STEP: Score-Based Temporal Energy for Human Pose Video Anomaly Detection

Jakub Micorek[1], Mateusz Koziński[2], and Horst Possegger[1]

[1] Institute of Visual Computing, Graz University of Technology, Austria
{jakub.micorek,possegger}@tugraz.at
[2] Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria

Abstract. Skeleton-based Video Anomaly Detection (VAD) offers a robust, privacy-preserving solution for identifying abnormal behaviors. To model the distribution of normal static and moving poses, recent methods train Energy-Based Models (EBMs) via Denoising Score Matching (DSM). However, directly injecting noise, required for training, into raw joint coordinates creates physically impossible poses, and this structural collapse severely worsens as the temporal window expands. To address this, we introduce STEP, a simple framework that utilizes Principal Component Analysis (PCA) to project pose sequences into a compact, whitened PC-space. Learning the data density within this well-behaved PC-space ensures that the injected noise translates into physically plausible variations, which allows the model to process longer video sequences without the performance collapse of raw coordinate baselines. Additionally, to mitigate inherent pose estimation inaccuracies arising from occlusions or motion blur, we integrate a sequence-level weighting mechanism based on the estimator's confidence scores. Operating at real-time computational efficiency, our simple and lightweight framework outperforms the previous skeleton-based state-of-the-art by 12.2% (90.1% AUROC) on the challenging UBnormal dataset and achieves highly competitive results by improving on the ShanghaiTech benchmark.

Keywords: Score-based Video Anomaly Detection · Human Pose · PCA

1 Introduction

Detecting anomalies in human motion sequences is pivotal for enabling a timely response to emergencies: accidents in production facilities, patient falls in elderly care institutions, or acts of violence in public spaces. However, algorithmic detection of anomalous motion sequences remains an open problem. Its inherent challenge stems from the lack of anomalous training data: only normal motion sequences are available during training. This prohibits framing the problem as event detection or motion sequence classification.

The classical anomaly detection approach relies on training a deep model in self-supervised tasks, like auto-encoding motion sequences [7–9, 12] or predicting masked data fragments [7, 16, 22]. Training data is anomaly-free, but at

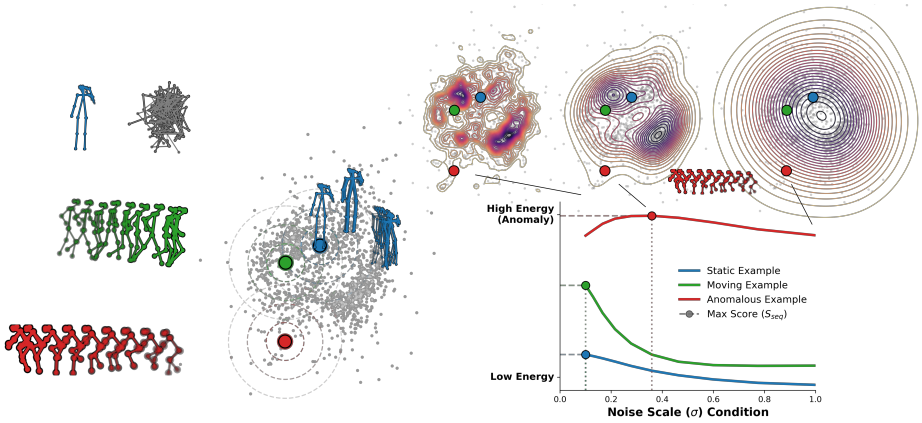

Fig. 1: The STEP Pipeline. Left: Human poses are extracted from the raw video. We highlight a static person (blue), a walking person (green), and an anomalous cyclist (red). Naively adding independent Gaussian noise to pixel coordinates on the static person destroys the kinematic structure, creating physically impossible poses (gray). **Middle:** In contrast, by projecting the sequences into a compact, whitened PC-space, the injected isotropic noise translates into semantically meaningful action variations (*e.g.*, smoothly inducing motion in the static blue pose sequence) while yielding structurally plausible bone lengths and human proportions. **Right:** Consequently, the Energy-Based Model learns smoothed, well-behaved energy landscapes at different σ scales (top). During inference, we evaluate the sequence across distinct noise scales (σ_{low} to σ_{high}). Normal actions (static and walking) fall into lower-energy basins across these scales, while the anomalous cyclist yields a high-energy score.

test time, the model performs the task on possibly anomalous data items, and a task-specific error measure is used as an anomaly indicator. Unfortunately, this method is limited by an inherent flaw: it is predicated on the assumption that a model trained on normal data fails for anomalous input. In reality, deep networks generalize beyond their training set, and there is no guarantee that all anomalies deteriorate the network's performance.

