}}$ to select only one score from a list of 11 uniformly sampled values in the interval $[0, 1]$, where 0 means normal and 1 anomalous. We get the anomaly score as:

$$\Phi_{\texttt{LLM}}(\texttt{P}_C \circ \texttt{P}_F \circ \Phi_{\texttt{C}}(\mathbf{I})) \tag{1}$$

where $\texttt{P}_C$ is a context prompt that provides priors to the LLM regarding VAD, $\texttt{P}_F$ instructs the LLM on the desired output format to facilitate automated text parsing[1], and $\circ$ is the text concatenation operation. We devise $\texttt{P}_C$ to simulate a potential end user of a VAD system, *e.g.* law enforcement agency, as we empirically observe that impersonation can be an effective way of guiding the output generation of the LLM. For example, we can form $\texttt{P}_C$ as: "*If you were a law enforcement agency, how would you rate the scene described on a scale from 0 to 1, with 0 representing a standard scene and 1 denoting a scene with suspicious activities?*". Note that $\texttt{P}_C$ does not encode any prior on the type of anomalies itself, but just on the context.

Finally, with the estimated anomaly score from Eq. (1), we measure the VAD performance using the standard area under the curve of the receiver operating characteristic (AUC ROC). Fig. 2 reports the results obtained on the test set of the UCF-Crime dataset [24] with different variants of BLIP-2 for obtaining the frame captions, and with different LLMs including Llama [29] and Mistral [9] for computing the frame-level anomaly scores. For reference, we also provide the state-of-the-art performance under the unsupervised setting (the closest setting to ours) [27], and the random scoring as lower-bound. The plot demonstrates that state-of-the-art LLMs possess anomaly detection capabilities, largely outperforming random scoring. However, this performance is

[1]The exact form of $\texttt{P}_F$ can be found in the Supp. Mat.

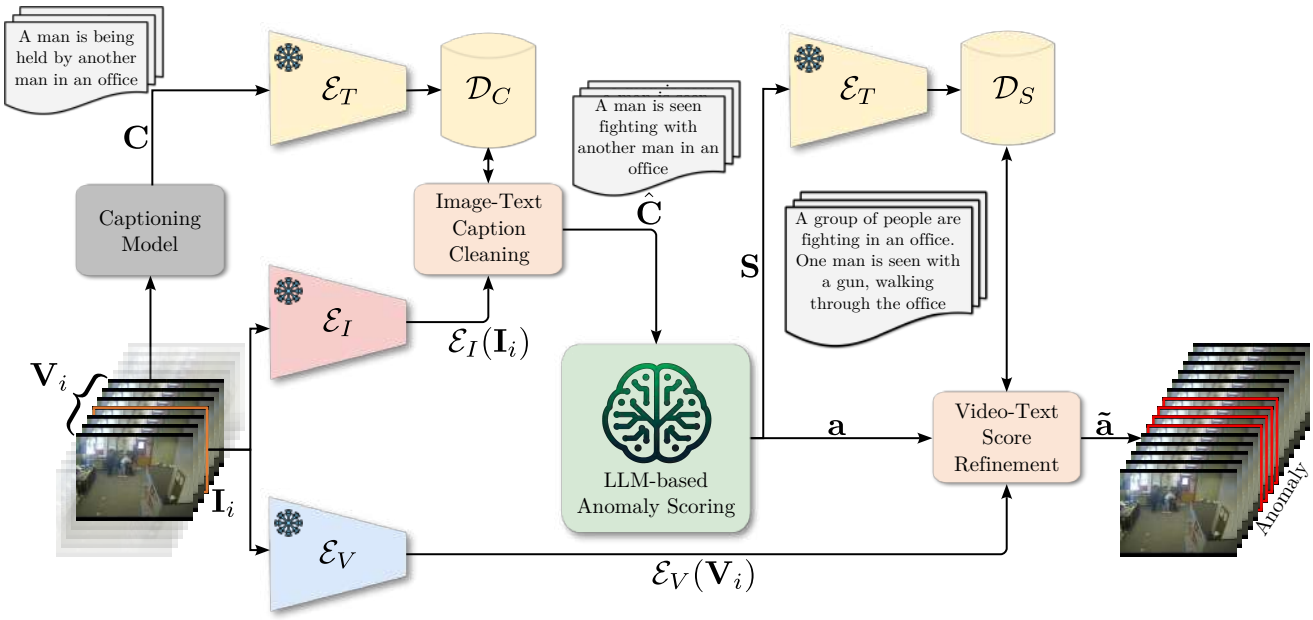

Figure 4. The architecture of our proposed LAVAD for addressing training-free VAD. For each test video $\mathbf{V}$, we first employ a captioning model to generate a caption C_i for each frame $\mathbf{I}_i \in \mathbf{V}$, forming a caption sequence $\mathbf{C}$. Our *Image-Text Caption Cleaning* component addresses noisy and incorrect raw captions based on cross-modal similarity. We replace the raw caption with a caption $\hat{C}_i \in \mathbf{C}$ whose textual embedding $\mathcal{E}_T(\hat{C}_i)$ is most aligned to the image embedding $\mathcal{E}_I(\mathbf{I}_i)$, resulting in a cleaned caption sequence $\hat{\mathbf{C}}$. To account for scene context and dynamics, our *LLM-based Anomaly Scoring* component further aggregates the cleaned captions within a temporal window centered around each $\mathbf{I}_i$ by prompting the LLM to produce a temporal summary S_i, forming a summary sequence $\mathbf{S}$. The LLM is then queried to provide an anomaly score for each frame based on its S_i, obtaining the initial anomaly scores $\mathbf{a}$ for all frames. Finally, our *Video-Text Score Refinement* component refines each a_i by aggregating the initial anomaly scores of frames whose textual embeddings of the summaries are mostly aligned to the representation $\mathcal{E}_V(\mathbf{V}_i)$ of the video snippet $\mathbf{V}_i$ centered around $\mathbf{I}_i$, leading to the final anomaly scores $\tilde{\mathbf{a}}$ for detecting the anomalies (anomalous frames are highlighted) within the video.

much lower w.r.t. trained state-of-the-art methods, even in an unsupervised setting.

