This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Denoise, Divide, Distill, and Predict ($\mathcal{D}^3\mathcal{P}$): Towards Forecasting Long-horizon Real-world Anomaly from Normalcy

Quentin Mérilleau[1,2], Snehashis Majhi[1,2], Antitza Dantcheva[1,2], Quan Kong[3], Lorenzo Garattoni[4], Gianpiero Francesca[4], François Brémond[1,2]

[1] INRIA [2] Côte d'Azur University [3] Woven by Toyota [4] Toyota Motor Europe

Code: https://github.com/quentinmerilleau/D3P

Abstract

Forecasting abnormal human behavior (AHB) in unconstrained real-world environments is critical for enabling proactive safety interventions. Unlike short-term anomaly detection, long-horizon forecasting offers a vital reaction window but remains underexplored due to three core challenges: (i) noisy, complex human–agent interactions; (ii) weak temporal coupling between normal observations and distant anomalies; and (iii) data scarcity limiting the scalability of autoregressive models. To address these, we propose $\mathcal{D}^3\mathcal{P}$ (Denoise, Divide, Distill, and Predict), a novel encoder–decoder framework that bridges denoised pasts with distilled autoregressive futures. Our Differential Past Encoder (DiPE) disentangles scene-level and object-level dynamics via differential attention, suppressing irrelevant interactions and enhancing discriminative cues. The Distilled Future Auto-Regressive Decoder (D-FAD) adopts a divide-and-conquer strategy, segmenting future queries into temporal chunks for sequential prediction, while leveraging distillation to balance robustness and latency. We validate our approach on the AHB-F benchmark, the only dataset dedicated to abnormal behavior forecasting, and further integrate D-FAD with several state-of-the-art methods. In all cases, our framework consistently outperforms prior work in both forecasting accuracy and computational efficiency.

1. Introduction

Abnormal human behavior (AHB) in real-world environments, such as theft, aggression, or panic poses severe risks to both individuals and infrastructure. These events are rare, subtle, complex, and deeply entangled with normal interactions, making them notoriously difficult to forecast. While video-based anomaly detection has traditionally been active, recent efforts have shifted towards proactive forecasting, seeking to recognize the early signs of anomalies before they unfold. In real-world scenarios, long-horizon forecasting is far more valuable as it provides the critical reaction window needed for preventive intervention.

Despite its importance, robust long-term anomaly forecasting from normal-only observations remains underexplored, hindered by three fundamental challenges: **(I) Complexity in human–agent interactions and surrounding noise :** Real-world scenes are crowded, unstructured, and filled with noisy dynamics. Human behaviors are unconstrained and context-dependent, with subtle object-level cues buried under noisy scene dynamics. For instance, predicting a shoplifting event emerging from normal shopping activity requires disentangling discriminative cues from overwhelming background noise. **(II) Weak coupling between present and distant future:** When only normal frames are observed, the link to a distant abnormal outcome is negligible. This exacerbates uncertainty and makes models prone to error propagation in long-horizon prediction. **(III) Data scarcity vs. autoregressive capacity :** While large language models (LLMs) like GPT [3] and LLaMA [49] excel at long-range autoregression, their success hinges on massive training datasets. In anomaly forecasting, domain-specific video data is scarce, and naively applying large autoregressive decoders incurs severe data–parameter imbalance and huge inference latency. These challenges lead to the central question: **How can we design a robust, low-latency framework for long-horizon anomaly forecasting in unconstrained, real-world settings?**

To address this, we propose $\mathcal{D}^3\mathcal{P}$ (Denoise, Divide, Distill, and Predict), a novel encoder–decoder framework that bridges denoised pasts with distilled autoregressive futures (see Fig. 1). It is powered by two complementary components:**(A) Differential Past Encoder (DiPE) :** Designed to suppress irrelevant interactions and emphasize discriminative cues. DiPE disentangles temporal–spatial dynamics by separately modeling scene-level and object-level interactions. Through differential attention, it sharpens the rep-

resentation of normal behavior by eliminating the irrelevant features, thereby reducing uncertainty about their abnormal continuations. **(B) Distilled Future Auto-Regressive Decoder (D-FAD) :** To combat the weak coupling between present and distant future, D-FAD adopts a divide-and-conquer autoregressive strategy. Future queries are split into smaller temporal chunks; each is predicted locally and passed forward sequentially, enabling predictions at earlier horizons to bootstrap accuracy at later horizons. To balance performance and efficiency, D-FAD incorporates distillation against a classical transformer decoder, achieving robustness in long-term forecasting while mitigating parameter overhead and latency. Additionally, D-FAD can be adopted to any state-of-the-art methods in a *plug-and-play* manner to further enhance the long-horizon prediction efficacy. To the best of our knowledge, $\mathcal{D}^3\mathcal{P}$ is the first framework to systematically bridge denoised pasts with distilled autoregressive futures for real-world long-horizon anomaly forecasting, offering both predictive accuracy and practical efficiency.

To summarize, our contributions are in three-folds.

- We propose $\mathcal{D}^3\mathcal{P}$, the first encoder–decoder framework that bridges denoised pasts (DiPE) with distilled autoregressive futures (D-FAD) for robust long-horizon anomaly forecasting from normal-only observations.
- We design D-FAD, a divide-and-conquer autoregressive decoder that predicts future queries sequentially, mitigates weak past–future coupling, and leverages knowledge distillation for efficiency. D-FAD is *plug-and-play* and consistently improves SoTA models.
- We conduct extensive experiments on the AHB-F benchmark, the only available dataset for abnormal human behavior forecasting, and further validate D-FAD across multiple SoTA architectures. Our method consistently outperforms prior approaches in both forecasting accuracy and efficiency.