Recently, a less heuristic approach has received increasing attention: both training and non-anomalous test data are treated as instantiations of a random variable with a fixed probability distribution, and anomaly detection is posed as estimating the energy of each test item [20]. To that end, an energy-based model is trained on anomaly-free training data and used to predict the energy of the possibly anomalous test items. Test samples with high estimated energy, in other words, ones unlikely under the model of the non-anomalous data distribution, are deemed anomalous. Samples with low estimated energy are considered normal. The advantage of this approach is that it casts the practical challenge of accurate anomaly detection into a mathematically well-defined objective of developing accurate approximations to the energy function.

Unfortunately, training energy-based anomaly detectors remains challenging. A typical approach is to harness the Denoising Score Matching (DSM) [27] objec-

tive, which relies on injecting noise into training samples. However, naively applying DSM to human pose sequences, with their complex pattern of joint dependencies, displays two fundamental deficiencies. First, as illustrated in the left part of Fig. 1, injecting isotropic Gaussian noise into joint coordinates destroys the kinematic structure of the human body: bone lengths warp, symmetries break, and the sequence degrades into a cloud of independent joints. Crucially, this structural collapse compounds as the temporal window T expands. This drastically inflates the space of physically impossible poses, thereby severely degrading the model's ability to learn a meaningful energy landscape. Consequently, the expressive power of the energy model is wasted on modeling impossible body configurations. Second, existing approaches treat pose sequences extracted from images as perfect ground truth. In reality, off-the-shelf pose estimators applied to real-world surveillance footage suffer from coordinate jitter, missing joints, and occlusion, which causes current models to warp energy landscapes around tracking failures.

To address these deficiencies, we propose Score-based Temporal Energy for Poses (STEP), a novel energy-based approach to anomaly detection in human motion sequences. STEP is trained with Denoising Score Matching (DSM). However, in contrast to previous approaches leveraging DSM, STEP foregoes modeling the data distribution in the original space of joint coordinate sequences in favor of a compact space resulting from Principal Component Analysis (PCA). As shown in the middle of Fig. 1, injecting isotropic noise into this representation perturbs human motion sequences along learned basis trajectories, generating semantically meaningful action variations while preserving natural human proportions. As a result, the expressive power of the distribution model is spent on representing feasible motion sequences, as opposed to kinematically impossible noise. Consequently, as depicted in the right part of Fig. 1, our Energy-Based Model learns a smoothed, well-behaved energy landscape where normal actions naturally fall into low-energy basins.

Furthermore, to account for the inherent unreliability of pose estimators, we introduce a soft confidence weighting mechanism into the DSM training objective and test-time energy estimation. This discounts the impact of tracking failures without perturbing the temporal data structure or provoking the omission of critical anomalies, yielding a robust and accurate anomaly detector.

The technical contributions behind STEP are summarized as follows:

- **PC-Space Density Estimation:** We demonstrate that naively applying Denoising Score Matching to raw coordinates yields suboptimal performance and worsens as the temporal dimension T grows. We resolve this by operating in a compact, whitened PC-space, ensuring noise injection translates to meaningful pose sequence variations.
- **Confidence-Aware Anomaly Scoring:** We introduce a sequence-level confidence weighting mechanism into the DSM framework, both at training and test time. This accounts for low-confidence poses and provides a significant boost to overall AUROC.

– **Architectural Conditioning:** We show that introducing a σ-modulated Residual MLP, which utilizes skip-connections and per-layer noise conditioning, consistently outperforms the architectures used in previous work.

In exhaustive experiments on two challenging human motion datasets, UBnormal and ShanghaiTech, the STEP framework outperforms the previous skeleton-based state-of-the-art by 12.2% and 0.3% AUROC, respectively. Crucially, these performance gains require negligible computational overhead. Our lightweight architecture processes dense crowds in under 1 ms, leaving overall pipeline latency dominated solely by the upstream pose extraction.

2 Related Work

Anomaly detection in human motion sequences has become a prominent focus, as it abstracts away the background clutter and illumination changes that typically challenge traditional pixel-space approaches. Historically, one way to detect anomalies in human motion is directly from raw videos. This task, often framed as Video Anomaly Detection (VAD) under one-class classification (OCC), relies on training exclusively on normal data [1, 26, 30, 32]. These standard approaches typically categorize into frame-centric models analyzing global features [29,30,32] and object-centric models estimating bounding box abnormalities [2, 5, 7, 12, 28]. Such methods often solve proxy tasks like auto-encoding or future-frame prediction [9,16,17,23,31]. However, multimodal and appearance-based modeling from videos remains vulnerable to demographic bias, privacy concerns, and scene-specific artifacts, motivating a strong shift toward purely skeletal kinematics.