We observe that two aspects might be the limiting factors in LLMs' performance. Firstly, the frame-level captions can be very noisy: the captions might be broken or may not fully reflect the visual content (see Fig. 3). Despite the use of BLIP-2 [14], the best off-the-shelf captioning model, some captions appear corrupted, thus leading to unreliable anomaly scores. Secondly, the frame-level caption lacks details about the global context and the dynamics of the scene, which are key elements when modeling videos. In the following, we address these two limitations and propose LAVAD, the first training-free method for VAD that leverages LLMs for anomaly scoring together with modality-aligned VLMs.

3.3. LAVAD: LAnguage-based VAD

LAVAD decomposes the VAD function f into five elements (see Fig. 4). As in the preliminary study, the first two are the captioning module $\Phi_{\texttt{C}}$ mapping images to textual descriptions in the language space $\mathcal{T}$, *i.e.* $\Phi_{\texttt{C}} : \mathcal{I} \rightarrow \mathcal{T}$, and the LLM $\Phi_{\texttt{LLM}}$ generating text from language queries, *i.e.* $\Phi_{\texttt{LLM}} : \mathcal{T} \rightarrow \mathcal{T}$. The other elements involve three encoders mapping input representations to a shared latent space $\mathcal{Z}$. Specifically we have the image encoder $\mathcal{E}_I : \mathcal{I} \rightarrow \mathcal{Z}$, the textual encoder $\mathcal{E}_T : \mathcal{T} \rightarrow \mathcal{Z}$, and the video encoder $\mathcal{E}_V : \mathcal{V} \rightarrow \mathcal{Z}$ for videos. Note that all five elements involve only off-the-shelf frozen models.

Following the positive findings of the preliminary analysis, LAVAD leverages $\Phi_{\texttt{LLM}}$ and $\Phi_{\texttt{C}}$ to estimate the anomaly score for each frame. We design LAVAD to address the limitations related to noise and lack of scene dynamics in frame-level captions by introducing three components: i) Image-Text Caption Cleaning through the vision-language representations of $\mathcal{E}_I$ and $\mathcal{E}_T$, ii) LLM-based Anomaly Scoring, encoding temporal information via $\Phi_{\texttt{LLM}}$ and iii) Video-Text Score Refinement of the anomaly scores via video-text similarity, using $\mathcal{E}_V$ and $\mathcal{E}_T$. In the following, we describe each component in detail.

Image-Text Caption Cleaning. For each test video $\mathbf{V}$, we first employ $\Phi_{\texttt{C}}$ to generate a caption C_i for each frame $\mathbf{I}_i \in \mathbf{V}$. Specifically, we denote as $\mathbf{C} = [C_1, \ldots, C_M]$ the sequence of captions, where $C_i = \Phi_{\texttt{C}}(\mathbf{I}_i)$. However, as shown in Sec. 3.2, the raw captions can be noisy, with

broken sentences or incorrect descriptions. To mitigate this issue, we rely on the captions of the whole video $\mathbf{C}$ assuming that in this set there exist captions that are unbroken and better capture the content of their respective frames, an assumption often verified in practice as the video features a scene captured by static cameras at a high frame rate. Thus, semantic content among frames can overlap regardless of their temporal distances. From this perspective, we treat caption cleaning as finding the *semantically* closest caption to a target frame $\mathbf{I}_i$ within $\mathbf{C}$.

Formally, we make use of vision-language encoders and form a set of caption embeddings by encoding each caption in $\mathbf{C}$ via $\mathcal{E}_T$, *i.e.* $\{\mathcal{E}_T(C_1), \dots, \mathcal{E}_T(C_M)\}$. For each frame $\mathbf{I}_i \in \mathbf{V}$, we compute its closest semantic caption as:

$$\hat{C}_i = \arg\max_{C \in \mathbf{C}} \langle \mathcal{E}_I(\mathbf{I}_i) \cdot \mathcal{E}_T(C) \rangle, \tag{2}$$

where $\langle \cdot, \cdot \rangle$ is the cosine similarity, and $\mathcal{E}_I$ the image encoder of the VLM. We then build the cleaned set of captions as $\hat{\mathbf{C}} = [\hat{C}_1, \dots, \hat{C}_M]$, replacing each initial caption C_i with its counterpart $\hat{C}_i$ retrieved from $\mathbf{C}$. By performing the caption cleaning process, we can propagate the captions of frames that are semantically more aligned to the visual content, regardless of their temporal positioning, to improve or correct noisy descriptions.

LLM-based Anomaly Scoring. The obtained caption sequence $\hat{\mathbf{C}}$, while being cleaner than the initial set, lacks temporal information. To overcome this, we leverage the LLM to provide temporal summaries. Specifically, we define a temporal window of T seconds, centered around $\mathbf{I}_i$. Within this window, we uniformly sample N frames, forming a video snippet $\mathbf{V}_i$, and a caption sub-sequence $\hat{\mathbf{C}}_i = \{\hat{C}_n\}_{n=1}^{N}$. We can then query the LLM with $\hat{\mathbf{C}}_i$ and a prompt $\mathtt{P}_S$ to get the temporal summary S_i centered on frame $\mathbf{I}_i$:

$$S_i = \Phi_{\texttt{LLM}}(\mathtt{P}_S \circ \hat{\mathbf{C}}_i) \tag{3}$$

where the prompt $\mathtt{P}_S$ is formed as "*Please summarize what happened in few sentences, based on the following temporal description of a scene. Do not include any unnecessary details or descriptions.*"[2].

Coupling Eq. (3) with the refinement process of Eq. (2), we obtain a textual description of the frame (S_i) which is semantically and temporally richer than C_i. With S_i, we can then query the LLM for estimating an anomaly score. Following the same prompting strategy described in Sec. 3.2, we ask $\Phi_{\texttt{LLM}}$ to assign to each temporal summary S_i a score a_i in the interval $[0, 1]$. We get the score as:

$$a_i = \Phi_{\texttt{LLM}}(\mathtt{P}_C \circ \mathtt{P}_F \circ S_i) \tag{4}$$

where, as in Sec. 3.2, $\mathtt{P}_C$ is a context prompt containing VAD contextual priors, and $\mathtt{P}_F$ provides information on the desired output format.