2. Related Work

Action Anticipation. Action anticipation has been studied in both third-person [1, 6, 16, 41, 50] and egocentric settings [11, 12, 14, 18, 39]. Early approaches relied on recurrent models [12, 13, 36, 45], where a rolling LSTM encodes observations and an unrolled LSTM predicts future actions. While effective on short clips, RNNs struggle to preserve long-range temporal dependencies and multi-actor interactions typical of real footage. Transformers have therefore become the de-facto backbone [17, 18, 20, 33, 34, 39, 43, 44, 52, 54, 61], leveraging global attention, intention conditioning, and hierarchical aggregation. However, these architectures are commonly validated on relatively simple activity sets and often assume a single dominant event stream. In cluttered scenes with concurrent actions, limited supervision, and noisy dynamics, their global attention becomes diffuse and optimization favors salient but not necessarily predictive cues, leading to brittle long-horizon anticipation.

Generative Anticipation. Generative formulations synthesize multiple plausible futures using U-Net/GAN/diffusion variants [4, 15, 57, 58, 63]. By explicitly modeling uncertainty, they can capture diverse outcomes; yet, the same past can map to many equally likely futures, yielding non-deterministic predictions and evaluation ambiguity. Moreover, pixel- or token-level generation can be computationally heavy and may drift from the task objective (action or anomaly labels), especially when training data are sparse.

Auto-Regressive and LLM-Driven Forecasting. Inspired by large language models [3, 10, 49], recent works forecast with textual reasoning only (e.g., antGPT [60], PaLM-based [21]) or multimodally while retaining visual tokens (ICVL [5], PlausiVL [35]). These approaches benefit from world knowledge and semantic generalization, but remain highly prompt- and context-dependent. In real-world anticipation where future events are open-ended, concurrent, and weakly specified LLM outputs can be inconsistent across prompts and sensitive to dataset biases, limiting reliable deployment.

Video Anomaly Understanding. VAD has progressed along online/offline regimes with weakly supervised MIL formulations [2, 9, 24, 26–32, 37, 46, 51, 53, 56, 59, 62], often evaluated on UCF-Crime [46], XD-Violence [53], ShanghaiTech [25], and IITB-Corridor [42]. Classical pipelines emphasize motion magnitude (e.g., TCN [23] with optical flow [47]) and thus favor sharp scene anomalies, while missing subtle human-centric cues. Context-relation models (e.g., MTN [48]) and more recent transformer designs [7] introduce global–local reasoning; nevertheless, their magnitude-driven objectives continue to overweight conspicuous scene changes, under-representing fine-grained human interactions that often precede anomalies. Crucially, prior VAD work focuses on detection, not anticipation leaving the question of *when* future anomalies will occur largely unexplored.

Across these threads, three limitations persist: (i) anticipation models excel on simple, single-stream activities but degrade with concurrent events ; (ii) generative and LLM-driven approaches produce non-deterministic, prompt-sensitive forecasts misaligned with label-level objectives; and (iii) VAD pipelines emphasize detection of salient scene shocks rather than forecasting subtle, human-driven precursors. These gaps motivate our framework, which targets *real-world anomaly anticipation* by (a) extracting discriminative, interaction-aware representations from normal pasts, (b) performing label-level, auto-