Early skeleton-based methods framed motion anomaly detection primarily as an unsupervised reconstruction or prediction problem [13–15,19,21,24]. However, standard deep networks tend to over-generalize, successfully reconstructing even unseen anomalous behaviors. To better capture the multimodal diversity of human motion, contemporary models have increasingly adopted explicit probabilistic density estimation. For instance, while MoCoDAD [5] utilizes diffusion models to forecast future poses, it ultimately falls back on measuring coordinate-level reconstruction errors to score anomalies rather than evaluating the true probability of the sequence. Conversely, frameworks that do compute true probabilities rely on highly specialized architectures to model skeletal dependencies before applying density estimation. For instance, alongside earlier spatial networks like Normal Graph [18] and COSKAD [6], STG-NF [10] couples Spatio-Temporal Graph Convolutional Networks (ST-GCNs) with Normalizing Flows, whereas SeeKer [3] pairs autoregressive keypoint factorization with explicit per-keypoint likelihoods. Alternatively, EBMs trained via DSM provide a highly adaptable approach to density estimation [11, 25, 27]. As demonstrated by MULDE [20], this feature-agnostic technique can be applied directly to multimodal video features and individual poses. However, as discussed, naively applying this framework to raw coordinates is fundamentally flawed due to noise-induced structural collapse. Instead, STEP ensures robust density estimation by operating in a structure-preserving, whitened PC space.

3 Method

Problem formulation We are given a training set D of human motion sequences extracted by an off-the-shelf pose tracker from a video recording. Each sequence $X \in D$ takes the form of a tensor $X \in \mathbb{R}^{2\times V\times T}$, where the first dimension corresponds to the two image coordinates, V is the number of joints, and T is the number of frames included in the sequence. Our goal is to train an anomaly detector to classify test sequences of the same form as either normal or anomalous.

Overview of our approach We assume that both training samples and non-anomalous test samples originate from the same data distribution and train a deep network f to estimate sample energy. At test time, we use f as an anomaly indicator. Our training procedure relies on denoising score matching (DSM), which we introduce in Sec. 3.1. However, since a direct application of DSM to human motion sequences results in subpar performance, we propose modifications: As described in Sec. 3.2, we model the distribution of the motion sequences in the space of principal components, as opposed to the space of raw joint coordinates; We introduce pose confidence weighting into training and prediction, as outlined in Sec. 3.3, and verify a new architecture of the energy-based model, described in Sec. 3.4.

3.1 Denoising score matching for anomaly detection

Our work builds on denoising score matching for anomaly detection [20]. Denoising score matching [27] is a formulation for learning the energy gradient of a distribution given in the form of a data sample, and [20] harnessed this approach for learning the energy function itself. Below, we first introduce DSM and then describe its use for anomaly detection in previous work.

Given a training set D of data items $\mathbf{x}$ drawn from a fixed distribution with probability density p, DSM lets one train a deep network to approximate the energy gradient of the training distribution injected with noise. Formally, the noise-injected distribution is defined as

$$q(\tilde{\mathbf{x}}) = \int \rho(\tilde{\mathbf{x}}|\mathbf{x})p(\mathbf{x})d\mathbf{x}, \tag{1}$$

where $\rho(\tilde{\mathbf{x}}|\mathbf{x})$ denotes the conditional distribution of a noisy sample $\tilde{\mathbf{x}}$ given a noise-free sample $\mathbf{x}$, universally taken to be an i.i.d. Gaussian centered at $\mathbf{x}$. The main idea behind DSM is that q preserves the shape of p but leads to a simpler and more effective training formulation. The energy of a sample $\tilde{\mathbf{x}}$ is defined as $E_q(\tilde{\mathbf{x}}) = -\log q(\tilde{\mathbf{x}})$. DSM trains a neural network s_θ, parameterized with a vector θ, to approximate the energy gradient $\nabla_{\tilde{\mathbf{x}}} E_q(\tilde{\mathbf{x}})$ in the sense of solving

$$\min_\theta \mathbb{E}_{\tilde{\mathbf{x}}\sim q(\tilde{\mathbf{x}})} \left\| s_\theta(\tilde{\mathbf{x}}) - \nabla_{\tilde{\mathbf{x}}} E_q(\tilde{\mathbf{x}}) \right\|_2^2 . \tag{2}$$

Directly evaluating (2) is impossible because $q(\tilde{\mathbf{x}})$ is not known analytically, but Vincent et al. [27] showed that it is equivalent, up to an additive constant, to

$$\min_{\theta} \mathbb{E}_{\substack{\mathbf{x}\sim p(\mathbf{x}) \\ \tilde{\mathbf{x}}\sim\mathcal{N}(\tilde{\mathbf{x}}|\mathbf{x},\sigma\mathbf{I})}} \left\| s_\theta(\tilde{\mathbf{x}}) - \frac{\tilde{\mathbf{x}}-\mathbf{x}}{\sigma^2} \right\|_2^2, \tag{3}$$

which can be evaluated efficiently. DSM owes its name to the fact that training s with the objective (3) resembles training the model for denoising.