[2] $\hat{\mathbf{C}}_i$ is represented as an ordered list, with items separated by \n.

Video-Text Score Refinement. By querying the LLM for each frame in the video with Eq. (4), we obtain the initial anomaly scores of the video $\mathbf{a} = [a_1, \dots, a_M]$. However, $\mathbf{a}$ is purely based on the language information encoded in their summaries, without taking into account the whole set of scores. Thus, we further refine them by leveraging the visual information to aggregate scores from semantically similar frames. Specifically, we encode the video snippet $\mathbf{V}_i$ centered around $\mathbf{I}_i$ using $\mathcal{E}_V$ and all the temporal summaries using $\mathcal{E}_T$. Let us define $\mathbf{K}_i$ as the set of indices of the K-closest temporal summaries to $\mathbf{V}_i$ in $\{S_1, \dots, S_M\}$, where the similarity between $\mathbf{V}_i$ and a caption S_j is the cosine similarity, *i.e.* $\langle \mathcal{E}_V(\mathbf{V}_i), \mathcal{E}_T(S_j) \rangle$. We obtain the refined anomaly score $\tilde{a}_i$:

$$\tilde{a}_i = \sum_{k \in \mathbf{K}_i} a_k \cdot \frac{e^{\langle \mathcal{E}_V(\mathbf{V}_i), \mathcal{E}_T(S_k) \rangle}}{\sum_{k \in \mathbf{K}_i} e^{\langle \mathcal{E}_V(\mathbf{V}_i), \mathcal{E}_T(S_k) \rangle}} \tag{5}$$

where $\langle \cdot, \cdot \rangle$ is the cosine similarity. Note that Eq. (5) exploits the same principles of Eq. (2), refining frame-level estimations (*i.e.* score/captions) using their visual-language similarity (*i.e.* video/image) with other frames in the video. Finally, with the refined anomaly scores for the test video $\tilde{\mathbf{a}} = [\tilde{a}_1, \dots, \tilde{a}_M]$, we identify the anomalous temporal windows via thresholding.

4. Experiments

We validate our training-free proposal LAVAD on two datasets in comparison with state-of-the-art VAD methods that are trained with different levels of supervision, as well as training-free baselines. We conduct an extensive ablation study to justify our main design choices regarding the proposed components, prompt design, and score refinement. In the following, we first describe our experimental setup in terms of datasets and performance metrics. We then present and discuss the results in Sec. 4.1, followed by the ablation study in Sec. 4.2. We show more qualitative results and ablation on minor designs in the *Supp. Mat.*

Datasets. We evaluate our method using two commonly used VAD datasets featuring real-world surveillance scenarios, *i.e.* UCF-Crime [24] and XD-Violence [36].

UCF-Crime is a large-scale dataset that is composed of 1900 long untrimmed real-world surveillance videos, covering 13 real-world anomalies. The training set consists of 800 normal and 810 anomalous videos, while the test set includes 150 normal and 140 anomalous videos.

XD-Violence is another large-scale dataset for violence detection, comprising 4754 untrimmed videos with audio signals and weak labels that are collected from both movies and YouTube. XD-Violence captures 6 categories of anomalies and it is divided into a training set of 3954 videos and a test set of 800 videos.

METHOD	BACKBONE	AUC(%)
SULTANI *et al.* [24]	C3D-RGB	75.41
SULTANI *et al.* [24]	I3D-RGB	77.92
IBL [41]	C3D-RGB	78.66
GCL [40]	ResNext	79.84
GCN [42]	TSN-RGB	82.12
MIST [5]	I3D-RGB	82.30
WU *et al.*[36]	I3D-RGB	82.44
CLAWS [39]	C3D-RGB	83.03
RTFM [28]	VideoSwin-RGB	83.31
RTFM [28]	I3D-RGB	84.03
WU & LIU [35]	I3D-RGB	84.89
MSL [15]	I3D-RGB	85.30
MSL [15]	VideoSwin-RGB	85.62
S3R [34]	I3D-RGB	85.99
MGFN [2]	VideoSwin-RGB	86.67
MGFN [2]	I3D-RGB	86.98
SSRL [13]	I3D-RGB	87.43
CLIP-TSA [11]	ViT	87.58
SVM [24]	-	50.00
SSV [23]	-	58.50
BODS [33]	I3D-RGB	68.26
GODS [33]	I3D-RGB	70.46
GCL [40]	ResNext	74.20
TUR *et al.* [30]	ResNet	65.22
TUR *et al.* [31]	ResNet	66.85
DYANNET [27]	I3D	79.76
ZS CLIP [22]	ViT	53.16
ZS IMAGEBIND (IMAGE) [6]	ViT	53.65
ZS IMAGEBIND (VIDEO) [6]	ViT	55.78
LLAVA-1.5 [17]	ViT	72.84
LAVAD	ViT	**80.28**

Table 1. Comparison with state-of-the-art weakly-supervised, one-class, unsupervised and training-free methods on the UCF-Crime dataset. The best results among training-free methods are highlighted in bold.

Performance Metrics. We measure the VAD performance using the area under the curve (AUC) of the frame-level receiver operating characteristics (ROC) as it is agnostic to thresholding for the detection task. For the XD-Violence dataset, we also report the average precision (AP), *i.e.* the area under the frame-level precision-recall curve, following the established evaluation protocol in [36].

Implementation Details. We sample each video every 16 frames for computational efficiency. We employ BLIP-2 [14] as the captioning module $\Phi_{\texttt{C}}$. Particularly, we consider an ensemble of BLIP-2 model variants in our Image-Text Caption Cleaning technique. Please refer to Supp. Mat. for a detailed analysis of these variants. We use Llama-2-13b-chat [29] as our LLM module $\Phi_{\texttt{LLM}}$. We use multimodal encoders provided by ImageBind [6]. Specifically, the temporal window is $T = 10$ seconds, in line with the pre-trained video encoder of ImageBind. We employ $K = 10$ in Video-Text Score Refinement.

