SAPIENZA
UNIVERSITÀ DI ROMA

Improving context in Video Anomaly Detection with Diffusion Models and State Space Models

Faculty of Information Engineering, Computer Science and Statistics
Master Degree in Computer Science (XXXVII cycle)

Riccardo Capobianco
ID number 1884636

Advisor
Prof. Fabio Galasso

Academic Year 2024/2025

Thesis not yet defended

Improving context in Video Anomaly Detection with Diffusion Models and State Space Models

Master thesis. Sapienza University of Rome

This thesis has been typeset by LaTeX and the Sapthesis class.

Author's email: capobianco.1884636@studenti.uniroma1.it

Abstract

Video Anomaly Detection (VAD) aims to identify unusual spatio-temporal patterns in video data, a task requiring long-range temporal modeling, spatial coherence, and sensitivity to fine-grained motion. This thesis explores state-space modeling alternatives to attention within diffusion-based generative models for unsupervised VAD. Specifically, we adapt and optimize the Random-Mask Video Diffusion (RaMViD) framework, originally designed for video prediction and infilling, by re-purposing it for the VAD setting and replacing attention modules with Mamba state-space blocks, enabling more efficient modeling of long temporal sequences. To further improve reliability, we introduce a custom anomaly scoring function tailored to stabilize detection across heterogeneous scenes. Experiments on the Avenue, ShanghaiTech Campus, and UBnormal datasets demonstrate competitive accuracy and consistent anomaly localization, highlighting the benefits of adapting RaMViD with Mamba to the VAD scenario and of our custom scoring function over existing ones. These results suggest state-space diffusion models as a promising direction for scalable and robust VAD.

Contents

Chapter 1

Introduction

1.1 Motivation

Video Anomaly Detection (VAD) has emerged as a critical research area in computer vision, with applications spanning public safety, surveillance, traffic monitoring, and industrial automation. The central objective of VAD is to identify unusual spatio-temporal patterns in videos that deviate from an established notion of "normality." This requires simultaneously modeling spatial coherence, fine-grained motion, and long-range temporal dependencies. Despite substantial progress in the field, the reliable detection of anomalies in complex, real-world scenarios remains a difficult challenge.

The ability to detect anomalies is of both practical and scientific importance. In public security, rapid identification of unusual events such as fights or accidents can trigger immediate interventions. In traffic monitoring, detecting abnormal pedestrian or vehicle trajectories improves safety and efficiency. In industrial settings, anomaly detection can prevent costly breakdowns by identifying malfunctioning machinery. These use cases highlight the necessity of robust and scalable VAD models.

1.2 Open Challenges

A pervasive limitation in existing VAD approaches lies in their reliance on short temporal contexts. Many state-of-the-art systems only process a few frames at a time, which restricts their ability to capture long-range dependencies across a video sequence. This constraint significantly reduces robustness in scenes characterized by complex dynamics, where anomalies may unfold gradually or depend on distant temporal context.

Another key challenge concerns computational efficiency. Attention-based architectures, while powerful, suffer from quadratic complexity with respect to sequence length, which makes them unsuitable for long videos or resource-constrained environments. This bottleneck not only limits scalability but also prevents the adoption

of larger temporal windows that are often required to capture realistic anomaly dynamics.

Finally, VAD is complicated by dataset heterogeneity and scene variance. Models trained on specific environments often generalize poorly when applied to new settings. This variance leads to unstable detection and requires additional mechanisms to ensure reliability across diverse video sources.

To address these challenges, this work explores the integration of state-space models within diffusion-based frameworks, aiming to extend temporal context without incurring quadratic costs. By replacing attention modules with efficient state-space mechanisms and introducing a custom scoring function to stabilize detection across heterogeneous scenes, we propose a scalable and robust approach to long-range Video Anomaly Detection.

1.3 Objectives and Contributions

This thesis addresses the above challenges by proposing a diffusion-based generative framework tailored to unsupervised video anomaly detection. The key contributions can be summarized as follows:

- **Optimization of RaMViD for VAD.** We adapt the Random-Mask Video Diffusion (RaMViD) framework, originally designed for video prediction and infilling, to the anomaly detection setting. The modifications include replacing random masking with a fixed selective masking strategy, which reduces training time and improves inference stability, and a sliding window at inference.
- **Replacing attention with Mamba state-space blocks.** We substitute costly attention mechanisms with state-space sequence modeling using bidirectional Mamba blocks. This change enables modeling longer temporal contexts with linear complexity in sequence length, while preserving the spatial inductive bias of 3D convolutional layers.
- **Custom anomaly scoring function.** To further stabilize detection across heterogeneous scenes, we introduce a novel Blob-Loss scoring function. This custom score emphasizes spatially coherent differences rather than pixel-level noise, improving the robustness of anomaly localization.
- **Comprehensive evaluation.** We validate the proposed approach on three widely used benchmarks: CUHK Avenue, ShanghaiTech Campus (STC), and UBnormal. Results demonstrate competitive accuracy and consistent anomaly localization, highlighting the effectiveness of combining diffusion models with state-space modeling.

Chapter 2

Background

2.1 Fundamentals of Video Anomaly Detection

Video anomaly detection aims to identify spatio-temporal patterns in video sequences that deviate from a learned representation of normality. Unlike supervised classification problems, anomaly detection is typically framed as a one-class learning task: models are trained only on normal data and are required to detect deviations without direct examples of anomalies. This formulation reflects the inherent scarcity and unpredictability of abnormal events in real-world scenarios.

The central challenge of VAD lies in modeling the joint distribution of spatial appearance and temporal dynamics. Normal videos display coherent motion patterns, such as pedestrians walking or vehicles moving along roads, whereas anomalies introduce deviations such as unexpected behaviors, abnormal trajectories, or rare events. An effective VAD model must therefore preserve spatial consistency across frames while capturing temporal dependencies that extend beyond local neighborhoods.

2.2 Diffusion Models for Generative Learning

Diffusion models are a class of generative frameworks that have proven extremely effective in creating high-quality data such as images, audio, and video. The key idea is to start from real data and gradually add noise until the original structure is completely destroyed, reaching a state that looks like random noise. The model is then trained to reverse this process step by step, learning the noise to be removed in order to recover the original data. Once trained, this allows the model to start from pure noise and progressively denoise it, producing realistic new samples.

This iterative refinement process gives diffusion models a unique advantage: instead of generating everything in a single step, they build the result gradually, which helps capture fine details and complex structures that are often missed by other generative models such as GANs or VAEs. Furthermore, diffusion models can be conditioned on additional information, for example a class label, a text

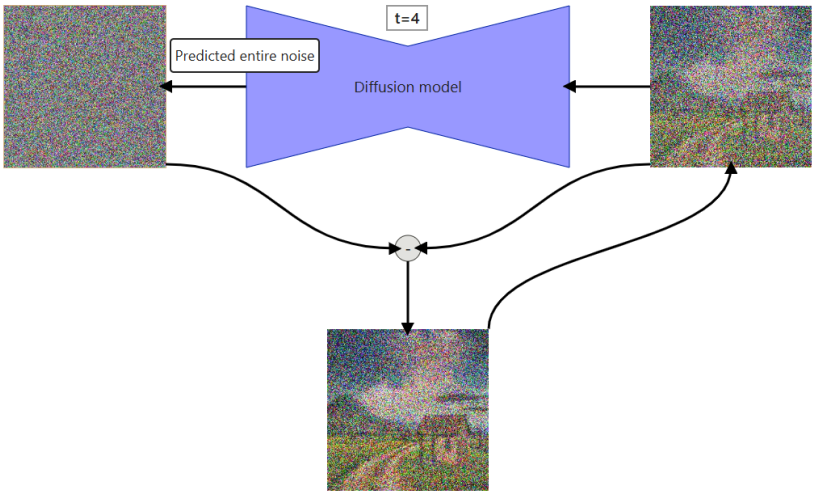

Figure 2.1. Example of the denoising process in a diffusion model.

description, or the surrounding frames in a video, making them highly flexible and adaptable to many tasks.