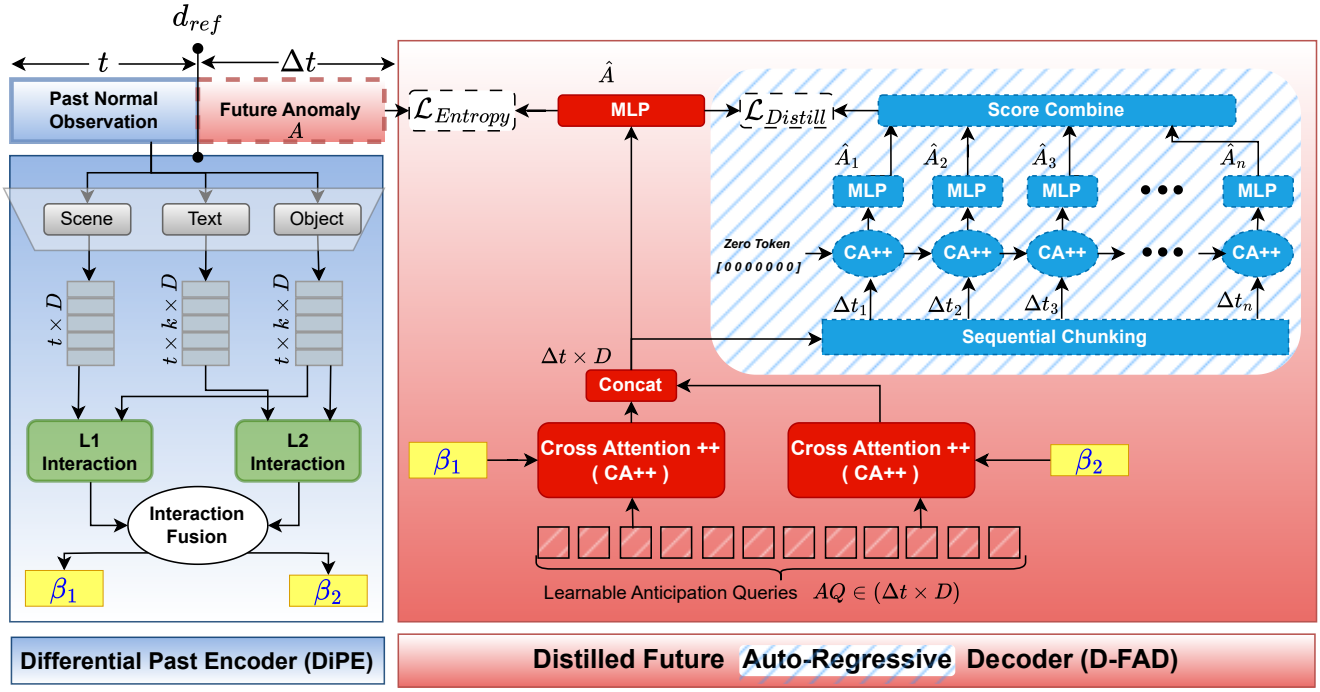


Figure 1. **Overview** of the proposed **D^{3}P** framework for anomaly forecasting. The Differential Past Encoder (DiPE) captures discriminative cues from past normal observations by modeling multi-level interactions (scene, text, object) and fusing them into robust representations. The Distilled Future Auto-Regressive Decoder (D-FAD) then generates sequential future predictions via a divide-and-conquer strategy, where learnable anticipation queries are refined through stacked cross-attention++ modules and distilled into compact predictions

regressive forecasting that remains deterministic and data-efficient, and (c) explicitly accounting for long-horizons while preserving subtle human-centric cues.

3. Preliminaries

Long-horizon anomaly forecasting from normal-only observations poses a unique challenge in video understanding, where the goal is to anticipate future abnormal events solely on temporally distant, non-anomalous precursors. Formally, given a sequence of normal observations $\mathcal{X} = \{x_1, x_2, \ldots, x_t\}$, the task is to predict a future sequence $\hat{\mathcal{Y}} = \{y_{t+1}, \ldots, y_{t+\Delta t}\}$ using a forecasting function $\mathcal{F}$ such that $\hat{\mathcal{Y}} = \mathcal{F}(\mathcal{X}; \theta)$, where θ denotes the model parameters. As the anticipation horizon Δt increases, the semantic coupling between $\mathcal{X}$ and $\mathcal{Y}$ weakens, leading to elevated predictive uncertainty. This uncertainty quantified by the entropy $\mathcal{H}(\hat{\mathcal{Y}}) = -\sum_{i=t+1}^{t+\Delta t} \sum_{c \in \mathcal{C}} p(y_i = c) \log p(y_i = c)$ over behavior classes $\mathcal{C}$—is exacerbated by noise amplification and semantic drift in unconstrained environments. Minimizing this long-term temporal uncertainty is crucial for reliable forecasting, as it enables the model to suppress irrelevant dynamics and focus on subtle, discriminative cues that precede anomalies. The D^{3}P framework is explicitly designed to address this challenge by denoising past representations and distilling future predictions through autoregressive refinement, thereby enhancing both accuracy and temporal stability in anomaly anticipation.

4. Proposed D^{3}P Framework

We present **D^{3}P (Denoise, Divide, Distill, and Predict)**, a novel encoder-decoder architecture tailored for long-horizon anomaly forecasting from normal-only observations. The key principle behind $\mathcal{D}^3\mathcal{P}$ is twofold: (i) *denoise past representations* to filter irrelevant human-agent interactions, and (ii) *distill autoregressive futures* to stabilize long-range predictions. This design explicitly addresses the three fundamental challenges outlined earlier: (1) noisy and redundant dynamics, (2) weak temporal coupling between present and distant abnormal futures, and (3) limited data availability versus the heavy capacity of autoregressive models. As illustrated in Fig. 1, the framework consists of two complementary modules: the **Differential Past Encoder (DiPE)** and the **Distilled Future Auto-Regressive Decoder (D-FAD)**.

4.1. Feature Extraction

Given a temporal observation window of t frames, we extract multi-modal features that capture complementary se-

mantics at the scene, object, and language levels. We employ Mask2Former [8] for panoptic segmentation and CLIP [40] for shared vision-language encoding:

- **Object Features:** Object instances are cropped and passed through CLIP's image encoder to obtain embeddings $F_O \in \mathbb{R}^{t\times k\times D}$, where k is the number of detected objects per frame.
- **Text Features:** Object category labels are encoded with CLIP's text encoder, yielding $F_{\text{txt}} \in \mathbb{R}^{t\times k\times D}$.
- **Scene Features:** Holistic frame-level embeddings $F_S \in \mathbb{R}^{t\times D}$ are extracted using CLIP's image encoder.

This tri-modal feature space provides the foundation for DiPE, enabling fine-grained modeling of spatio-temporal context and early discriminative cues within normal observations.

4.2. Differential Past Encoder (DiPE)

The goal of DiPE is to construct a *denoised and discriminative representation of the past*. Unlike conventional temporal encoders that aggregate features indiscriminately, DiPE explicitly disentangles:

- **Global Interaction Module:** Encodes cross-temporal dependencies between coarse scene features F_S and pooled object features F_O. This emphasizes scene evolution while filtering out redundant motion.
- **Local Interaction Module:** Captures cross-spatial interactions among objects and their textual labels. Temporal pooling highlights persistent cues, ensuring subtle signals (e.g., abnormal object handling) are retained while transient noise is suppressed.

Both modules are instantiated with **Mamba blocks** [19], a state-space sequence model that scales linearly with input length while preserving long-range dependencies, a key requirement for extended horizons. For input $x_t \in \mathbb{R}^D$, the latent state evolves as:

$$h_t = \bar{\mathbf{A}} h_{t-1} + \bar{\mathbf{B}} x_t, \tag{1}$$

$$y_t = \mathbf{C} h_t, \tag{2}$$

where $\bar{\mathbf{A}} = \exp(\Delta\mathbf{A})$ encodes discretized dynamics.