Detecting anomalies requires approximating the energy $E_q(\tilde{\mathbf{x}})$, as opposed to its gradient. To that end, recent work [20] modified the DSM objective (3) to train the gradient of f, instead of the network itself,

$$\min_{\theta} \mathbb{E}_{\substack{\mathbf{x}\sim p(\mathbf{x}) \\ \tilde{\mathbf{x}}\sim\mathcal{N}(\tilde{\mathbf{x}}|\mathbf{x},\sigma\mathbf{I})}} \left\| \nabla_{\tilde{\mathbf{x}}} f_\theta\left(\tilde{\mathbf{x}}\right) - \frac{\tilde{\mathbf{x}}-\mathbf{x}}{\sigma^2} \right\|_2^2, \tag{4}$$

making $\nabla_{\tilde{\mathbf{x}}} f_\theta\left(\tilde{\mathbf{x}}\right)$ approximate $\nabla_{\tilde{\mathbf{x}}} E_q(\tilde{\mathbf{x}})$, thereby aligning f with $E_q(\tilde{\mathbf{x}})$ up to a constant bias. In practice, instead of approximating $E_q(\tilde{\mathbf{x}})$ for a fixed noise scale σ, the network f_θ is trained to approximate a family of energies $E_{q_\sigma}(\tilde{\mathbf{x}})$, parameterized by σ, with a multi-scale objective

$$\min_{\theta} \mathbb{E}_{\substack{\mathbf{x}\sim p(\mathbf{x}) \\ \tilde{\mathbf{x}}\sim\mathcal{N}(\tilde{\mathbf{x}}|\mathbf{x},\sigma^2 I) \\ \sigma\sim\mathcal{U}(\{\sigma_i\}_{i=1}^L)}} \lambda(\sigma) \left\| \nabla_{\tilde{\mathbf{x}}} f_\theta(\tilde{\mathbf{x}},\sigma) - \frac{\tilde{\mathbf{x}}-\mathbf{x}}{\sigma^2} \right\|_2^2, \tag{5}$$

where $\mathcal{U}(\{\sigma_i\}_{i=1}^L)$ denotes a uniform distribution over L predefined noise scales σ_i, and $\lambda(\sigma) = \sigma^2$ is a scale-specific weight. At test time, energy estimates corresponding to different noise levels are aggregated into a single anomaly indicator.

3.2 Denoising score matching in the space of principal components

The disadvantage of DSM in the context of human motion sequences is that injecting i.i.d. Gaussian noise into sequences of joint coordinates ignores the structure of the human body and destroys the pattern of statistical dependencies between joint positions. The resulting probability distribution assigns mass to kinematically implausible motion patterns, like the one shown in gray in the left part of Fig. 1. This defeats the purpose of training the model to capture the manifold of human motion: Modelling capacity is spent on accommodating implausible poses instead of outlining the limits of normal, kinematically feasible sequences.

We address the problem by moving away from the original space of joint coordinates and modeling the distribution of the motion sequences in a more compact space, where Gaussian noise corresponds to valid poses, as opposed to kinematically implausible ones. We tested three alternative techniques of projecting motion sequences to such a space: Principal Component Analysis, a plain autoencoder and a variational autoencoder. We selected PCA, because projecting motion sequences to the whitened space of principal components (PC) yielded

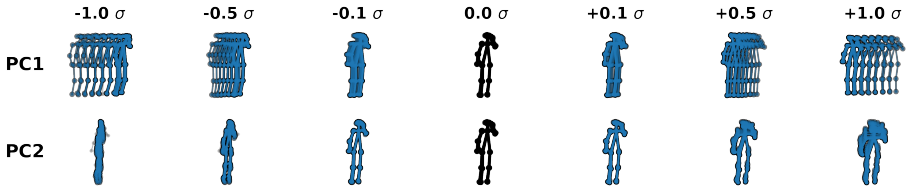

Fig. 2: The principal components correspond to semantically meaningful actions: The image shows the effect of shifting a static pose along the first two principal components. PC1 corresponds to walking across the field of view, while PC2 represents walking towards or away from the camera.

the highest anomaly detection accuracy (see detailed experiments in the supplementary material). To formalize this approach, we denote the projection matrix of the top K principal components, estimated on the training set, by W_K, the diagonal matrix of the corresponding eigenvalues by Λ_K and the training data mean by μ, and formalize the projection to the whitened PC space as:

$$\pi(\mathbf{x}) = \Lambda_K^{-\frac{1}{2}} W_K^T (\mathbf{x} - \mu) \; . \tag{6}$$

The projection effectively scales the latent space into an isotropic hypersphere.