In video applications, diffusion models are particularly powerful because they can fill in missing or corrupted frames by using the temporal context provided by neighboring frames. This is especially relevant for VAD, where the goal is to detect unusual events that do not fit the normal dynamics of a scene. If a diffusion model has learned how "normal" frames typically look and evolve, it will reconstruct them accurately. When faced with abnormal or unexpected events, however, the reconstruction will be less accurate, producing larger errors that can be used as anomaly indicators.

An example of this approach is the Random-Mask Video Diffusion (RaMViD) model [1]. Instead of always reconstructing the entire video sequence, RaMViD randomly hides some frames and forces the model to predict them from the visible ones. This strategy trains the model to reason over time and to become very good at infilling missing parts. For anomaly detection, this ability is crucial: while the model easily reconstructs typical patterns of motion and appearance, it struggles with unusual events, and this difference provides a reliable anomaly signal.

In summary, diffusion models combine a noise-based training scheme with an iterative reconstruction process. Their ability to condition on temporal context and to generate missing content makes them an ideal foundation for robust video anomaly detection systems.

2.3 Attention vs. State-Space Models

Transformer architectures based on self-attention have been the dominant paradigm in sequence modeling for the past years, achieving state-of-the-art performance in

domains such as natural language processing, speech recognition, and computer vision. The self-attention mechanism computes interactions between all pairs of elements in a sequence, allowing the model to capture rich contextual relationships. However, this comes at a computational cost: the attention operation scales quadratically with sequence length. As a result, when applied to long video sequences, attention becomes prohibitively expensive in both memory and time. This limitation forces practitioners to restrict the temporal context, which hinders the ability to capture long-range dependencies that are critical in Video Anomaly Detection (VAD).

State-Space Models (SSMs) [2] have recently re-emerged as a powerful alternative for sequence modeling. Instead of explicitly computing pairwise interactions, SSMs rely on a continuous-time dynamical system defined by hidden states that evolve over time and are updated based on input signals. Importantly, many modern SSM formulations achieve *linear* computational complexity with respect to sequence length, making them highly efficient for processing long sequences. This efficiency enables the use of much larger temporal windows, which is particularly valuable in video applications where anomalous events may span hundreds of frames.

Among the most recent advances in this family of models is the **Mamba** architecture [3], a selective state-space model designed to combine the strengths of attention with the efficiency of SSMs. Mamba introduces a mechanism that dynamically selects relevant information to propagate through its state updates, effectively filtering out irrelevant parts of the sequence while maintaining global contextual awareness. This selective mechanism allows Mamba to focus computational resources on informative regions, similar to how attention highlights relevant tokens, but without incurring quadratic cost. As a result, Mamba achieves linear scaling while retaining the modeling power typically associated with attention.

For video tasks, this property is particularly attractive. Video data inherently involves both spatial and temporal dependencies, and the ability to process long sequences is crucial to detect subtle or delayed anomalies. By replacing attention modules with Mamba blocks in the diffusion framework, this thesis leverages the efficiency of state-space models while preserving the capacity to model complex spatio-temporal patterns. In practice, Mamba enables diffusion-based architectures to handle much longer video contexts, providing a richer temporal representation that improves robustness in Video Anomaly Detection.

Chapter 3

Related Work

Video anomaly detection (VAD) has been extensively explored through prediction, reconstruction, and memory-driven approaches, each making different assumptions about normality and how to capture deviations. In this chapter, we analyze three representative algorithms, MA-PDM, STATE, and LGN-Net, detailing their architectures and highlighting the main advantages and limitations with respect to our method.

3.1 Prediction-based Approaches

Prediction models aim to forecast future frames under the assumption that anomalies manifest as deviations from normal temporal dynamics. A recent example is the Motion and Appearance guided Patch Diffusion Model (MA-PDM) by Zhou et al. [4]. This method reframes VAD as a generation problem within a diffusion framework. Instead of reconstructing the entire frame, MA-PDM decomposes each sequence into *appearance* (the first frame) and *motion* (temporal differences between consecutive frames), which act as conditions guiding the diffusion process.

The architecture combines three main modules: (i) a patch-cropping strategy that extracts local regions, (ii) an appearance encoder with a learnable patch-memory bank that preserves semantic normality, and (iii) a U-Net noise estimation network that integrates motion and appearance cues to predict the denoising trajectory. During training, noise is injected at the patch level, and the model learns to predict the forward noise. At inference, overlapping patches are reconstructed via a sliding-window strategy and merged to recover the full frame. Anomalies are detected by computing the mean squared error (MSE) between predicted and ground-truth frames.

By operating on patches, MA-PDM captures fine-grained local deviations that may otherwise be diluted in full-frame models. However, the method typically conditions on only six past frames, which restricts its ability to capture long-range temporal dependencies.

3.2 Reconstruction-based Approaches

Reconstruction-based methods learn to reproduce normal patterns and detect anomalies when reconstruction errors are high. Wang et al. [5] propose the Spatio-Temporal Auto-Trans-Encoder (STATE), which rethinks reconstruction with a transformer-inspired architecture.

STATE operates at the object level: given a frame, a pre-trained detector extracts bounding boxes, and each object is expanded into a *Spatio-Temporal Context Cube* (STCC) that spans several preceding and following frames. Each STCC is processed by a convolutional encoder–decoder, where temporal reasoning is handled by a stack of *multi-head learnable convolutional attention* layers. These layers allow features from different time steps to attend to each other, providing richer temporal modeling compared to conventional autoencoders.

The model reconstructs object-centric patches along two parallel branches: one for raw appearance (RGB) and one for motion (optical flow, computed via FlowNet2). During testing, Wang et al. introduce an *input perturbation* strategy: frames are perturbed via the gradient of the reconstruction loss, which reduces errors on normal samples but not on anomalies, thereby amplifying their separability. The final anomaly score combines reconstruction errors from both branches.

STATE addresses the weaknesses of convolutional AEs, overfitting and poor temporal reasoning, by integrating attention and perturbation. However, its attention layers scale quadratically with sequence length, which limits its scalability to long temporal contexts.

3.3 Memory based Approaches

Memory-based methods explicitly store representations of normality and use them to constrain reconstruction or prediction. Zhao et al. [6] introduce LGN-Net (Local-Global Normality Network), which combines local spatio-temporal modeling with a global memory of prototypes.

The architecture has a dual-branch design: (i) a convolutional encoder with LSTM units extracts local motion-sensitive features, and (ii) an external memory module stores prototypical embeddings of normal events. During inference, the model retrieves the most relevant prototypes from memory and fuses them with local features to reconstruct or predict the next frame. This mechanism helps suppress false positives by anchoring predictions to stored normal patterns, especially when the local features are ambiguous.

LGN-Net effectively balances local detail with global context, reducing overgeneralization that often plagues AE-based methods. However, its reliance on relatively short input windows (five frames) limits its capacity to model long-term dependencies. Moreover, the memory bank must be carefully managed to avoid redundancy or drift, which can reduce effectiveness in highly heterogeneous datasets.

3.4 Hybrid and State-of-the-Art Models

Several recent works combine prediction, reconstruction, and memory in order to exploit their complementary strengths. While these hybrids improve robustness, they remain constrained by limited context length and computational overhead.