To further suppress redundancy and noise, we take inspiration from the recent success of the differential transformer [55] in LLMs and propose a **Cross-Attention++ (CA++)** that uses a differential attention operator:

$$\text{CA++} = \left(\text{softmax}(Q_1 K_1^\top) - \lambda\,\text{softmax}(Q_2 K_2^\top)\right) V, \tag{3}$$

where λ is a learnable weight. Unlike vanilla cross-attention, CA++ explicitly subtracts redundant alignments, sharpening cross-modal correlations.

Outputs from the two modules are integrated via an Interaction fusion that takes inspiration from NULL module of [33] block, which adaptively weights scene- and object-level cues based on predictive confidence. This produces embeddings $\beta_1, \beta_2 \in \mathbb{R}^d$, which serve as denoised past representations and input to D-FAD.

4.3. Distilled Future Auto-Regressive Decoder (D-FAD)

While DiPE suppresses noise in the past encoder states, the **Distilled Future Auto-Regressive Decoder (D-FAD)** is designed to *stabilize long-horizon forecasting*. Unlike conventional transformer decoders that predict all future queries jointly, often compounding errors when the correlation between past and future is weak, D-FAD adopts a **divide-and-conquer autoregressive strategy** that refines conventional decoding with sequential distillation.

Anticipation Queries with Conventional Decoder. A set of learnable anticipation queries is defined as $AQ \in \mathbb{R}^{\Delta t\times D}$, where Δt corresponds to the prediction horizon and d is the embedding dimension. These queries interact with encoded features β_1, β_2 through a stacked cross-attention mechanism (CA++). The attended features are then concatenated and projected via an MLP classifier to produce initial logits $\hat{A}$:

$$\hat{A} = MLP(\text{concat}(CA{+}{+}(\beta_1, AQ),\, CA{+}{+}(\beta_2, AQ))). \tag{4}$$

Sequential Chunking with Autoregression. To mitigate long-term error accumulation, D-FAD divides the full horizon Δt into N sequential chunks $\{P_i\}_{i=1}^N$, each responsible for a short-term forecast. For training, chunk predictions are conditioned autoregressively:

$$P_{i+1} = CA{+}{+}(P_i, \beta_1, \beta_2), \quad i = 0, \ldots, N-1, \tag{5}$$

where $P_0 = AQ$. This recursive conditioning converts long-horizon forecasting into a chain of near-term subproblems, reducing semantic drift and leveraging short-term certainty to bootstrap long-range accuracy. However, such sequential decoding increases inference latency due to cumulative autoregression.

Prediction Distillation. To preserve long-range reasoning while maintaining low latency, we distill autoregressive predictions against the conventional decoder. Specifically, given autoregressive logits $\hat{A}_{AR}$ and conventional decoder logits $\hat{A}$, we define the distillation loss as

$$\mathcal{L}_{distill} = \|\hat{A}_{AR} - \hat{A}\|_2^2, \tag{6}$$

where $\|\cdot\|_2^2$ denotes the `mean-squared-error`. This distillation enforces that autoregressive forecasts inherit the stability and long-term consistency of full decoding, while reducing parameters and latency at inference time.

Training Objective. The final loss for D-FAD combines the action anticipation loss with the distillation regularizer:

$$\mathcal{L}_{D-FAD} = \lambda_1 \mathcal{L}^A(\hat{A}, A) + \lambda_2 \mathcal{L}_{distill}, \quad (7)$$

where λ_1, λ_2 are balancing hyperparameters. The $\mathcal{L}^A$ loss is defined with standard `cross-entropy` between action A and logits $\hat{A}$. This ensures that D-FAD achieves a trade-off between short-term stability, long-term consistency, and computational efficiency.

Together, DiPE and D-FAD form $\mathcal{D}^3\mathcal{P}$ a robust, low-latency framework that systematically denoises complex pasts and distills autoregressive futures, enabling accurate long-horizon anomaly forecasting in real-world environments.

5. Experiments

5.1. Dataset

We experiment our method on the Abnormal Human Behavior Forecast (AHB-F) dataset [33], the only AHB forecasting dataset available. Collecting abnormal behaviors perpetuated by humans in real-world scenarios and scenes. AHB-F contains 1470 videos separated with a 3/1 ratio in 1091 videos (31.03h) for training and 379 (9.90h) for testing. The videos average 102.04 seconds for the training part and 94.13 seconds for the testing part. AHB-F contains 26 types of human anomalies recorded in Close-Circuit Television scenarios.

5.2. Evaluation Metrics

To investigate the performance of our method, we use an evaluation protocol based on deterministic reference points ($\mathbf{D_{ref}}$) splitting in two parts a video, the observation frames and the future prediction frames. Aiming, for this task, to observe only normal frames to predict the future occurrence of an anomaly, the reference point ($\mathbf{D_{ref}}$) is placed such that the observation frames only contains normal frames and the future prediction frames contains abnormal frames, normal frames or both. In order to compare our results with previous we keep frame-level *Top-1* mean average precision (mAP) for evaluation on the test set. We advocate that our work is over-performing previous approaches on both short- and long-term anticipation. In this optic we evaluate our method on 3 "short indicators" the mean mAP for 1,2 and 3 seconds in the future and 2 "long indicators" the mean mAP for 4 and 8 seconds. With this protocol, we are confident to prove the robustness of $\mathcal{D}^3\mathcal{P}$ on all the anomaly instances regardless of their length.

5.3. Implementation Details

We implement our model using Pytorch [38]. All our experiences are performed on a system of 4 Nvidia V100 GPU. We evaluated our methods with 1,2,3 and 4 layers of Transformer Decoder when performing Cross-Attention++ (CA++) for D-FAD. For all our Transformer units, we set the number of heads to 8 and hidden units to 128. For the mamba units, we set the number of SSM state expansion factor to 256, the convolution width to 16 and the expansion factor to 4. Our method is trained for 100 epochs using the Adam [22] optimizer with a learning rate of 0.0001.