Modelling the distribution of motion sequences in the whitened space of principal components has two advantages. First, injecting noise into this representation results in plausible human motion sequences, like the one shown in blue in the left part of Fig. 1. In Fig. 2, we show that principal components represent meaningful actions, like walking. Second, PCA acts as a denoiser: limiting the representation to top K PCs removes high-frequency motion, letting the energy model focus on the overarching motion trajectory instead of frame-to-frame fluctuation of joint positions. Our experiments in Sec. 4.3 show that moving the model to the whitened space results in a considerable performance boost.

3.3 Accounting for uncertainty in pose extraction

The DSM training objective (4) treats all data items equally. In our context of human motion anomaly detection, this ignores the fact that pose sequences extracted from videos are associated with uncertainty. In consequence, the capacity of the energy model is spent on trying to accommodate erroneous poses.

To ensure the learned energy landscape is more robust against severe tracking and detection failures, we discount the influence of uncertain pose sequences through confidence weighting. We denote an average joint position confidence score provided by the pose estimator for motion sequence $\mathbf{x}$ by $c(\mathbf{x}) \in [0, 1]$. We use $c(\mathbf{x})$ to reweight loss terms corresponding to individual training sequences. To that end, we define a weighted multi-scale DSM loss as

$$\min_{\theta} \mathbb{E}_{\substack{\mathbf{x}\sim p(\mathbf{x}) \\ \mathbf{z}\sim\mathcal{N}(\mathbf{z}|\pi(\mathbf{x}),\sigma^2 I) \\ \sigma\sim\mathcal{U}(\{\sigma_i\}_{i=1}^{L})}} c(\mathbf{x})\lambda(\sigma) \left\| \nabla_{\mathbf{z}} f_\theta(\mathbf{z}, \sigma) - \frac{\mathbf{z} - \pi(\mathbf{x})}{\sigma^2} \right\|_2^2 , \tag{7}$$

where $\mathbf{z}$ denotes a noisy vector in the PC space. $c(\mathbf{x})$ downweights low-confidence measurements, preventing tracking failures from distorting the energy landscape.

At test-time, given a motion sequence $\mathbf{x}$ and a sequence of pre-defined noise levels σ_i, we use the trained energy model f to compute noise-level-specific energy estimates $f(\pi(\mathbf{x}), \sigma_i)$. To aggregate them into a single anomaly estimator, we standardize the energy at the i-th noise level using the mean energy $\bar{E}_i$ and energy variance $\mathrm{Var}(E_i)$, computed over the training set, and then select the largest standardized energy estimate. To accommodate the uncertainty of the pose extractor also at test time, we weight the standardized energy estimates by the confidence estimate c. Formally, the aggregated anomaly indicator is

$$A(\mathbf{x}) = c(\mathbf{x}) \max_i \frac{f(\pi(\mathbf{x}), \sigma_i) - \bar{E}_i}{\sqrt{\mathrm{Var}(E_i)}}. \tag{8}$$

Our ablation study shows that discounting the anomaly score of uncertain motion sequences often prevents a common error consisting of declaring a sequence anomalous due to tracking errors.

3.4 Architecture of the Energy-Based Model

Previous skeleton-based energy models, such as MULDE [20], typically rely on standard MLPs where the noise scale σ is only integrated via early-fusion concatenation at the input. We argue that this leaves the architectural design space for noise-dependent density estimation largely unexplored. Instead, we propose a σ-modulated Residual MLP that treats σ as a global conditioning signal. As shown in Fig. 3, σ is explicitly routed to every hidden layer to modulate activations within each residual block. Our ablation study (Table 3) confirms that this dense conditioning approach consistently outperforms the vanilla MLP baseline.

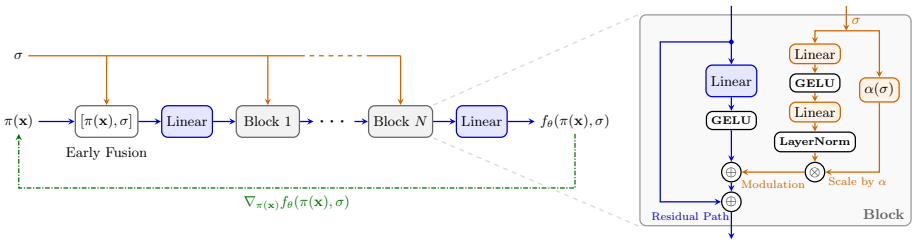

Fig. 3: STEP Architecture. Left: Overview of our energy model. The noise scale σ is integrated via early fusion with the whitened pose sequence $\pi(\mathbf{x})$ and additionally input to every stacked block. The network $f_\theta(\pi(\mathbf{x}), \sigma)$ outputs a scalar energy estimate. The dashed green path illustrates the energy gradient calculation, which is required only during training. **Right:** Internal structure of the modulated block, combining conditioning on σ with a residual skip connection.