Our method advances this line of research by extending the effective temporal horizon to 30 frames per generation, significantly longer than MA-PDM, STATE, or LGN-Net, while maintaining efficiency. Furthermore, we replace self-attention with state-space models, which scale linearly with sequence length and improve stability during training and inference. This design allows us to bridge the gap between spatial fidelity and long-range temporal reasoning in VAD.

Chapter 4

Methodology

4.1 Overview

We formulate VAD as a *conditional video infilling* task within a diffusion framework. Given a temporal window of L frames, a subset of indices C is selected as the *anchor set*, while the remaining frames $U = \{0, \ldots, L-1\} \setminus C$ are corrupted by the diffusion forward process and subsequently reconstructed by the model. The discrepancy between the reconstructed frames and the ground-truth observations provides an *anomaly score*, with larger reconstruction errors indicating a higher likelihood of abnormality.

Our approach builds upon the Random-Mask Video Diffusion (RaMViD) framework [1], originally introduced for video prediction and infilling, but introduces several key modifications tailored to anomaly detection:

- **State-space modeling with Mamba.** We replace standard attention mechanisms with Mamba blocks, a class of state-space models that scale efficiently to long temporal sequences while preserving spatial-temporal coherence. This allows the model to capture long-range dependencies critical for anomaly detection, mitigating the context limitations of attention-based diffusion models.

- **Selective masking.** Instead of training with random masking, which forces the model to handle a wide variety of conditioning patterns, we adopt a deterministic selective masking strategy aligned with the masking pattern used at inference. This focuses learning on the exact infilling regime required for VAD, improving reconstruction quality and inference efficiency.

- **Sliding-window inference.** During evaluation, we process the video stream with overlapping windows of length L, enabling continuous anomaly scoring across long sequences. This sliding-window mechanism ensures temporal coverage and smooth detection without loss of context.

- **Task-aware anomaly scoring.** We introduce a custom scoring function that refines the raw reconstruction error into a more stable and discriminative anomaly measure. This adjustment improves robustness across scenes, yielding consistent detection performance.

Together, these modifications transform RaMViD into a diffusion-based framework specifically optimized for video anomaly detection, balancing modeling capacity, temporal scalability, and inference efficiency.

4.2 From Random to Selective Masking

The masking strategy is a central component of the RaMViD framework, as it defines how context is provided to the model during training and inference. By determining which frames remain observable and which are corrupted, masking directly shapes the learning objective and the reconstruction regime. In this section, we first formalize the original *random masking* strategy, which was introduced in RaMViD to promote generalization across tasks, and then we motivate and describe our proposed refinement, *selective masking*, designed specifically for video anomaly detection (VAD), where the inference masking pattern is known in advance.

4.2.1 Random Masking in RaMViD

A fundamental feature of the RaMViD framework is the use of *random masking* during training. Formally, let a video be represented as a sequence of L consecutive frames:

$$x = \{x_0, x_1, \ldots, x_{L-1}\}, \quad x_t \in \mathbb{R}^{H \times W \times C},$$

where H and W denote the spatial resolution of each frame and C the number of channels (typically $C = 3$ for RGB). The random variable x is assumed to be sampled from the underlying data distribution $p(x)$.

At each training step, a subset of frame indices is selected to form the *anchor set*:

$$C \subset \{0, \ldots, L-1\},$$

which contains the indices of frames that remain uncorrupted and serve as conditioning context. Its complement is defined as:

$$U = \{0, \ldots, L-1\} \setminus C,$$

corresponding to the set of masked indices whose content is noised and reconstructed during training.

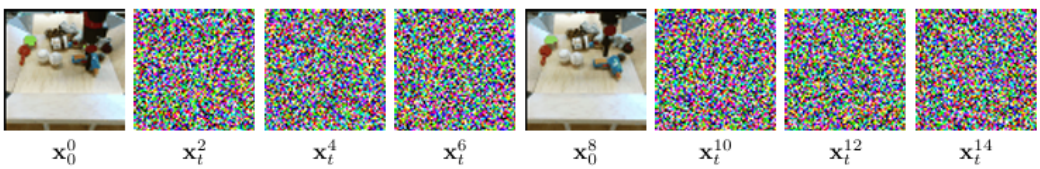

Figure 4.1. Illustration of the random masking strategy in RaMViD. The anchor set C (clean frames) provides context, while the complement U (masked frames) is corrupted and reconstructed. Figure taken from RaMViD paper.

The corrupted input at diffusion time step t can be written as:

$$x_t = x_t^U \oplus x_0^C,$$

where x_t^U denotes the frames in U perturbed by the forward diffusion process, x_0^C are the clean anchor frames, and $\oplus$ represents the concatenation of the two subsets into a single mixed sequence.

The model is then trained to reconstruct the original frames in U, conditioned on the clean anchors in C. The reconstruction loss is therefore computed only on U.

The intuition behind this procedure is that random masking exposes the model to a wide variety of reconstruction patterns during training. This was originally motivated by the goal of better generalization: at inference time, different tasks such as video prediction or video infilling require different masking patterns, and a model trained with random conditioning is expected to adapt to all of them.

4.2.2 Motivation for Selective Masking

While random masking encourages generalization, it is not always optimal in the context of video anomaly detection (VAD). In this setting, the masking pattern used at inference time is known *a priori*. Training the model with arbitrary random conditioning forces it to allocate capacity to masking patterns that will never occur at test time, thereby wasting representational power and possibly degrading anomaly detection performance.

For this reason, we replace random conditioning with a *deterministic selective masking* strategy. Unlike in RaMViD, where the anchor set C is resampled at each step, selective masking fixes C deterministically within every training window.

4.2.3 Selective Masking in VAD

In our selective masking strategy, the anchor set C is chosen according to the same regime that will be used during inference. Specifically, C includes:

- the first two frames of the sequence,
- every 10th and 11th frame,
- the last two frames.

The unknown set is then given by:

$$U = \{0, \dots, L-1\} \setminus C,$$

and is always corrupted by the forward diffusion process. The network input retains the same mixed structure:

$$x_t = x_t^U \oplus x_0^C,$$

but now the distribution of the anchor set C is no longer random; it is deterministically fixed. The reconstruction loss is computed exclusively on U, ensuring that the learning process focuses on the precise infilling regime used during evaluation.

The diffusion objective and noise schedule remain unchanged with respect to RaMViD; the only modification lies in the distribution of anchor indices C.

(a) Random masking

(b) Selective masking

Figure 4.2. Comparison between random and selective masking both at 200k training steps and 200 diffusion steps at inference

4.2.4 Empirical Observations

Early qualitative experiments revealed immediate benefits of selective masking over random masking. One of the most important advantages of selective masking is its impact on inference efficiency. The improved reconstruction quality enabled us to reduce the number of diffusion steps from 1000 to 200 without any measurable loss in anomaly detection metrics, yielding an approximate 5× speedup. In contrast, when trained with random masking, the same 200-step sampler produced highly noisy reconstructions, rendering them less suitable for evaluation.

4.3 Replacing Attention with Mamba State-Space Blocks

To overcome the quadratic complexity of attention, we integrate *bidirectional Mamba* state-space blocks at selected resolutions of the diffusion U-Net. Input features $\mathbf{x} \in \mathbb{R}^{B \times C \times F \times H \times W}$ are reshaped into sequences of length $L = FHW$, and Mamba layers perform linear-time sequence mixing. The output is merged with the original representation via additive residual connections, while 3D convolutional blocks preserve local spatial inductive bias. This hybrid design allows efficient modeling of spatio-temporal dependencies with $\mathcal{O}(L)$ complexity, supporting longer sequences without prohibitive computational cost.