6. Ablation Study

To systematically evaluate the design choices within our D^{3}P framework, we conduct a series of ablation studies focusing on both architectural components and training strategies. These experiments are designed to isolate the contributions of key modules namely the **Differential Past Encoder (DiPE)** and the **Distilled Future Auto-Regressive Decoder (D-FAD)** and to quantify their impact on forecasting accuracy across short and long temporal horizons. By varying encoder backbones, attention mechanisms, autoregressive chunking strategies, and distillation objectives, we aim to uncover the optimal configuration for robust anomaly anticipation. All results are reported on the AHB-F benchmark using Top-1 mAP, and demonstrate that each component of D^{3}P plays a critical role in enhancing temporal stability, discriminative power, and computational efficiency.

6.1. Impact of Mamba in Past Interaction Encoding

To assess the effectiveness of Mamba blocks within the Differential Past Encoder (DiPE), we compare three configurations: a baseline transformer-layer encoder (T.), a distillation setup from transformer to Mamba (Distill(T.,M.)), and a pure Mamba-layer encoder (M.). As shown in Table 1, the Mamba-layer variant achieves superior performance across both short- and long-horizon forecasting. Specifically, it yields a short-term average Top-1 mAP of 66.22% and a long-term average of 63.19%, outperforming the transformer baseline (64.81% short, 61.74% long) and the distillation variant (65.24% short, 62.39% long). These improvements highlight Mamba's ability to model long-range temporal dependencies with linear complexity, which is critical for capturing subtle behavioral cues in extended video sequences. Furthermore, the integration of Cross-Attention++ (CA++) with Mamba enhances cross-modal alignment by subtracting redundant attention weights, thereby sharpening discriminative features and suppressing irrelevant dynamics. This synergy between state-space modeling and differential attention enables DiPE to produce robust, noise-resistant representations essential for anomaly anticipation.

6.2. Impact of Auto-Regression for Long-Horizon Temporal Modeling

We further investigate in Table 1 the role of autoregressive chunking in the Distilled Future Auto-Regressive Decoder (D-FAD) by comparing three decoder variants: a con-

DiPE			D-FAD			Short (Top-1 mAP %)				Long (Top-1 mAP %)		
T.	**Distill(T., M.)**	**M.**	**TD.**	**AR.**	**Distill(TD., AR.)**	**1 sec.**	**2 sec.**	**3 sec.**	**Avg**	**4 sec.**	**8 sec.**	**Avg**
✓	-	-	✓	-	-	65.96	64.64	63.85	64.81	63.85	59.63	61.74
-	✓	-	✓	-	-	66.22	64.90	64.64	65.24	63.85	60.94	62.39
-	-	✓	✓	-	-	65.96	66.49	66.22	66.22	65.96	60.42	63.19
-	-	✓	-	✓	-	22.42	38.25	49.07	36.58	40.36	35.35	37.85
-	-	✓	-	-	✓	66.75	67.81	67.28	67.28	66.49	61.21	63.85

Table 1. Ablation study evaluating the impact of architectural choices in DiPE and D-FAD. We compare transformer-based encoding **(T.)**, Mamba-based encoding **(M.)**, and distillation strategies **(Distill(T.,M.))** for DiPE, alongside conventional Transformer decoding **(TD.)**, autoregressive chunking **(AR.)**, and distillation **(Distill(TD.,AR.))** for D-FAD.

D-FAD Chunks	$\mathcal{L}_{distill}$			Short (Top-1 mAP %)				Long (Top-1 mAP %)		
N =	**MSE**	**KL-Div.**	**MAE.**	**1 sec.**	**2 sec.**	**3 sec.**	**Avg**	**4 sec.**	**8 sec.**	**Avg**
2	✓	-	-	66.49	65.43	64.90	65.60	63.85	60.15	62.00
4	✓	-	-	65.96	67.81	64.37	66.04	64.37	59.89	62.13
8	✓	-	-	66.75	64.64	65.17	65.52	63.85	61.21	62.53
16	✓	-	-	65.96	64.90	64.90	65.25	64.64	60.94	62.79
32	✓	-	-	66.22	64.90	64.37	65.16	64.11	60.68	62.39
8	-	✓	-	64.90	64.37	64.64	64.63	64.64	59.36	62.00
8	-	-	✓	66.22	65.17	64.90	65.43	64.11	60.15	62.13

Table 2. Ablation study on the design of D-FAD, analyzing the impact of autoregressive chunk granularity (N) and distillation loss formulation ($\mathcal{L}_{distill}$).

ventional transformer decoder (TD.), a pure autoregressive strategy (AR.), and a distillation setup from autoregressive to full transformer decoding refinement (Distill(TD., AR.)). Notably, the autoregressive-only configuration performs poorly (36.58% short, 37.85% long), underscoring the difficulty of learning stable long-horizon forecasts without global guidance. In contrast, the distilled transformer decoder achieves the highest performance (67.28% short, 63.85% long), surpassing both the conventional decoder and the non-distilled autoregressive variant. This demonstrates that neither sequential chunking nor global temporal consistency alone is sufficient; distillation from autoregressive modeling to full decoding is essential to share global temporal consistency with local short-term certainty while preserving efficiency. By recursively conditioning predictions across temporal chunks, D-FAD reduces semantic drift and leverages short-term certainty to bootstrap long-range accuracy. These results validate that combining autoregression with knowledge distillation is key to achieving robust, low-latency forecasting in real-world anomaly anticipation scenarios.