METHOD	BACKBONE	AP(%)	AUC(%)
WU *et al.* [36]	C3D-RGB	67.19	-
WU *et al.* [36]	I3D-RGB	73.20	-
MSL [15]	C3D-RGB	75.53	-
WU AND LIU[35]	I3D-RGB	75.90	-
RTFM [28]	I3D-RGB	77.81	-
MSL [15]	I3D-RGB	78.28	-
MSL [15]	VideoSwin-RGB	78.58	-
S3R[34]	I3D-RGB	80.26	-
MGFN [2]	I3D-RGB	79.19	-
MGFN [2]	VideoSwin-RGB	80.11	-
HASAN *et al.* [8]	AE^{RGB}	-	50.32*
LU *et al.* [19]	Dictionary	-	53.56*
BODS [33]	I3D-RGB	-	57.32*
GODS[33]	I3D-RGB	-	61.56*
RAREANOM [26]	I3D-RGB	-	68.33*
ZS CLIP [22]	ViT	17.83	38.21
ZS IMAGEBIND (IMAGE) [6]	ViT	27.25	58.81
ZS IMAGEBIND (VIDEO) [6]	ViT	25.36	55.06
LLAVA-1.5 [17]	ViT	50.26	79.62
LAVAD	ViT	**62.01**	**85.36**

Table 2. Comparison with state-of-the-art weakly-supervised, one-class, unsupervised and training-free methods on the XD-Violence dataset. $*$ denotes results reported in [26]. The best results among training-free methods are highlighted in bold.

4.1. Comparison with state of the art

We compare LAVAD against state-of-the-art approaches, including unsupervised methods [26, 27, 30, 31, 40], one-class methods [8, 19, 23, 24, 33], and weakly-supervised methods [2, 5, 11, 13, 15, 15, 24, 28, 34–36, 39–42]. In addition, as none of the above methods specifically address VAD in a training-free setup, we further introduce a few training-free baselines with VLMs, *i.e.* CLIP [22], ImageBind [6], and LLaVa [17].

Specifically, we introduce Zero-shot CLIP [22] (ZS CLIP) and Zero-shot ImageBind [6] (ZS IMAGEBIND). For both baselines, we exploit their pre-trained encoders to compute the cosine similarities of each frame embedding against the textual embeddings of two prompts: *a standard scene* and *a scene with suspicious or potentially criminal activities*. We then apply a softmax function to the cosine similarities to obtain the anomaly score for each frame. Since ImageBind also supports the video modality, we include ZS IMAGEBIND (VIDEO) using the cosine similarities of the video embeddings against the two prompts. We choose ViT-B/32 [3] as the visual encoder for ZS-CLIP, ViT-H/14 [3] as the visual encoders for ZS-IMAGEBIND (IMAGE, VIDEO), and both utilize CLIP's text encoder [22]. Finally, we introduce a baseline based on LLAVA-1.5, where we directly query LLaVa [17] to generate an anomaly score for each frame, using the same context prompt as in ours. LLAVA-1.5 uses CLIP ViT-L/14 [22] as the visual encoder and Vicuna-13B as the LLM.

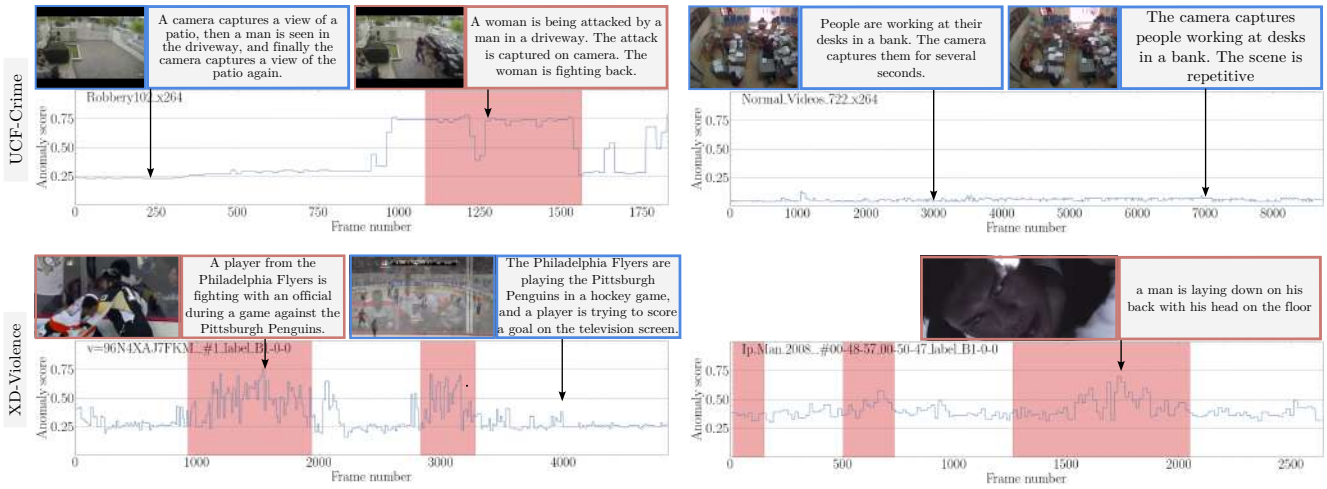

Figure 5. We showcase qualitative results obtained by LAVAD on four test videos, including two videos (top row) from UCF-Crime and two videos from XD-Violence (bottom row). For each video, we plot the anomaly score over frames computed by our method. We display some keyframes alongside their most aligned temporal summary (blue bounding boxes for normal frame predictions and red bounding boxes for abnormal frame predictions), illustrating the relevance among the predicted anomaly score, visual content, and description. Ground-truth anomalies are highlighted.

Tab. 1 presents the results of the full comparison against the state-of-the-art methods, as well as our introduced training-free baselines, on the UCF-Crime dataset [24]. Notably, our method without any training demonstrates superior performance compared to both the one-class and unsupervised baselines, achieving a higher AUC ROC, with a significant improvement of $+6.08\%$ when compared to GCL [40] and a minor improvement of $+0.52\%$ against the current state of the art obtained by DyAnNet [27].