4.3.1 U-Net Backbone in RaMViD

The diffusion model relies on a *U-Net* backbone [7], an encoder–decoder architecture with skip connections that preserve spatial details during reconstruction. In RaMViD,

3D convolutional blocks are used within the U-Net to encode local spatio-temporal structure. By replacing self-attention with Mamba, we maintain the U-Net's inductive biases while enhancing its ability to model long-range dependencies efficiently. This combination of convolutional locality and state-space global mixing provides a strong backbone for video anomaly detection.

4.3.2 Replacing Self-Attention with Mamba

In the original RaMViD architecture [1], self-attention blocks are inserted at selected U-Net resolutions to capture long-range dependencies. However, self-attention scales quadratically with the flattened sequence length $L = FHW$, leading to high computational and memory cost:

$$\text{Attention}: \quad \mathcal{O}(L^2).$$

We replace these blocks with bidirectional Mamba layers, which scan the sequence in both directions with linear complexity:

$$\text{Mamba}: \quad \mathcal{O}(L).$$

Concretely, the replacement works as follows:

1. Features $\mathbf{x} \in \mathbb{R}^{B\times C\times F\times H\times W}$ are flattened along space and time into $\tilde{\mathbf{x}} \in \mathbb{R}^{B\times L\times C}$ with $L = FHW$.
2. The bidirectional Mamba block processes this sequence, yielding $\tilde{\mathbf{y}} = \text{Mamba}_{\leftrightarrow}(\tilde{\mathbf{x}})$.
3. The output is reshaped back to the original 5D layout, $\text{reshape}(\tilde{\mathbf{y}}) \in \mathbb{R}^{B\times C\times F\times H\times W}$.
4. Finally, the result is combined with the original features through an additive residual connection:

$$\mathbf{y} = \mathbf{x} + \text{reshape}\big(\text{Mamba}_{\leftrightarrow}(\tilde{\mathbf{x}})\big).$$

This residual connection means that the model does not discard the original convolutional features $\mathbf{x}$, but instead enriches them with the global spatio-temporal mixing provided by Mamba. In practice, $\mathbf{x}$ retains local spatial structure, while $\text{reshape}(\tilde{\mathbf{y}})$ contributes long-range dependencies, and the sum $\mathbf{y}$ fuses the two.

4.3.3 Bidirectional Sequence Mixing

The bidirectional formulation ensures that each position in the sequence has access to both past and future context. For each index $\ell \in \{1, \dots, L\}$:

$$\mathbf{h}_\ell^{\rightarrow} = \mathcal{S}(\tilde{\mathbf{x}}_{1:\ell}), \qquad \mathbf{h}_\ell^{\leftarrow} = \mathcal{S}^{\leftarrow}(\tilde{\mathbf{x}}_{\ell:L}),$$

and the two directional states are fused as

$$\tilde{\mathbf{y}}_\ell = \phi\big(\mathbf{h}_\ell^{\rightarrow}, \mathbf{h}_\ell^{\leftarrow}\big).$$

This bidirectional mechanism is crucial for conditional video infilling in anomaly detection, since masked frames benefit from both preceding and succeeding temporal context. In practice, we build upon the publicly available implementation from the *VideoMamba* paper [8], which we extend to integrate a bidirectional Mamba module within our diffusion-based framework.

4.3.4 Residual Connections and Gating

Residual connections stabilize training by letting each layer learn a *correction* to its input rather than a complete replacement. In our case:

$$\mathbf{y} = \mathbf{x} + \text{reshape}(\tilde{\mathbf{y}}),$$

acts as a fusion of local and global information: $\mathbf{x}$ comes from 3D convolutional residual blocks (local inductive bias), while reshape($\tilde{\mathbf{y}}$) injects long-range temporal dependencies captured by Mamba.

In addition, Mamba employs an *input-dependent gating mechanism*, which dynamically modulates internal parameters based on the input sequence. Intuitively, gating acts as a filter: it amplifies informative temporal patterns while suppressing irrelevant ones. This makes the sequence mixing more adaptive and effective than static recurrences, especially in complex video streams.

4.3.5 Impact of Replacing Attention with Mamba

Replacing the attention modules with Mamba state-space blocks proved highly effective in extending the temporal context of the model. Specifically, this substitution enabled the model to handle approximately 20% more frames within the same memory budget, directly addressing the quadratic cost bottleneck of self-attention. Notably, even employing the bidirectional Mamba formulation, which requires nearly twice the number of parameters compared to the standard unidirectional version, the framework maintained this efficiency gain. This demonstrates that Mamba-based sequence modeling can scale to longer contexts without incurring prohibitive memory usage, while still benefiting from richer bidirectional temporal representations crucial for anomaly detection.

4.4 Overlapping Infilling and Max Aggregation

At inference time, the video is divided into overlapping temporal windows of fixed length $L = 30$ frames. For each window, only a subset of frames (the anchors C) is revealed to the model, while the remaining frames $U = \{0, \ldots, L-1\} \setminus C$ are reconstructed through the diffusion process. To ensure broader temporal coverage, we slide the window along the sequence with stride $S = 15$. This overlapping strategy guarantees that each frame x_t appears at least once as part of the masked set U.

Formally, let $\hat{x}_t^{(1)}$ and $\hat{x}_t^{(2)}$ denote the two reconstructions of the same frame x_t, obtained from the two overlapping windows in which t is included. Each reconstruction is evaluated with a frame-wise reconstruction loss function $\mathcal{L}(\cdot, \cdot)$, typically defined as the mean squared error (MSE) or the custom Blob-Loss. We then define the final anomaly score for frame t as:

$$s_t = \max\left(\mathcal{L}(x_t, \hat{x}_t^{(1)}),\ \mathcal{L}(x_t, \hat{x}_t^{(2)})\right).$$