4 Experiments

To validate the effectiveness of our approach, we evaluate our STEP framework, comparing its performance against the state-of-the-art, ablating its core architectural components, and analyzing its real-time computational efficiency.

4.1 Experimental Setup

Datasets. We evaluate our method on two popular and challenging Video Anomaly Detection benchmarks: ShanghaiTech [17] and UBnormal [1]. ShanghaiTech is a real-world dataset comprising 330 training and 107 test videos across 13 scenes, featuring anomalous events such as robbery, fighting, and cycling. UBnormal is a synthetic, photorealistic dataset composed of 543 videos (268 training, 64 validation, and 211 test sequences) spanning 29 virtual scenes. It introduces complex human-related anomalies, such as fighting and running, alongside non-human and environmental anomalies, and is highly accurately labeled. Because both datasets contain events that are not explicitly driven by human pose (*e.g.*, fire or weather), we follow established protocols [5, 24] and report results on both the Full test and the Human-Related (HR) splits, which isolate anomalies directly involving human actions.

Baselines. We compare STEP against a broad range of skeleton-based VAD algorithms. While we compare our framework to a wide range of traditional reconstruction and prediction models [5, 6, 13–15, 18, 19, 21, 24], our primary focus is on probabilistic density estimation methods. While STG-NF [10] and SeeKer [3] define the current performance ceiling on ShanghaiTech and UBnormal for skeleton-based VAD, MULDE [20] serves as our most direct methodological baseline. By exposing the structural collapse inherent in MULDE's DSM framework when applied to raw coordinates, we demonstrate how our PC-space and confidence weighting approach effectively resolves these fundamental limitations. A broader comparison against additional multi-modal and pixel-based methods is provided in the supplementary material.

Evaluation Metrics. We use the standard Area Under the Receiver Operating Characteristic curve (AUROC) as our evaluation metric. To ensure a fair comparison with the state-of-the-art, we compute the AUROC jointly over all test frames (micro-averaging) without any video-level normalization. We follow the established evaluation protocol, applying temporal Gaussian 1D-smoothing on the aggregated anomaly scores [3, 10, 20]. Additional results using Average Precision (AP) are provided in the supplementary material.

Implementation Details. To ensure a direct comparison to our closest competitors, we utilize the same extracted poses (via AlphaPose [4]), raw confidences, and per-segment standardizations as the current state-of-the-art models, SeeKer [3] and STG-NF [10]. Notably, this pose estimator introduces an 18-point skeleton. Each sequence is scored based on the current frame and the preceding $T-1$ frames. As illustrated in Figure 3, our EBM is built with four hidden blocks, utilizing a hidden dimension of 1024. To construct the multiscale DSM objective,

we sample uniformly from a discrete geometric sequence of $L = 10$ noise scales bounded between $\sigma_{low} = 0.1$ and $\sigma_{high} = 1.0$, unless stated otherwise.

The network is trained exclusively on normal data for up to 400 epochs with a batch size of 1024 using the AdamW optimizer (weight decay 10^{-2}, $\beta_1 = 0.5$, $\beta_2 = 0.9$). The learning rate follows a Cosine Annealing schedule with a linear warmup during the first epoch. The initial learning rate is set to 5×10^{-4} for UBnormal and 2×10^{-4} for ShanghaiTech, smoothly decaying to a minimum of half the initial rate. Because UBnormal provides a dedicated validation split, we employ early stopping based on validation AUROC; for ShanghaiTech, models are evaluated exactly at the end of the 400-epoch training schedule. Unless stated otherwise, we report AUROC for $T = 12, K = 48$ for both ShanghaiTech and UBnormal, selected based on the best performance on the UBnormal validation set. Finally, we maintain a secondary Exponential Moving Average (EMA) model, initialized as an exact copy of the actively trained model, whose weights are updated after each optimization step using a decay factor of 0.999. By averaging the weight updates over time, the EMA model smooths out the slight performance fluctuations observed in the active model during training, promoting a stable and highly reproducible model state. Consequently, this EMA model is strictly used for all our evaluations. To ensure rigorous evaluation against the state-of-the-art, we evaluate our main STEP framework across 20 independent training runs to report mean and standard deviation metrics.