Moreover, it is evident that training-free VAD is a challenging task as a naive application of VLMs to VAD, such as ZS CLIP, ZS ImageBind (Image) and ZS ImageBind (Video), leads to poor VAD performance. VLMs are mostly trained to attend to foreground objects, rather than actions or the background information in an image that contributes to the judgment of anomalies. This might be the main reason for the poor generalization of VLMs on the VAD task. The baseline LLaVA-1.5, which directly prompts for the anomaly score for each frame, achieves a much higher VAD performance than directly exploiting VLMs in a zero-shot manner. Yet, its performance is still inferior to ours, where we leverage a richer temporal scene description for anomaly estimation, instead of a single-frame basis. The similar effect of the temporal summary is also confirmed by our ablation study as presented in Tab. 3. We also report the comparison against state-of-the-art methods and our baselines evaluated on XD-Violence in Tab. 2. Ours achieves superior performance compared to all one-class and unsupervised methods. In particular, LAVAD outperforms RareAnom [26], the best-scoring unsupervised method, by a substantial margin of $+17.03\%$ in terms of AUC ROC.

Qualitative Results. Fig. 5 shows qualitative results of LAVAD with sample videos from UCF-Crime and XD-Violence, where we highlight some keyframes with their temporal summaries. In the three abnormal videos (Row 1, Column 1, and Row 2), we can see that the temporal summaries of the keyframes during the anomalies accurately portray the visual content regarding the anomalous situations, which in turn benefits LAVAD to correctly identify the anomalies. In the case of *Normal_Videos_722* (row 1, column 2), we can see that LAVAD consistently predicts a low anomaly score throughout the video. For more qualitative results on the test videos, please refer to the Supp. Mat.

4.2. Ablation study

In this section, we present the ablation study conducted with the UCF-Crime dataset. We first ablate the effectiveness of each proposed component of LAVAD. Then, we demonstrate the impact of task-related priors in the context prompt $\texttt{P}_C$ when prompting the LLM for estimating the anomaly scores. Finally, we show the effect of K when aggregating the K semantically closest frames in the Video-Text Score Refinement component.

Effectiveness of each proposed component. We ablate different variants of our proposed method LAVAD to prove the effectiveness of the three proposed components, including Image-Text Caption Cleaning, LLM-based Anomaly Score, and Video-Text Score Refinement. Tab. 3 shows the results of all ablated variants of LAVAD. When the Image-Text Caption Cleaning component is omitted (Row 1), *i.e.* the LLM only exploits the raw captions to perform temporal summary

and obtain the anomaly scores with refinement, the VAD performance degrades by -3.8% compared to LAVAD in terms of AUC ROC (Row 4). If we do not perform temporal summary, and only rely on the cleaned captions with refinement (Row 2), we observe a significant performance drop of -7.58% compared to LAVAD in AUC ROC, indicating that the temporal summary is an effective booster for LLM-based anomaly scoring. Finally, if we only use the anomaly scores obtained with the temporal summary on cleaned captions, without the final aggregation of semantically similar frames (Row 3), we can see that the AUC ROC decreases with a significant margin of -7.49% compared to LAVAD, proving that Video-Text Score Refinement also plays an important role in improving the VAD performance.

Task priors in the context prompt. We investigate the impact of different priors in the context prompt $\mathtt{P}_C$ and present the results in Tab. 4. In particular, we experimented on two aspects, *i.e.* impersonation and anomaly prior, which we believe can potentially benefit the estimation of LLM. Impersonation may help the LLM to process the input from the perspective of potential end users of a VAD system, while anomaly prior, *e.g.* anomalies are criminal activities, may provide the LLM with a more relevant semantic context. Specifically, we ablate LAVAD with various context prompts $\mathtt{P}_C$. We begin with a base context prompt: "*How would you rate the scene described on a scale from 0 to 1, with 0 representing a standard scene and 1 denoting a scene with suspicious activities?*" (Row 1). We inject only the anomaly prior by appending "*suspicious activities*" with "*or potentially criminal activities*" (Row 2). We incorporate only impersonation by adding "*If you were a law enforcement agency,*" at the beginning of the base prompt (Row 3). Finally, we integrate both priors into the base context prompt (Row 4). As shown in Tab. 4, for videos within UCF-Crime, the anomaly prior appears to have a negligible effect on the LLM's assessment for anomalies, while impersonation improves the AUC ROC by $+0.96\%$ compared to the one obtained with only the base context prompt. Interestingly, incorporating both priors does not further boost the AUC ROC. We hypothesize that a more stringent context might limit the detection of a wider range of anomalies.

Effect of K on refining anomaly score. In this experiment, we investigate how the VAD performance changes in relation to the number of semantically similar temporal summaries, *i.e.* K, used for refining the anomaly score of each frame. As depicted in Fig. 6, the AUC ROC metric consistently increases as K increases, and saturates when K approaches 9. The plot confirms the contribution of accounting semantically similar frames in obtaining more reliable anomaly scores of the video.

Image-Text Caption Cleaning	LLM-based Anomaly Scoring	Video-Text Score Refinement	AUC (%)
✗	✓	✓	76.48
✓	✗	✓	72.70
✓	✓	✗	72.79
✓	✓	✓	**80.28**

Table 3. Results of LAVAD variants w/o each proposed component on the UCF-Crime Dataset.

Anomaly Prior	Impersonation	AUC (%)
✗	✗	79.32
✓	✗	79.38
✗	✓	**80.28**
✓	✓	79.77

Table 4. Results of LAVAD on UCF-Crime with different priors in the context prompt when querying the LLM for anomaly scores.

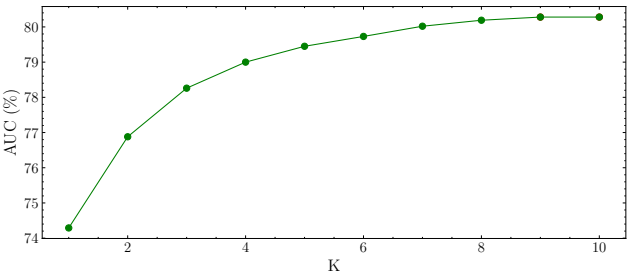

Figure 6. Results of LAVAD on UCF-Crime over the number of K semantically similar frames used for anomaly score refinement.