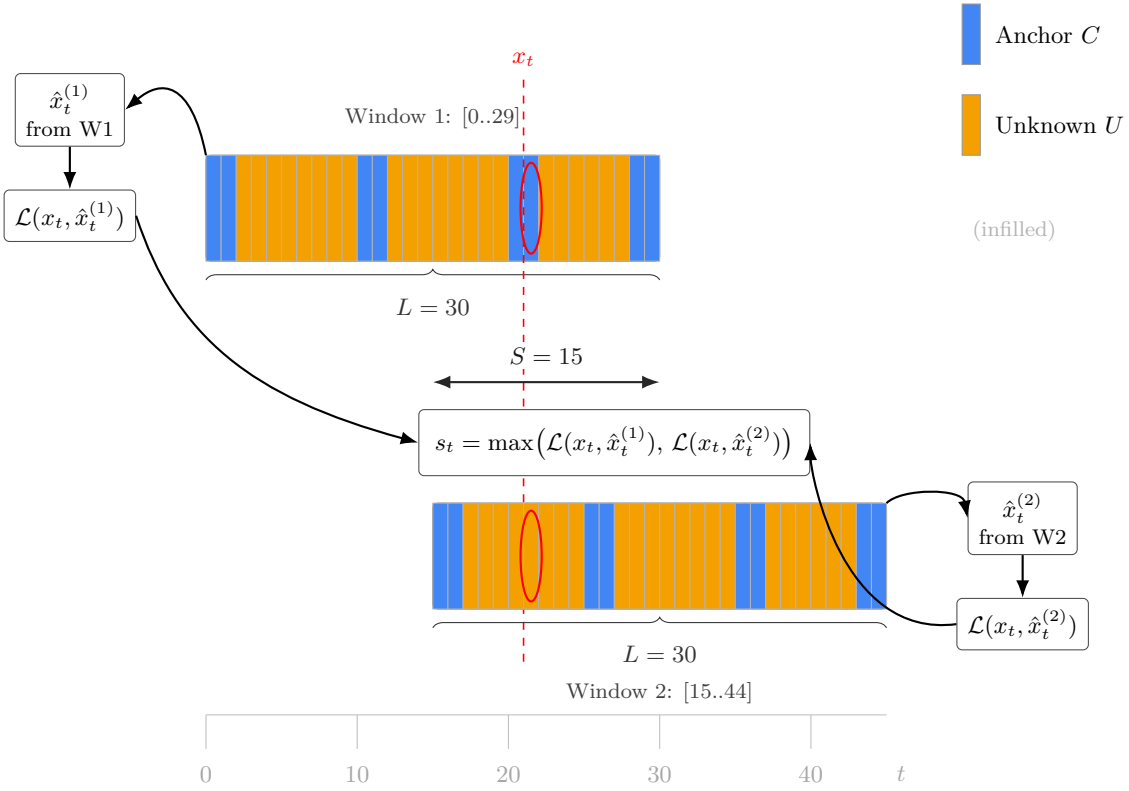

Figure 4.3. Windows adiacenti: L=30 per finestra, S=15 tra gli inizi. x_t (linea rossa tratteggiata) è *anchor* in W1 e *masked* in W2. Il punteggio è $s_t = \max(\mathcal{L}(x_t, \hat{x}_t^{(1)}), \mathcal{L}(x_t, \hat{x}_t^{(2)}))$.

This max-aggregation mechanism is crucial for anomaly detection. When x_t belongs to the anchor set C in one window, its reconstruction error in that context is trivially low, since the ground-truth frame was directly observed by the model. However, in the overlapping window where $x_t \in U$, the model must generate it, and the reconstruction loss is informative of whether x_t conforms to the learned distribution of normality. By taking the maximum across the two candidates, we discard uninformative near-zero errors from anchor appearances, while retaining the residuals from generative reconstructions.

The resulting frame-level anomaly sequence provides a robust measure of abnormality over the entire video. As shown in our experiments, this overlapping infilling combined with max aggregation stabilizes the anomaly score curve across different scenes, enhancing both sensitivity to subtle deviations and robustness to trivial low-error reconstructions.

4.5 Custom Blob-Loss Scoring Function

Why standard scores can miss coherent anomalies. Before introducing WMSE, it is important to recall how commonly used frame-wise reconstruction scores operate, and why they may under-emphasize *spatially connected* residuals that characterize anomalous events.

- **MSE (Mean Squared Error).** MSE computes the squared difference between input and reconstructed pixels, averaged across the frame. While simple and widely adopted, MSE treats each pixel independently and equally, so a large number of scattered pixel errors may dominate the score as much as a compact anomalous region. This sensitivity to pixel noise limits its discriminative power in heterogeneous datasets.

- ℓ_1 **loss.** The ℓ_1 norm measures the average absolute difference across pixels. Compared to MSE, it is more robust to outliers and less sensitive to very large single-pixel deviations. However, like MSE, it ignores spatial relationships, failing to distinguish isolated pixel errors from spatially coherent anomaly patterns.

- **SSIM (Structural Similarity Index).** SSIM [9] compares local patches of the input and reconstruction across three components: luminance, contrast, and structural similarity. It correlates more closely with perceptual quality than MSE or ℓ_1, since it emphasizes structural information. However, SSIM is computed locally with sliding windows, and the averaging process can dilute the contribution of small but spatially coherent anomalies (e.g., a person running in Avenue).

- **MS–SSIM (Multi-Scale SSIM).** MS–SSIM [10] extends SSIM by computing similarity at multiple resolutions and combining them in a weighted fashion. This increases robustness to scale variations and captures both global and local structures. Yet, the pooling across scales may further downweight subtle, fine-grained deviations that are spatially compact but critical for anomaly detection.

- **PSNR (Peak Signal-to-Noise Ratio).** PSNR [11] is derived directly from MSE and expresses error as a logarithmic ratio between the maximum pixel value and reconstruction error. It is commonly used in image/video compression but, its validity as a perceptual quality metric is limited. Since it inherits MSE's lack of spatial selectivity, PSNR does not provide additional benefits for anomaly detection.

- **CW–SSIM (Complex Wavelet SSIM).** CW–SSIM [12] evaluates similarity in the complex wavelet domain, where phase information captures structural alignment. This metric is robust to small geometric distortions such as translations, rotations, and contrast changes. While beneficial for recognition tasks, these invariances can actually hide subtle yet spatially coherent anomalies, making CW–SSIM less effective for detecting small but meaningful deviations in video anomaly detection.

In short, these standard scores do not explicitly *reward spatial connectivity*: they either focus on pixel-wise differences (MSE, ℓ_1, PSNR) or rely on perceptual models (SSIM, MS–SSIM, CW–SSIM) that smooth out fine localized anomalies.

Our approach. To overcome these limitations, we designed a blob-aware loss,that explicitly emphasizes connected residual regions while suppressing isolated noise. This ensures that anomalies manifesting as coherent structures (e.g., a moving vehicle or an abnormal pedestrian trajectory) contribute strongly to the anomaly score, while scattered pixel errors are discounted. As shown in our experiments, Blob-Loss consistently outperforms all the standard alternatives across Avenue, STC, and UBnormal.

Notation. Let $x_t, \hat{x}_t \in [0,1]^{H\times W\times 3}$ be the ground-truth and reconstructed RGB frame at time t, $\Omega = \{1,\dots,H\} \times \{1,\dots,W\}$ the pixel grid.

Step 1: grayscale discrepancy and activation mask. Define the grayscale discrepancy

$$d(p) = \left| \frac{1}{3} \sum_{c\in\{R,G,B\}} x_t(p,c) \; - \; \frac{1}{3} \sum_{c\in\{R,G,B\}} \hat{x}_t(p,c) \right|, \quad p \in \Omega,$$

and the binary activation mask

$$M(p) = \mathbf{1}\{\, d(p) \geq \tau \,\}, \qquad \tau > 0.$$

Step 2: connected components and small-blob rejection. Compute connected components of M to obtain blobs $\{B_k\}_{k=1}^{K}$; discard small blobs with area $|B_k| < a_{\min}$.

Step 3: blob weight map. Assign a scalar weight $w_k > 0$ to each remaining blob and define

$$W(p) = \begin{cases} w_k, & p \in B_k, \\ 0, & \text{otherwise.} \end{cases}$$

We use one of the following rules: (i) $w_k = |B_k|$ (area-proportional), (ii) $w_k = |B_k| \log(1+|B_k|)$ (super-linear), (iii) $w_k = 1$ (uniform on valid blobs). Pixels outside valid blobs receive weight 0.

Step 4: weighted average of per-pixel MSE. Let the per-pixel difference be the RGB mean-squared error

$$\delta^2(p) = \frac{1}{3} \sum_{c\in\{R,G,B\}} \big(x_t(p,c) - \hat{x}_t(p,c)\big)^2.$$

The frame score is the *weighted average* of this per-pixel MSE using the weight map W:

$$\mathrm{WMSE}(x_t, \hat{x}_t) = \frac{\displaystyle\sum_{p\in\Omega} W(p)\, \delta^2(p)}{\displaystyle\sum_{p\in\Omega} W(p)}.$$

If no valid blob exists (i.e., $\sum_p W(p) = 0$), the score is defined to be 0.