4.2 State-of-the-Art Comparison

Table 1 compares our proposed STEP framework against existing state-of-the-art skeleton-based methodologies. First, we highlight the performance of our direct methodological baseline, MULDE [20]. Surprisingly, adapting their score-matching framework to raw joint coordinates using only isolated, single-frame poses ($T = 1$, lacking any temporal modeling) yields an AUROC of 80.6% on UBnormal. This already surpasses previous complex temporal models like SeeKer (77.9%). However, as we will show in Sec. 4.3, this raw coordinate approach suffers from structural collapse when temporal windows are expanded. By projecting the poses into our whitened PC-space, STEP drastically resolves these limitations and unlocks the full potential of temporal density estimation.

On the synthetic UBnormal dataset, STEP achieves an average AUROC of 90.1% ($\pm$0.4) across 20 independent training runs on the Full test set, and 90.9% on the Human-Related split. This represents a substantial absolute improvement of over 12% compared to the previous state-of-the-art, setting a new benchmark for the dataset. Furthermore, STEP demonstrates robust generalization on the real-world ShanghaiTech dataset. It achieves a leading average performance of 86.2% ($\pm$0.1) on the Full set and 87.7% on the HR split, on par with STG-NF (85.9% and 87.4%, respectively).

Notably, our pose-only method outperforms several established methods that rely on heavier, multi-modal inputs (such as raw pixels and optical flow). A comprehensive comparison against these multi-modal approaches is provided in the supplementary material. A qualitative illustration of STEP's detection

Table 1: State-of-the-art comparison on ShanghaiTech and UBnormal. STEP sets a new state-of-the-art on UBnormal and matches performance on ShanghaiTech. Results are reported for both the Full test set and the Human-Related (HR) splits. To ensure rigorous evaluation, the STEP results are reported as the mean and standard deviation across 20 independent training runs. Best results per column are in **bold**, second best are underlined. Results show AUROC (%). †reproduced by us.

Method	ShanghaiTech		UBnormal	
	Full	HR	Full	HR
BiPOCO [14]	-	74.9	50.7	52.3
MPED-RNN [21]	73.4	75.4	60.6	61.2
MTP [24]	76.0	77.0	-	-
GEPC [19]	76.1	74.8	53.4	55.2
PoseCVAE [13]	-	75.5	-	-
Normal Graph [18]	-	76.5	-	-
COSKAD [6]	-	77.1	65.0	65.5
GCAE-LSTM [15]	-	77.2	-	-
MoCoDAD [5]	-	77.6	68.3	68.4
MULDE (Pose, $T = 1$) [20]	78.5	-	80.6†	-
STG-NF [10]	85.9	87.4	71.8	71.5
SeeKer [3]	85.5	86.9	77.9	78.9
STEP (Ours)	**86.2** ± 0.1	**87.7** ± 0.1	**90.1** ± 0.4	**90.9** ± 0.4

behavior on ShanghaiTech is provided in Fig. 4. We further evaluate on the MSAD dataset [33] and its Human-Related split (MSAD-HR), first explored for skeleton-based VAD by SeeKer [3]; STEP achieves 74.1% AUROC on MSAD-HR, outperforming SeeKer (61.1%) by 13% and STG-NF (55.7%) by 18%. A detailed study on the MSAD dataset is provided in the supplementary material.

4.3 Ablation Studies

Resolving Structural Temporal Collapse. Our core hypothesis is two-fold: (1) Denoising Score Matching fails on temporal pose sequences because isotropic coordinate noise generates physically impossible poses in high dimensions, and (2) off-the-shelf pose estimators introduce severe tracking artifacts that corrupt the learned energy. To cleanly isolate the relative impact of individual architectural choices, all ablation studies in this section are conducted using a fixed random seed. We note that our selected baseline configuration ($T = 12, K = 48$) achieves 90.4% AUROC on UBnormal under this fixed random seed, tightly aligning with the highly stable multi-run mean (90.1% ± 0.4) reported in Tab. 1.

Table 2 systematically validates our core claims by isolating both mechanisms. When applying the EBM directly to raw coordinate vectors (MULDE baseline), performance peaks at an extremely short temporal window ($T = 2$ at 80.7%) but degrades rapidly, dropping by 10% as the window expands to $T = 32$. Integrating our sequence-level confidence weighting on raw coordinates provides a massive baseline improvement (boosting $T = 2$ to 87.2%), demonstrating that the EBM is highly sensitive to noisy pose estimates. However, confidence weighting alone cannot prevent the fundamental high-dimensional collapse, as performance still sharply declines to 73.1% at $T = 32$.