6.3. Optimal Split for Auto-Regression

To determine the optimal temporal granularity for autoregressive forecasting, we vary the number of sequential chunks N in D-FAD and evaluate performance across short- and long-horizon settings. As shown in Table 2, chunking the prediction horizon into 8 segments yields the best overall performance, achieving a short-term average Top-1 mAP of 65.52% and a long-term average of 62.53%. While smaller splits ($N = 2$ or 4) offer competitive short-term accuracy, they suffer from reduced long-range stability due to insufficient temporal decomposition. Conversely, larger splits ($N = 16$ or 32) improve long-horizon consistency (e.g., 62.79% for $N = 16$) but slightly degrade short-term precision, likely due to over-fragmentation and reduced context per chunk. These results suggest that $N = 8$ strikes a favorable balance between temporal resolution and context retention, enabling D-FAD to bootstrap long-horizon predictions from stable short-term cues without incurring excessive autoregressive overhead.

6.4. Impact of Distillation Loss in D-FAD

We further analyze in Table 2 the role of the distillation loss $\mathcal{L}_{distill}$ in stabilizing autoregressive decoding. Specifically, we compare mean-squared-error (MSE), KL-divergence and mean-absolute-error (MAE) formulations under a fixed chunk size ($N = 8$). As shown in Table 2, the MSE-based distillation achieves superior performance (65.52% short, 62.53% long) compared to KL-divergence (64.63% short, 62.00% long) and MAE (65.43% short, 62.13% loss). This indicates that MSE provides a more stable alignment between autoregressive logits and full-decoder outputs, preserving both magnitude and directional consistency. KL-divergence, while effective for distribution matching, may introduce instability due to its sensitivity to softmax temperature and class imbalance. MAE, while giving competitive results, is experimentally not performing as good

Methods	Short (Top-1 mAP %)				Long (Top-1 mAP %)		
	1 sec.	2 sec.	3 sec.	Avg	4 sec.	8 sec.	Avg
*SoTA with **concat(Scene,Object)** Feature*							
OADTR [52]	63.07	62.40	62.58	62.68	62.06	59.13	60.59
FUTR [17]	61.47	61.21	61.74	61.47	62.79	56.46	59.62
LSTR [54]	63.06	62.00	63.06	62.70	63.06	59.63	61.34
JOADAA [20]	62.79	62.79	62.53	62.70	62.00	57.51	59.75
TesTra [61]	63.85	63.32	62.80	63.32	62.53	59.10	60.81
SIaT [33]	65.96	64.64	63.85	64.81	63.85	59.63	61.74
D^3P w/o. D-FAD	65.96	66.49	66.22	66.22	65.96	60.42	63.19
*SoTA with **our D-FAD** Decoder*							
OADTR [52] + D-FAD	64.37	64.64	65.69	64.90	64.11	60.15	62.13
FUTR [17] + D-FAD	63.58	63.58	64.11	63.75	63.85	57.78	60.81
LSTR [54] + D-FAD	65.69	64.37	64.37	64.81	63.58	59.10	61.34
JOADAA [20] + D-FAD	64.11	63.06	64.11	63.76	64.11	57.78	60.94
SIaT [33] + D-FAD	65.43	65.96	64.64	65.34	63.58	60.42	62.00
D^3P (our)	66.75	67.81	67.28	67.28	66.49	61.21	63.85

Table 3. State-of-the-art comparison on the AHB-F dataset abnormal human behavior forecasting. We report Top-1 mAP (%) across short-horizon (1–3 sec) and long-horizon (4–8 sec) intervals. Our full D^3P framework achieves the highest performance across all horizons, demonstrating superior short-term precision and long-term stability. Additionally, integrating our D-FAD decoder into existing SoTA models consistently boosts their forecasting accuracy, validating the robustness and plug-and-play versatility of our approach.

as MSE. These findings validate our choice of MSE as the preferred distillation objective, ensuring that D-FAD inherits the global consistency of conventional decoding while maintaining low-latency autoregressive inference.

7. State-of-the-Art Comparison

To rigorously evaluate the effectiveness and generalizability of our proposed D^3P framework, we conduct two distinct sets of experiments on the AHB-F dataset, the only benchmark dedicated to abnormal human behavior forecasting. These experiments are designed to assess both standalone performance and the plug-and-play versatility of our Distilled Future Auto-Regressive Decoder (D-FAD).

Experiment I: Comparison with Existing SoTA Architectures. In the first setup, we benchmark D^3P against leading action anticipation models including OADTR [52], FUTR [17], LSTR [54], JOADAA [20], TesTra [61], and SIaT [33], all operating on a unified feature space derived from concatenated scene and object-level embeddings. As shown in Table 3, our D^3P (w/o D-FAD) already surpasses these baselines, achieving a short-horizon average Top-1 mAP of 66.22% and a long-horizon average of 63.19%. This improvement stems from the Differential Past Encoder (DiPE), which disentangles noisy human–agent interactions and enhances discriminative cues through differential attention and Mamba-based temporal modeling. The consistent gains across all short-term intervals (1–3 sec) and long-term intervals (4–8 sec) highlight the encoder's ability to extract robust representations even in unconstrained environments.

Experiment II: Augmenting SoTA Models with D-FAD. In the second setup, we integrate our D-FAD decoder into each of the aforementioned SoTA models to evaluate its modularity and impact on long-horizon forecasting. Remarkably, all models exhibit performance improvements when equipped with D-FAD. For instance, OADTR+D-FAD improves from 62.68% to 64.90% in short-horizon average mAP, and from 60.59% to 62.13% in long-horizon average mAP. Similar trends are observed across FUTR, LSTR, JOADAA, and SIaT, validating the decoder's ability to stabilize predictions through autoregressive chunking and distillation. These results confirm that D-FAD is not only effective within our full pipeline but also enhances existing architectures in a plug-and-play fashion.