Hyperparameters. The activation threshold τ, the minimum blob area $a_{\min}$, and the weight rule w_k jointly control sensitivity to weak signals, robustness to tiny speckles, and emphasis on extended anomalies. In practice, area-based or super-linear weights strengthen truly coherent residuals, while isolated pixels are suppressed. Empirically, WMSE produces smoother anomaly curves and competitive metrics compared with MSE, ℓ_1, SSIM, MS–SSIM, PSNR, and CW–SSIM in our test suites.

4.6 Training and Inference Protocols

Data regime (one-class). Training is carried out in the one-class setting, where only videos depicting *normal* behavior are used. Each training window has a fixed length of L=30 frames, and a deterministic, task-specific masking pattern is applied. In every window, the anchor indices C are kept fixed, while the loss is computed exclusively on the complementary set $U = \{0, \ldots, L-1\} \setminus C$. This strategy ensures that the model is consistently exposed to the same infilling regime that will be encountered at test time, rather than wasting capacity on arbitrary or less relevant masking patterns.

Pre-processing. Before training, all frames are resized with bilinear interpolation to a resolution of 128×128, followed by normalization and batching. The same procedure is applied during inference, ensuring consistency between training and evaluation while also reducing both memory usage and computational overhead.

Masking pattern (selective vs. random). In contrast to RaMViD, which selects anchor frames C at random during training, we adopt a *selective and fixed* masking pattern. Specifically, the first two frames, every 10th and 11th frame, and the final two frames are designated as anchors, while all remaining frames are masked. This deterministic choice leads to sharper and more stable reconstructions, and it also allows the model to converge with significantly fewer diffusion steps at inference time.

Objective (masked diffusion loss). Let x_t denote a clean training window, and $x_t^U \oplus x_0^C$ its corrupted version, where frames in U are noised while those in C remain clean. The optimization objective follows the standard RaMViD denoising loss. The essential point is that gradients are accumulated only over the masked frames, while conditioning always relies on the unmasked anchors.

Optimization schedule. Training proceeds for up to roughly one million steps, depending on the dataset. For instance, we train for approximately 448k steps on Avenue, 978k on STC, and 1,028k on UBnormal. To provide a fair comparison across datasets of different scales, results are often reported in terms of *steps-per-scene* and *steps-per-frame*.

Inference configuration and post-processing. During inference, we employ a diffusion sampler with N_{diff}=200 steps per clip. After the diffusion process the raw anomaly scores are then smoothed with a Gaussian kernel and, if necessary, normalized before applying a threshold. Importantly, the selective masking pattern enables reducing the number of sampling steps from 1000 to 200 without any significant drop in performance.

Evaluation metrics. We evaluate the model using several complementary metrics. We report **frame-level accuracy**, which measures detection quality at the individual frame level. Second, we compute **average accuracy per video**, defined as

$$\text{AvgAcc}_{\text{video}} = \frac{1}{|V|} \sum_{v \in V} Acc(v),$$

where $Acc(v)$ is the frame-level accuracy for video v and V is the set of test videos.

Chapter 5

Experimental Setup

5.1 Datasets

We evaluate our method on three widely used benchmarks for video anomaly detection:

- **CUHK Avenue** [13]: a single-scene dataset containing 16 training videos with only normal events and 21 testing videos with both normal and abnormal events. Abnormal behaviors include running, throwing objects, or loitering. Its simplicity makes it a standard entry point for VAD research.
- **ShanghaiTech Campus (STC)** [14]: a multi-scene dataset consisting of 330 training and 107 testing videos across 13 scenes. Abnormal events include biking in pedestrian areas, chasing, or sudden motion. The presence of diverse environments increases scene variance, making this dataset particularly challenging.
- **UBnormal** [15]: a recent benchmark explicitly designed for open-set anomaly detection. It includes 22 object categories, 29 scenes, and highly heterogeneous anomalies as shown in figure 5.1. Its diversity introduces strong generalization challenges, especially under unsupervised settings.

5.2 Evaluation Metrics

To assess model performance, we employ a set of complementary evaluation metrics that capture both frame-level accuracy and dataset-level consistency:

- **Frame-level Accuracy**. This measures the proportion of test frames that are correctly classified as normal or anomalous:

$$\mathrm{Acc}_{\mathrm{frame}} = \frac{TP + TN}{TP + TN + FP + FN},$$

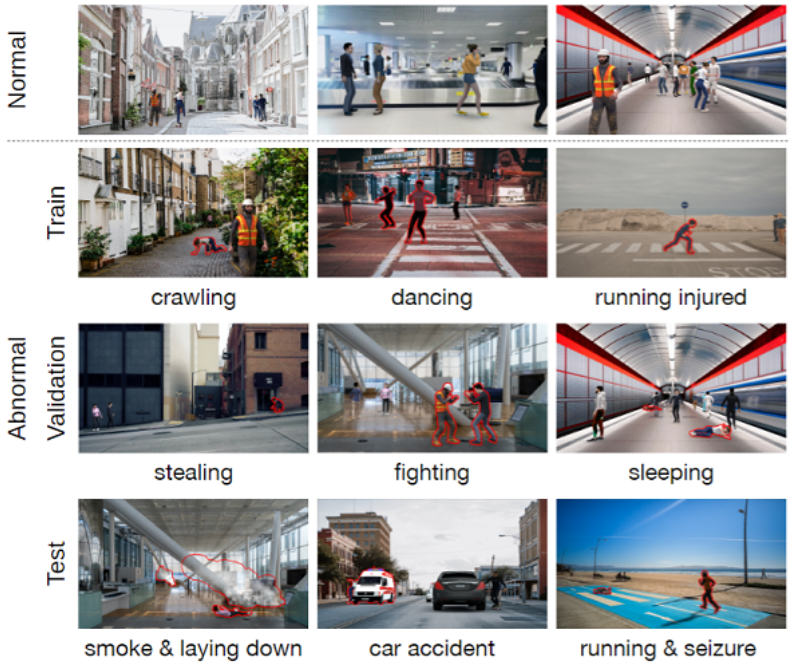

Figure 5.1. Example of actions from UBnormal.

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. Although intuitive, this metric can be biased by class imbalance, since anomalies are typically rare.

- **Average per-video Accuracy**. Instead of pooling all frames, we compute the frame-level accuracy $Acc(v)$ for each video v, and then average across the test set V:

$$\text{AvgAcc}_{\text{video}} = \frac{1}{|V|} \sum_{v \in V} Acc(v).$$

  This reduces bias from videos with disproportionate amounts of anomalies and highlights robustness across different scenes.

- **Best F1-score**. The F1-score is the harmonic mean of precision and recall:

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}},$$

  with

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}.$$

  We report the maximum F1-score achieved across all thresholds, which reflects the best trade-off between correctly detecting anomalies (high recall) and minimizing false alarms (high precision).

Chapter 6

Results

6.1 Quantitative Results

We evaluate our approach on three benchmark datasets: **Avenue**, **ShanghaiTech Campus (STC)**, and **UBnormal**, following their standard training and testing splits. Results are reported in terms of frame-level accuracy and average per-video accuracy, complemented by training statistics such as the number of optimization steps and the ratio of steps per scene in 6.1. This analysis highlights how dataset scale and heterogeneity affect performance, with Avenue benefiting from its single-scene setup, while STC and UBnormal exhibit lower frame-level scores due to higher variability and fewer training steps per scene.

On **Avenue**, the model achieves 85% frame-level accuracy and 81% average accuracy per video using Blob-Loss scoring at 448k steps. The very high steps-per-scene ratio ($448k$ for one scene) leads to rapid convergence and stable performance. In Figure 6.1, we illustrate the anomaly score trajectory for Avenue: the green threshold line successfully separates most of the red ground-truth anomalous segments, confirming that the numerical results correspond to meaningful separations in the anomaly score space.