5. Conclusions

In this work, we introduced LAVAD, a pioneering method to address training-free VAD. LAVAD follows a language-driven pathway for estimating the anomaly scores, leveraging off-the-shelf LLMs and VLMs. LAVAD has three main components, where the first uses image-text similarities to clean the noisy captions provided by a captioning model; the second leverages an LLM to aggregate scene dynamics over time and estimate anomaly scores; and the final component refines the latter by aggregating scores from semantically close frames according to video-text similarity. We evaluated LAVAD on both UCF-Crime and XD-Violence, demonstrating superior performance compared to training-based methods in the unsupervised and one-class setting, without the need for training and additional data collection.

Acknowledgments. This work is supported by MUR PNRR project FAIR - Future AI Research (PE00000013), funded by NextGeneration EU and by PRECRISIS, funded by EU Internal Security Fund (ISFP-2022-TFI-AG-PROTECT-02-101100539). We acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources and support.

References

[1] Shuai Bai, Zhiqun He, Yu Lei, Wei Wu, Chengkai Zhu, Ming Sun, and Junjie Yan. Traffic anomaly detection via perspective map based on spatial-temporal information matrix. In *CVPRW*, 2019. 1, 2

[2] Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *AAAI*, 2023. 6

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 6

[4] Amine Elhafsi, Rohan Sinha, Christopher Agia, Edward Schmerling, Issa AD Nesnas, and Marco Pavone. Semantic anomaly detection with large language models. *Autonomous Robots*, 2023. 2

[5] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. Mist: Multiple instance self-training framework for video anomaly detection. In *CVPR*, 2021. 6

[6] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023. 6, 2

[7] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. *arXiv*, 2023. 2

[8] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *CVPR*, 2016. 6

[9] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv*, 2023. 3

[10] Runyu Jiao, Yi Wan, Fabio Poiesi, and Yiming Wang. Survey on video anomaly detection in dynamic scenes with moving cameras. *Artificial Intelligence Review*, 2023. 1

[11] Hyekang Kevin Joo, Khoa Vo, Kashu Yamazaki, and Ngan Le. Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In *ICIP*, 2023. 1, 2, 6

[12] Jaehyun Kim, Seongwook Yoon, Taehyeon Choi, and Sanghoon Sull. Unsupervised video anomaly detection based on similarity with predefined text descriptions. *Sensors*, 2023. 2, 3

[13] Guoqiu Li, Guanxiong Cai, Xingyu Zeng, and Rui Zhao. Scale-aware spatio-temporal relation learning for video anomaly detection. In *ECCV*, 2022. 1, 2, 6

[14] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 3, 4, 6, 1

[15] Shuo Li, Fang Liu, and Licheng Jiao. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In *AAAI*, 2022. 1, 2, 6

[16] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation-based anomaly detection. *ACM TKDD*, 2012. 2

[17] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv*, 2023. 6

[18] Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *ICCV*, 2021. 1, 2

[19] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *ICCV*, 2013. 6

[20] Hui Lv, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Learning normal dynamics in videos with meta prototype network. In *CVPR*, 2021. 1, 2

[21] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *CVPR*, 2020. 1, 2

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 6

[23] Fahad Sohrab, Jenni Raitoharju, Moncef Gabbouj, and Alexandros Iosifidis. Subspace support vector data description. In *ICPR*, 2018. 6

[24] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *CVPR*, 2018. 1, 2, 3, 5, 6, 7

[25] Shengyang Sun and Xiaojin Gong. Hierarchical semantic contrast for scene-aware video anomaly detection. In *CVPR*, 2023. 1, 2

[26] Kamalakar Vijay Thakare, Debi Prosad Dogra, Heeseung Choi, Haksub Kim, and Ig-Jae Kim. Rareanom: A benchmark video dataset for rare type anomalies. *Pattern Recognition*, 2023. 2, 6, 7

[27] Kamalakar Vijay Thakare, Yash Raghuwanshi, Debi Prosad Dogra, Heeseung Choi, and Ig-Jae Kim. Dyannet: A scene dynamicity guided self-trained video anomaly detection network. In *WACV*, 2023. 2, 3, 6, 7

[28] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *ICCV*, 2021. 1, 2, 6

[29] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv*, 2023. 3, 6

[30] Anil Osman Tur, Nicola Dall'Asen, Cigdem Beyan, and Elisa Ricci. Exploring diffusion models for unsupervised video anomaly detection. In *ICIP*, 2023. 1, 2, 6

[31] Anil Osman Tur, Nicola Dall'Asen, Cigdem Beyan, and Elisa Ricci. Unsupervised video anomaly detection with diffusion models conditioned on compact motion representations. In *ICIAP*, 2023. 1, 2, 6

[32] Gaoang Wang, Xinyu Yuan, Aotian Zheng, Hung-Min Hsu, and Jenq-Neng Hwang. Anomaly candidate identification and starting time estimation of vehicles from traffic videos. In *CVPRW*, 2019. 1, 2

[33] Jue Wang and Anoop Cherian. Gods: Generalized one-class discriminative subspaces for anomaly detection. In *ICCV*, 2019. 6

[34] Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. Self-supervised sparse representation for video anomaly detection. In *ECCV*, 2022. 6

[35] Peng Wu and Jing Liu. Learning causal temporal relation and feature discrimination for anomaly detection. *IEEE TIP*, 2021. 1, 2, 6

[36] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *ECCV*, 2020. 2, 5, 6, 1, 3

[37] Cheng Yan, Shiyu Zhang, Yang Liu, Guansong Pang, and Wenjun Wang. Feature prediction diffusion model for video anomaly detection. In *ICCV*, 2023. 1, 2

[38] Muhammad Zaigham Zaheer, Jin-ha Lee, Marcella Astrid, and Seung-Ik Lee. Old is gold: Redefining the adversarially learned one-class classifier training paradigm. In *CVPR*, 2020. 1, 2

[39] Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee. Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In *ECCV*, 2020. 6

[40] M Zaigham Zaheer, Arif Mahmood, M Haris Khan, Mattia Segu, Fisher Yu, and Seung-Ik Lee. Generative cooperative learning for unsupervised video anomaly detection. In *CVPR*, 2022. 1, 2, 6, 7

[41] Jiangong Zhang, Laiyun Qing, and Jun Miao. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In *ICIP*, 2019. 6

[42] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *CVPR*, 2019. 6

Harnessing Large Language Models for Training-free Video Anomaly Detection

Supplementary Material

In this supplementary material, we first provide the exact form of the prompts employed in our method and then we present additional experimental analyses. Specifically, we first present the impact of the task-related priors in prompting the anomaly scores on XD-Violence [36]. We then present the impact of captioning models, *i.e*. different variants of BLIP-2 models, for the VAD performance of our method on both XD-Violence [36] and UCF-Crime [24] datasets. Finally, we ablate the hyperparameters in constructing temporal windows to justify our design choice. Moreover, we describe the limitations and broader social impacts of our work, and we showcase additional qualitative results that demonstrate temporal summaries and the detection results. More qualitative results in the form of videos can be conveniently accessed on the project website at https://lucazanella.github.io/lavad/.