Overall Performance. Our complete D^3P framework, combining DiPE and D-FAD, achieves the highest performance across all horizons 67.28% short-horizon and 63.85% long-horizon average mAP, demonstrating superior precision, temporal stability, and computational efficiency. These results underscore the robustness and scalability of our approach, offering a unified solution for real-world anomaly anticipation that generalizes across models and temporal scales.

8. Conclusion

We introduced D^{3}P, a novel encoder–decoder framework for long-horizon abnormal human behavior forecasting in unconstrained real-world environments. By bridging denoised pasts with distilled autoregressive futures, D^{3}P directly addresses three core challenges: noisy human–agent interactions, weak temporal coupling between normal observations and distant anomalies, and data scarcity in autoregressive modeling. Our Differential Past Encoder (DiPE) disentangles scene-level and object-level dynamics using differential attention and Mamba-based temporal modeling, while our Distilled Future Auto-Regressive Decoder (D-FAD) stabilizes long-range predictions through sequential chunking and knowledge distillation. Extensive experiments on the AHB-F benchmark demonstrate that D^{3}P consistently outperforms state-of-the-art methods across both short and long forecasting horizons. Moreover, D-FAD enhances existing architectures in a plug-and-play fashion, validating the robustness and modularity of our design. These results establish D^{3}P as a scalable and efficient solution for proactive anomaly anticipation in safety-critical applications.

In the future, we envision integrating large language models (LLMs) to complement visual forecasting with semantic reasoning and interpretability. By leveraging LLMs to generate causal narratives or counterfactual explanations for predicted anomalies, future systems could offer human-understandable justifications for long-horizon forecasts, bridging the gap between predictive performance and actionable insight. This opens promising directions for explainable anticipation in surveillance.

Acknowledgments: This work was supported by Toyota Motor Europe (TME) and the French government, through the 3IA Cote d'Azur Investments managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002.

References

[1] Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what?-anticipating temporal occurrences of activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5343–5352, 2018. 2

[2] Giacomo D' Amicantonio, Snehashis Majhi, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, Francois Bremond, and Egor Bondarev. Mixture of experts guided by gaussian splatters matters: A new approach to weakly-supervised video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10275–10285, 2025. 2

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 2

[4] Congqi Cao, Yue Lu, Peng Wang, and Yanning Zhang. A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20392–20401, 2023. 2

[5] Congqi Cao, Lanshu Hu, Yating Yu, and Yanning Zhang. Vision and intention boost large language model in long-term action anticipation. *arXiv preprint arXiv:2505.01713*, 2025. 2

[6] Junwen Chen, Gaurav Mittal, Ye Yu, Yu Kong, and Mei Chen. Gatehub: Gated history unit with background suppression for online action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19925–19934, 2022. 2

[7] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 387–395, 2023. 2

[8] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. 2021. 4

[9] MyeongAh Cho, Minjung Kim, Sangwon Hwang, Chaewon Park, Kyungjae Lee, and Sangyoun Lee. Look around for anomalies: Weakly-supervised anomaly detection via context-motion relational learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12137–12146, 2023. 2

[10] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113, 2023. 2

[11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018. 2

[12] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2022. 2

[13] Antonino Furnari and Giovanni Maria Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4021–4036, 2020. 2

[14] Antonino Furnari and Giovanni Maria Farinella. Towards streaming egocentric action anticipation. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1250–1257. IEEE, 2022. 2

[15] Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Predicting the future: A jointly learnt model for action anticipation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5562–5571, 2019. 2

[16] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. Red: Reinforced encoder-decoder networks for action anticipation. *arXiv preprint arXiv:1707.04818*, 2017. 2

[17] Dayoung Gong, Joonseok Lee, Manjin Kim, Seong Jong Ha, and Minsu Cho. Future transformer for long-term action anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3052–3061, 2022. 2, 7

[18] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 2

[19] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 4

[20] Mohammed Guermal, Abid Ali, Rui Dai, and François Brémond. Joadaa: Joint online action detection and action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6889–6898, 2024. 2, 7

[21] Sanghwan Kim, Daoji Huang, Yongqin Xian, Otmar Hilliges, Luc Van Gool, and Xi Wang. Palm: Predicting actions through language models. In *European Conference on Computer Vision*, pages 140–158. Springer, 2024. 2

[22] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[23] Colin Lea, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks: A unified approach to action segmentation. In *European conference on computer vision*, pages 47–54. Springer, 2016. 2

[24] Shuheng Lin, Hua Yang, Xianchao Tang, Tianqi Shi, and Lin Chen. Social mil: Interaction-aware for crowd anomaly detection. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE, 2019. 2

[25] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013. 2

[26] Snehashis Majhi, Ratnakar Dash, and Pankaj Kumar Sa. Temporal pooling in inflated 3dcnn for weakly-supervised video anomaly detection. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE, 2020. 2

[27] Snehashis Majhi, Srijan Das, François Brémond, Ratnakar Dash, and Pankaj Kumar Sa. Weakly-supervised joint anomaly detection and classification. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–7. IEEE, 2021.

[28] Snehashis Majhi, Srijan Das, and François Brémond. Dam: Dissimilarity attention module for weakly-supervised video anomaly detection. In *2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8, 2021.

[29] Snehashis Majhi, Deepak Ranjan Nayak, Ratnakar Dash, and Pankaj Kumar Sa. Multi-level 3dcnn with min-max ranking loss for weakly-supervised video anomaly detection. In *International Conference on Neural Information Processing*, pages 25–37. Springer, 2022.