To further validate robustness, we evaluated the model on Avenue using alternative scoring metrics and regularization levels. Results (Table 6.2) indicate that performance remains consistently strong across metrics, with accuracy values ranging from 79% to 83%. In particular, MSE with σ=6 achieves 85%, while perceptual metrics such as CW-SSIM and L1 yield accuracies of 81%. This confirms that our method generalizes well across distinct evaluation criteria, not being overly sensitive to a single scoring function.

On **STC**, performance reaches 60% frame accuracy and 76% video-level accuracy at 978k steps. Despite the larger dataset, the lower ratio of steps per scene ($\sim$75k) introduces cross-scene variance that depresses frame-level accuracy.

Finally, on **UBnormal**, the model obtains 54% frame accuracy and 87% average per video. Here, the dataset's heterogeneity and a low steps-per-scene ratio ($\sim$35k) hinder frame-level detection but still allow competitive video-level accuracy.

These findings suggest a clear correlation between training steps per scene and achieved accuracy, and the Avenue score plot further demonstrates how anomaly scoring can provide interpretable results in practice.

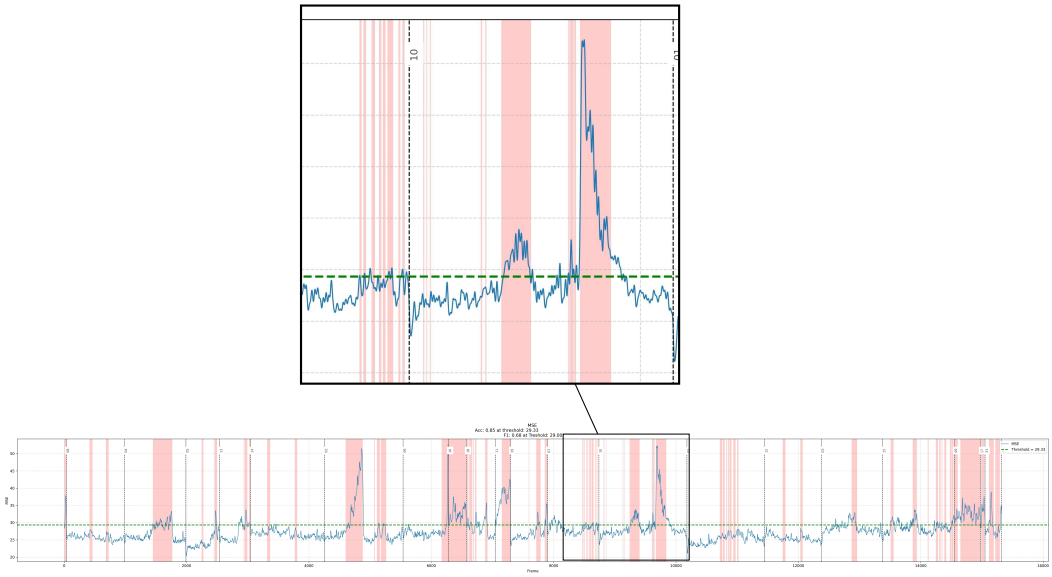

Figure 6.1. Anomaly score curve on Avenue with zoomed-in detail (WMSE).

Table 6.1. Summary of results across datasets. "Steps/Scene" is total training steps divided by the number of scenes.

Dataset	Scenes	Steps (k)	Steps/Scene (k)	Score	Frame Acc.	Avg/Video
Avenue	1	448	448.0	WMSE (σ=4, no norm)	85%	81%
STC	13	978	75.2	WMSE (σ=6, no norm)	60%	76%
UBnormal	29	1028	35.4	MSE (σ=6, norm)	54%	87%

On Avenue, the model thus achieves 85% accuracy with Blod-Loss but maintains stable performance across metrics (79–85%), indicating that the detected anomalies are consistent under both pixel-wise and perceptual similarity measures. On STC, performance decreases to 60% frame accuracy and 76% per-video, primarily due to scene variance and fewer training steps per scene. On UBnormal, the model reaches 54% frame accuracy but 87% per-video, demonstrating that while individual frame decisions are noisy, video-level aggregation remains strong.

6.2 Qualitative Analysis

A key strength of our approach lies in the interpretability of reconstructed frames, which provide intuitive visual evidence of anomalies. Figure 6.2 shows qualitative results from the ShanghaiTech Campus (STC) dataset. Panel (a) presents an original anomalous frame containing a van on the walkway. When this frame is reconstructed by the model (panel (b)), the anomalous object fails to be reproduced correctly: the van is replaced by blurred and incoherent artifacts. This discrepancy between the input and the reconstruction is precisely what drives the anomaly score upward. In contrast, panel (c) shows the reconstruction of a normal frame, which

Table 6.2. Alternative scoring metrics on Avenue. Results show consistent accuracy across diverse measures, confirming robustness of anomaly detection.

Metric	Config.	Frame Acc.
Blob-Loss	σ=4	85%
MSE	σ=6	83%
L1	σ=6	81%
CW-SSIM	σ=6	81%
PSNR	σ=6	80%
MS-SSIM	σ=6	80%
SSIM	σ=6	79%

(a) Original frame X **(b)** Reconstruction of anomaly frame X **(c)** Reconstruction of normal frame Y

Figure 6.2. Frame reconstructions from STC. Anomalous objects such as vehicles are not faithfully reproduced, resulting in visible artifacts that raise anomaly scores, whereas normal frames are reconstructed with high fidelity.

remains faithful to the input scene with clear spatial details and no spurious artifacts. These examples highlight the fundamental principle of reconstruction-based VAD: anomalies correspond to patterns unseen during training, which the model struggles to reproduce, thereby generating detectable residuals.

Complementary insights arise from the UBnormal dataset, where the model faces higher variability and limited steps per scene. Figure 6.3 depicts a case where the reconstruction introduces spurious blurred regions, highlighted by red circles. These artifacts are not linked to real anomalies but to dataset complexity and insufficient per-scene training, which limit the model's generative fidelity. Such cases emphasize that while anomaly detection benefits from visible reconstruction gaps, excessive artifacts can hinder interpretability and inflate false positives.

Finally, in Avenue (Figure 6.1), anomalies such as running or object throwing correspond to distinct peaks in the anomaly score curve. The Blob-Loss scoring function produces sharper separations between normal and abnormal segments compared to traditional MSE or SSIM.

Together, these qualitative observations strengthen the quantitative results: on simple datasets such as Avenue or single-scene settings, anomalies emerge clearly both visually and numerically, while in complex multi-scene datasets like UBnormal, reconstruction artifacts may confound detection and call for additional regularization or increased training per scene.

Figure 6.3. Image from UBnormal. Red circles mark artifacts introduced during reconstruction, reflecting challenges of dataset heterogeneity and lower steps-per-scene ratio.

6.3 Comparison with State-of-the-Art

We compare our method against representative state-of-the-art approaches in video anomaly detection, spanning diffusion-based, reconstruction-based, and memory-augmented prediction models. Results in Table 6.3 summarize frame-level AUC across three benchmark datasets: **Avenue**, **ShanghaiTech Campus (STC)**, and **UBnormal**.

Diffusion-based approaches. Zhou et al. [4] introduce the Motion and Appearance Guided Patch Diffusion Model (MA-PDM), which operates at the patch level and jointly models RGB appearance and motion differences. This method achieves strong performance across datasets, reaching 91.3% on Avenue and 79.2% on STC.