A. Prompts

The prompts utilized in our approach serve distinct purposes. The contextual prompt $\texttt{P}_C$ provides priors to the LLM for VAD. In line with the findings of our ablation studies presented in Tab. 4 and in Tab. 5, this prompt differs for UCF-Crime [24] and XD-Violence [36]. For UCF-Crime, the prompt is structured as: "*If you were a law enforcement agency, how would you rate the scene described on a scale from 0 to 1, with 0 representing a standard scene and 1 denoting a scene with suspicious activities?*". In contrast, for XD-Violence, the prompt has the form: "*How would you rate the scene described on a scale from 0 to 1, with 0 representing a standard scene and 1 denoting a scene with suspicious or potentially criminal activities?*".

The prompt $\texttt{P}_F$ provides guidance to the LLM for the desired output format, aimed at facilitating automated text parsing. This prompt remains consistent across both datasets and is defined as follows: "*Please provide the response in the form of a Python list and respond with only one number in the provided list below [0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0] without any textual explanation. It should begin with '[' and end with ']'.*".

Lastly, the prompt $\texttt{P}_S$ is employed to obtain a temporal summary S_i for each frame $\mathbf{I}_i$. The prompt is formulated as follows: "*Please summarize what happened in few sentences, based on the following temporal description of a scene. Do not include any unnecessary details or descriptions.*".

B. Additional analyses

Task priors in the context prompt. In Tab. 5 we present the impact of different priors in the context prompt $\texttt{P}_C$, i.e. impersonation and anomaly priors, on XD-Violence [36]. This follows the same ablation design as presented in Tab. 4 in the main manuscript for UCF-Crime, with the priors added in the same way for both datasets. As shown in Tab. 5, for videos within XD-Violence, incorporating the anomaly prior (Row 2) improves the average precision (AP) by $+1.67\%$ compared to using only the base context prompt (Row 1). Conversely, introducing impersonation (Row 3) degrades the AP by -1.51% compared to not using it (Row 1). Videos in XD-Violence originate from various sources, including CCTV cameras, movies, sports, and games. The effectiveness of the impersonation prior might be limited to CCTV camera videos, given that the surveillance domain is more closely associated with the concept of *"law enforcement agency"* which is utilized for impersonation. Finally, combining both priors (Row 4) leads to improved performance compared to not utilizing any of them, primarily due to the positive impact of the anomaly prior.

Table 5. Results of LAVAD on XD-Violence with different priors in the context prompt when querying the LLM for anomaly scores.

ANOMALY PRIOR	IMPERSONATION	AP (%)	AUC (%)
✗	✗	60.34	84.42
✓	✗	**62.01**	**85.36**
✗	✓	58.83	84.50
✓	✓	60.78	85.26

Impact of different BLIP-2 models. As captioners, we consider different BLIP-2 [14] models and their ensemble for both UCF-Crime [24] and XD-Violence [36], and we present the results in Tables 6 and 7, respectively.

In Tab. 6, the most effective strategy for UCF-Crime videos is employing an ensemble of all BLIP-2 models (Row 6). This involves generating captions for all frames in a video using all BLIP-2 models and relying on the vision-language model (VLM) to identify the semantically closest captions for each frame. The effectiveness of the ensemble might be attributed to the challenges posed by UCF-Crime videos. These videos, characterized by low-resolution CCTV footage, often lead captioning models to hallucinate scene descriptions. For instance, it is common to encounter captions, such as *"a person riding a skateboard down a road"* when the image only depicts a road in the absence of any specific event. The ensemble approach, by allowing the selection from a larger set of candidates, increases the likelihood of choosing more correct captions and filtering incorrect ones.

For XD-Violence, as shown in Tab. 7, utilizing the captions generated by *flan-t5-xxl* (Row 3) yields the best average