[30] Snehashis Majhi, Rui Dai, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. Human-scene network: A novel baseline with self-rectifying loss for weakly supervised video anomaly detection. *Computer Vision and Image Understanding*, 241:103955, 2024.

[31] Snehashis Majhi, Rui Dai, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, and François Brémond. Oectst: Outlier-embedded cross temporal scale transformer for weakly-supervised video anomaly detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 8574–8583, 2024.

[32] Snehashis Majhi, Giacomo D'Amicantonio, Antitza Dantcheva, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, Egor Bondarev, and Francois Bremond. Just dance with pi! a poly-modal inductor for weakly-supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24265–24274, 2025. 2

[33] Snehashis Majhi, Mohammed Guermal, Antitza Dantcheva, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. Guess future anomalies from normalcy: Forecasting abnormal behavior in real-world videos. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 388–398. IEEE Computer Society, 2025. 2, 4, 5, 7

[34] Esteve Valls Mascaró, Hyemin Ahn, and Dongheui Lee. Intention-conditioned long-term human egocentric action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6048–6057, 2023. 2

[35] Himangi Mittal, Nakul Agarwal, Shao-Yuan Lo, and Kwonjoon Lee. Can't make an omelette without breaking some eggs: Plausible action anticipation using large video-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18580–18590, 2024. 2

[36] Nada Osman, Guglielmo Camporese, Pasquale Coscia, and Lamberto Ballan. Slowfast rolling-unrolling lstms for action anticipation in egocentric videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3437–3445, 2021. 2

[37] Seongheon Park, Hanjae Kim, Minsu Kim, Dahye Kim, and Kwanghoon Sohn. Normality guided multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2665–2674, 2023. 2

[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An im-

perative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[39] Zhaobo Qi, Shuhui Wang, Chi Su, Li Su, Qingming Huang, and Qi Tian. Self-regulated learning for egocentric video activity anticipation. *IEEE transactions on pattern analysis and machine intelligence*, 45(6):6715–6730, 2021. 2

[40] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[41] Mamshad Nayeem Rizve, Gaurav Mittal, Ye Yu, Matthew Hall, Sandra Sajeev, Mubarak Shah, and Mei Chen. Pivotal: Prior-driven supervision for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22992–23002, 2023. 2

[42] Royston Rodrigues, Neha Bhargava, Rajbabu Velmurugan, and Subhasis Chaudhuri. Multi-timescale trajectory prediction for abnormal human activity detection. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020. 2

[43] Debaditya Roy, Ramanathan Rajendiran, and Basura Fernando. Interaction visual transformer for egocentric action anticipation. *arXiv preprint arXiv:2211.14154*, 2022. 2

[44] Debaditya Roy, Ramanathan Rajendiran, and Basura Fernando. Interaction region visual transformer for egocentric action anticipation. *arXiv preprint arXiv:2211.14154*, 2022. 2

[45] Yuge Shi, Basura Fernando, and Richard Hartley. Action anticipation with rbf kernelized feature mapping rnn. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 301–317, 2018. 2

[46] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6479–6488, 2018. 2

[47] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2018. 2

[48] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4975–4986, 2021. 2

[49] 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 preprint arXiv:2302.13971*, 2023. 1, 2

[50] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 98–106, 2016. 2

[51] Boyang Wan, Yuming Fang, Xue Xia, and Jiajie Mei. Weakly supervised video anomaly detection via center-guided discriminative learning. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020. 2

[52] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Yuanjie Shao, Zhengrong Zuo, Changxin Gao, and Nong Sang. Oadtr: Online action detection with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7565–7575, 2021. 2, 7

[53] 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 *European Conference on Computer Vision*, pages 322–339. Springer, 2020. 2

[54] Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Xia, Zhuowen Tu, and Stefano Soatto. Long short-term transformer for online action detection. *Advances in Neural Information Processing Systems*, 34:1086–1099, 2021. 2, 7

[55] Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. Differential transformer. *arXiv preprint arXiv:2410.05258*, 2024. 4

[56] Muhammad Zaigham Zaheer, Arif Mahmood, Hochul Shin, and Seung-Ik Lee. A self-reasoning framework for anomaly detection using video-level labels. *IEEE Signal Processing Letters*, 27:1705–1709, 2020. 2

[57] Olga Zatsarynna, Emad Bahrami, Yazan Abu Farha, Gianpiero Francesca, and Juergen Gall. Gated temporal diffusion for stochastic long-term dense anticipation. In *European Conference on Computer Vision*, pages 454–472. Springer, 2024. 2

[58] Olga Zatsarynna, Emad Bahrami, Yazan Abu Farha, Gianpiero Francesca, and Juergen Gall. Manta: Diffusion mamba for efficient and effective stochastic long-term dense anticipation. *arXiv preprint arXiv:2501.08837*, 2025. 2

[59] Jiangong Zhang, Laiyun Qing, and Jun Miao. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4030–4034. IEEE, 2019. 2

[60] Qi Zhao, Shijie Wang, Ce Zhang, Changcheng Fu, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. Antgpt: Can large language models help long-term action anticipation from videos? *arXiv preprint arXiv:2307.16368*, 2023. 2

[61] Yue Zhao and Philipp Krähenbühl. Real-time online video detection with temporal smoothing transformers. In *European Conference on Computer Vision*, pages 485–502. Springer, 2022. 2, 7

[62] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[63] Zeyun Zhong, Chengzhi Wu, Manuel Martin, Michael Voit, Juergen Gall, and Jürgen Beyerer. Diffant: Diffusion models for action anticipation. *arXiv preprint arXiv:2311.15991*, 2023. 2