Reconstruction with attention. Wang et al. [5] propose an encoder–decoder with attention, explicitly reconstructing object-level regions of interest through two parallel branches (appearance and optical flow). This design yields competitive accuracy, e.g., 90.3% on Avenue and 77.8% on STC.

Prediction + memory-based methods. Zhao et al. [6] combine prediction with prototype memory encoding (LGN-Net), which allows reconstruction conditioned on learned normality prototypes. This approach reports 89.3% on Avenue and 73.0% on STC.

Ours (Diffusion with Mamba). Our method achieves 85% on Avenue, 60% on STC, and 54% on UBnormal. While Avenue performance is competitive, the gap on multi-scene datasets is attributable to the reduced steps-per-scene ratio (75k for STC and 35k for UBnormal vs. 620k for Avenue). In practice, this limits generalization across diverse scenes, yet confirms the robustness of our model in single-scene scenarios.

These comparisons reveal two main insights: (i) our framework is competitive on Avenue, confirming its ability to model anomalies when training data is concentrated

Table 6.3. Frame-level AUC (%) on CUHK *Avenue*, ShanghaiTech *STC*, and *UBnormal*. "–" = not reported in the cited work.

Method	Type	Avenue	STC	UBnormal
MA-PDM (Zhou *et al.* [4])	Diffusion (patch, motion+appearance)	**91.3**	**79.2**	**63.4**
Making reconstruction great again [5]	Reconstruction with attention	90.3	77.8	–
LGN-Net (Zhao *et al.* [6])	Prediction + Memory (local+global)	89.3	73.0	–
Ours (Mamba Video Diffusion)	Diffusion with Mamba	85.0	60.0	54.0

in a single-scene setup with a high steps-per-scene ratio; (ii) performance on STC and UBnormal highlights the trade-off between dataset heterogeneity and training budget. Unlike attention-based or patch-level diffusion models, our method scales linearly in context length, enabling longer temporal windows, but requires additional training resources to close the performance gap on complex datasets. This suggests that with increased computational budget (i.e., more steps per scene), our method could match or surpass existing state-of-the-art performance while retaining its efficiency and scalability advantages.

Chapter 7

Conclusions

7.1 Summary of Contributions

This thesis addressed the problem of video anomaly detection (VAD) through the development of a diffusion-based generative model augmented with state-space sequence modeling. The key contributions can be summarized as follows:

- We adapted and optimized the Random-Mask Video Diffusion (RaMViD) framework to the anomaly detection setting.
- We introduced a fixed selective masking strategy, replacing the random masking of RaMViD with a deterministic scheme aligned with the inference setup. This modification reduced artifacts, improved the consistency between training and testing, and allowed us to lower the diffusion steps from 1000 to 200 without sacrificing accuracy, yielding a $\sim 5\times$ speed-up at inference.
- We replaced computationally expensive attention mechanisms with Mamba state-space blocks. This substitution enabled linear-time complexity with respect to sequence length while preserving spatial inductive biases through 3D convolutions. As a result, our model can handle longer temporal contexts with improved efficiency.
- We introduced a custom anomaly scoring function, Weighted MSE (WMSE), which emphasizes coherent spatial differences and reduces sensitivity to noisy pixel-level deviations. This function consistently stabilized detection across heterogeneous datasets.
- We conducted comprehensive experiments on three standard benchmarks—CUHK Avenue, ShanghaiTech Campus (STC), and UBnormal—reporting frame-level and video-level accuracies. Our results highlight the importance of training steps per scene and demonstrate that state-space diffusion models are a promising direction for scalable VAD.

7.2 Discussion and Implications

The experimental results presented in this thesis provide a set of insights into the behavior of diffusion models with state-space components for video anomaly detection, while also revealing the main limitations that shaped their performance.

A first key observation concerns the **ratio of training steps per scene**, which emerges as the primary performance driver. On Avenue, where the entire training budget is concentrated on a single scene (448k steps, i.e., 28.9 steps per frame), the model achieves rapid convergence and stable performance. In contrast, STC and UBnormal distribute a higher number of steps across many more scenes (13 and 29, respectively), resulting in substantially fewer steps per scene (75k for STC and 35k for UBnormal). This imbalance depresses frame-level accuracy, even though video-level aggregation remains competitive. The correlation between steps per scene and accuracy, consistently reported across experiments, suggests that training allocation must be carefully scaled with dataset heterogeneity.

The second finding relates to **dataset structure and heterogeneity**. Avenue, being homogeneous and single-scene, enables sharp anomaly separation and smooth score curves. UBnormal, by contrast, introduces highly diverse environments and activities, which makes reconstruction more challenging. This variability reduces frame-level accuracy but simultaneously improves per-video robustness, as aggregation over longer clips compensates for local errors.

A third important contribution is the **effectiveness of the custom anomaly scoring function (Blob-Loss)**. Standard metrics such as MSE, SSIM, or PSNR often produce noisy signals dominated by pixel-level variations. Blob-Loss instead emphasizes spatially coherent residuals, filtering out isolated pixel differences and highlighting anomaly-related structures. Across all benchmarks, this tailored scoring proved more stable, enabling sharper anomaly localization and reducing false positives.

Finally, the study highlights the role of **resource constraints**. Training runs were capped at one million steps due to computational limits, particularly on STC and UBnormal. This prevented full convergence and constrained results by budget rather than by model capacity. With extended training budgets, performance on large and heterogeneous datasets would likely improve substantially, narrowing the current gap with state-of-the-art approaches.

In summary, the results validate the proposed framework as a scalable and efficient solution for video anomaly detection. While its strengths are evident on homogeneous datasets and its design offers linear-time scalability through state-space modeling, performance on complex, multi-scene benchmarks remains limited by training allocation and computational resources.

Bibliography

[1] T. Höppe, A. Mehrjou, S. Bauer, D. Nielsen, and A. Dittadi, "Diffusion models for video prediction and infilling," *Transactions on Machine Learning Research*, 2022.

[2] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," 2022.

[3] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[4] H. Zhou, J. Cai, Y. Ye, Y. Feng, C. Gao, J. Yu, Z. Song, and W. Yang, "Video anomaly detection with motion and appearance guided patch diffusion model," *arXiv preprint arXiv:2412.09026*, 2024. Accepted to AAAI 2025.

[5] Y. Wang, C. Qin, Y. Bai, Y. Xu, X. Ma, and Y. Fu, "Making reconstruction-based method great again for video anomaly detection," in *2022 IEEE International Conference on Data Mining (ICDM)*, pp. 1215–1220, 2022.

[6] M. Zhao, X. Zeng, Y. Liu, J. Liu, D. Li, X. Hu, and C. Pang, "Lgn-net: Local-global normality network for video anomaly detection," *arXiv preprint arXiv:2211.07454*, 2022.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[8] K. Li, X. Li, Y. Wang, Y. He, Y. Wang, L. Wang, and Y. Qiao, "Videomamba: State space model for efficient video understanding," 2024.

[9] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[10] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, vol. 2, pp. 1398–1402 Vol.2, 2003.

[11] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of psnr in image/video quality assessment," *Electronics Letters*, vol. 44, no. 13, pp. 800–801, 2008.

[12] Y. Gao, A. Rehman, and Z. Wang, "Cw-ssim based image classification," in *2011 18th IEEE International Conference on Image Processing*, pp. 1249–1252, 2011.

[13] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in MATLAB," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2720–2727, 2013.

[14] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection – a new baseline," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6536–6545, June 2018.

[15] A. Acsintoae, A. Florescu, M. Georgescu, T. Mare, P. Sumedrea, R. T. Ionescu, F. S. Khan, and M. Shah, "Ubnormal: New benchmark for supervised open-set video anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.