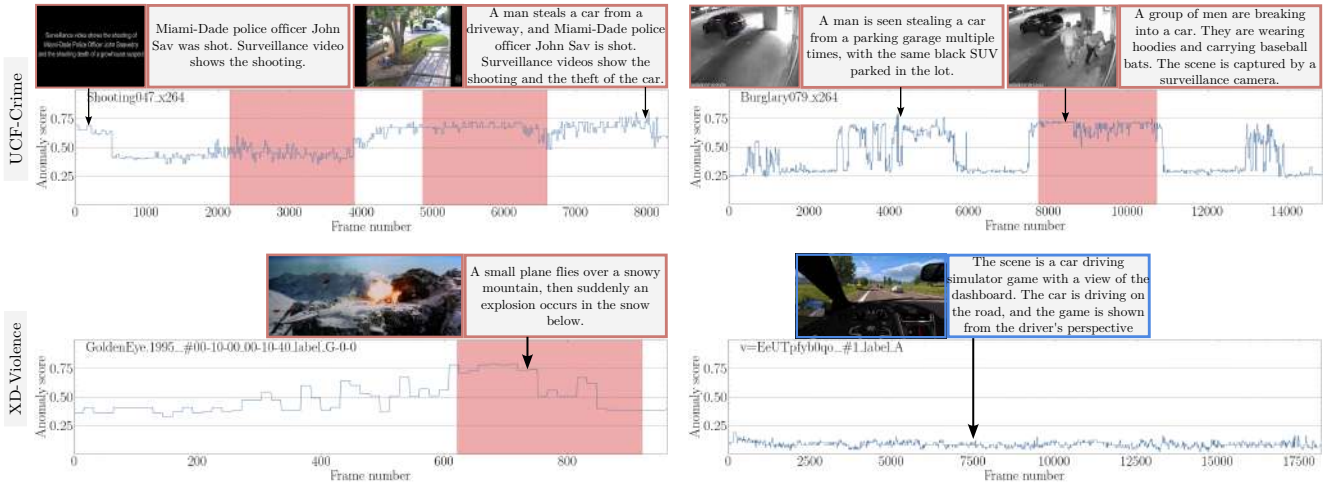

Figure 7. We showcase qualitative results obtained by LAVAD on four test videos, including two videos (top row) from UCF-Crime and two videos from XD-Violence (bottom row). For each video, we plot the anomaly score over frames computed by our method. We display some keyframes alongside their most aligned temporal summary (blue bounding boxes for normal frame predictions and red bounding boxes for abnormal frame predictions), illustrating the relevance among the predicted anomaly score, visual content, and description. Ground-truth anomalies are highlighted.

Table 6. Results of LAVAD on UCF-Crime with different BLIP-2 model variants in our Image-Text Caption Cleaning technique.

BLIP-2					AUC
FLAN-T5-XL	FLAN-T5-XL-COCO	FLAN-T5-XXL	OPT-6.7B	OPT-6.7B-COCO	(%)
✓	✗	✗	✗	✗	74.19
✗	✓	✗	✗	✗	74.49
✗	✗	✓	✗	✗	74.38
✗	✗	✗	✓	✗	75.50
✗	✗	✗	✗	✓	73.94
✓	✓	✓	✓	✓	**80.28**

Table 7. Results of LAVAD on XD-Violence with different BLIP-2 model variants in our Image-Text Caption Cleaning technique.

BLIP-2					AP	AUC
FLAN-T5-XL	FLAN-T5-XL-COCO	FLAN-T5-XXL	OPT-6.7B	OPT-6.7B-COCO	(%)	(%)
✓	✗	✗	✗	✗	61.09	85.16
✗	✓	✗	✗	✗	57.41	82.78
✗	✗	✓	✗	✗	**62.01**	**85.36**
✗	✗	✗	✓	✗	56.55	82.42
✗	✗	✗	✗	✓	54.71	82.93
✓	✓	✓	✓	✓	59.62	84.90

precision (AP). Other BLIP-2 variants for XD-Violence may provide captions that prioritize foreground objects, potentially overlooking background elements constituting anomalies (*e.g.* a vehicle enveloped in smoke on a busy street), yet better aligning with the VLM's representation of the video frames. Hence, when employing the ensemble of BLIP-2 models (Row 6), captions that specifically highlight elements constituting anomalies are not chosen as the semantically closest captions to video frames in the cleaning step, with a negative impact on the anomaly scoring phase.

Temporal window's duration and number of sampled frames. In Tab. 8, we evaluate the impact of varying the duration of the temporal window (T) and the number of sampled frames (N), which is used to query the LLM for the temporal summary S_i. Specifically, the temporal window duration T determines the time interval, while the number of sampled frames N determines the number of captions. First, we conduct experiments by adjusting the duration T to 2.5, 5, 10, and 20 seconds, while maintaining $N = 10$. The 10-second temporal window yields the highest AUC score (Row 3). This is in line with the fact that ImageBind [6] is trained with video clips of 10 seconds.

Table 8. Results of LAVAD on UCF-Crime with different combinations of temporal window duration (T) and number of sampled frames per window (N).

T(s)	N	AUC (%)
2.5	10	79.33
5	10	78.10
10	10	**80.28**
20	10	79.24
10	5	77.48
10	20	74.45

Subsequently, we maintain the temporal window's duration T at 10 seconds and vary the number of frames from 5 to 10 and 20. Notably, using 10 frames (Row 3), *i.e.* 1 frame every second, is the optimal choice within this experiment. Balancing the number of captions per snippet presents a trade-off with the quality of the summary. Too many captions may overwhelm with excessive and non-diverse content, while too few captions may result in limited coverage of the content.

C. Qualitative results

In Fig. 7, we present additional qualitative results demonstrating the effectiveness of our proposed LAVAD in detecting anomalies within a set of UCF-Crime [24] and XD-Violence [36] test videos. The figure showcases keyframes along with the most semantically similar temporal summaries. For example, in the video *Shooting047* (Row 1, Column 1), LAVAD assigns a high anomaly score when the video is labeled abnormal. However, it also assigns a high anomaly score during the initial and final segments, despite these parts being labeled as normal. This discrepancy arises because the video begins with text describing the subsequent content, leading the LLM to attribute a high anomaly score. In the final part, our method correctly identifies abnormality as the frame depicts a person on the ground who has been shot. In the video *Burglary079* (Row 1, Column 2), there is a false abnormal instance. This occurs because the temporal summary associated with that frame incorrectly suggests the presence of a man stealing a car. In reality, the video depicts a man behaving suspiciously near the car, leading to a wrong interpretation by the captioning module. In the XD-Violence videos (Row 2), an anomaly caused by an explosion is correctly detected (Row 2, Column 1), while a normal video is consistently predicted as normal for more than $17,500$ frames (Row 2, Column 2).

D. Limitations

We identify two main limitations of our work. Firstly, our method fully relies on pre-trained models from VLMs and LLMs, thus its performance greatly depends on i) how well the captioning model describes the visual content, ii) how reliable the LLM is when generating the anomaly scores, and iii) how aligned the multi-modal encoders are when processing videos from various domains. Secondly, our anomaly scores are primarily obtained via prompting LLMs. Although we conducted experiments investigating different prompting strategies, a systematic understanding of LLM prompting for VAD requires a community effort.

E. Broader Societal Impacts

While our work pioneers the technical aspect of leveraging LLMs for detecting anomalies in videos, there exist open ethical challenges for a broader concern. VAD systems are mostly applied to safety-related contexts, for private use or public interests. Prior to any deployment, it is crucial to first investigate the behaviors of LLM-based methods, mitigating any potential bias in LLMs and improving explainability. Our work serves as the first technical exploration of leveraging LLMs for training-free VAD, proving it as a competitive alternative. This is a necessary step to increase the awareness of the community on these important topics.