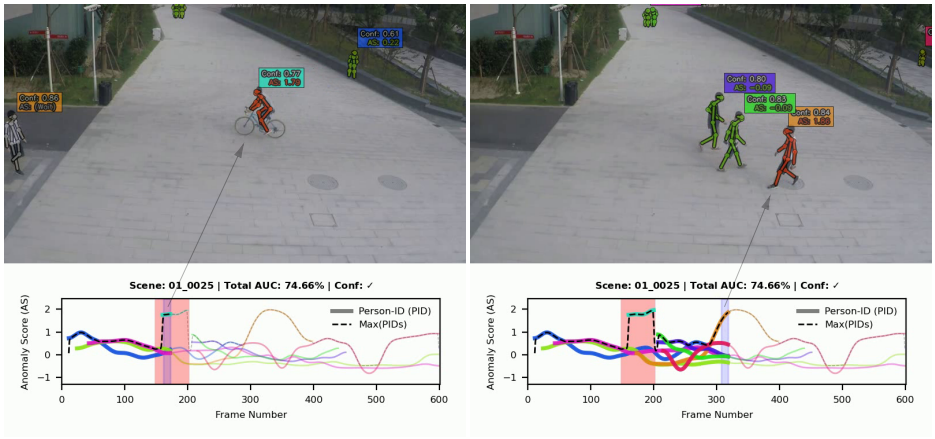

Fig. 4: Qualitative results on ShanghaiTech (clip `01_0025`). Each person's track is assigned a unique color (its text box matches the time-series plot below each frame); skeleton joints are colored from green (low anomaly score) to red (high anomaly score). The number shown above each person is the mean sequence confidence; below it is the confidence-weighted anomaly score. The temporal plot shows all active tracks' scores over time; the red-shaded region marks the ground-truth anomaly interval. **Left,** frames 150–200: STEP correctly detects a passing cyclist (cyan track) as anomalous, producing a sharp anomaly score spike. **Right,** frames 300–370: An insightful false positive: a pedestrian (orange track) turns and walks backwards, receiving an elevated score despite no anomaly label, revealing the model has learned forward-facing walking as normal.

Projecting the sequences onto a compact PC-space without confidence weighting ($K = 48, \times$) acts as a structural anchor. While absolute performance is bottlenecked by unmitigated tracking noise (∼84%), the model becomes more robust to temporal expansion, successfully halting the severe degradation seen in the raw coordinate baselines at high T.

By uniting the PCA approach with confidence weighting (Full STEP), we resolve both deficiencies. The table reveals a clear diagonal shift in performance based on the chosen capacity: overly tight bottlenecks ($K = 16, 32$) perform exceptionally well for short windows ($T = 4, T = 2$) but lose critical information as sequence length increases. Conversely, under-compressing the space ($K = 128$) leaves the EBM susceptible to the dimensionality curse at short windows. While multiple balanced configurations ($K \in \{48, 64\}$) achieve $> 90\%$ AUROC at longer temporal windows, we anchor our final evaluation to the $T = 12, K = 48$ configuration (yielding 90.4% on the test set), as it achieved the highest performance on the UBnormal validation set.

Component Breakdown and Noise Scale Sensitivity. To isolate the impact of our specific design choices, we ablate the core components of the STEP framework in Tab. 3. By reporting the degradation across our unified configuration ($T = 12, K = 48$) on both datasets, we demonstrate that these

Table 2: Ablation of temporal window (T), PCA dimension (K), and confidence weighting on UBnormal. Naively applying DSM to raw coordinates (MULDE) causes structural collapse as T expands. Projecting sequences onto our PC-space stabilizes this degradation, while confidence weighting mitigates tracking noise. The combination reveals a clear capacity trade-off, with the highest-performing configurations shifting diagonally as T increases. Results show AUROC (%). Best results per row are in **bold**, second best are underlined. Missing values (-) denote invalid configurations where the target PCA dimension K exceeds the input dimension. All results are reported for a fixed random seed. †reproduced by us.

Method	PCA (K)	Conf.	Temporal Window (T)							
			1	2	4	8	12	16	24	32
MULDE†	×	×	80.6	**80.7**	78.7	75.7	72.2	72.6	72.6	70.6
STEP	×	✓	86.0	**87.2**	86.3	82.4	81.3	80.5	78.3	73.1
STEP	48	×	-	82.1	83.2	**84.1**	84.0	83.6	82.3	81.3
STEP	16	✓	86.7	87.9	**88.0**	87.3	86.4	85.0	82.9	81.6
STEP	32	✓	86.7	**89.5**	88.4	88.6	88.9	89.0	87.5	85.7
STEP	48	✓	-	88.8	89.6	89.2	**90.4**	90.2	89.5	88.3
STEP	64	✓	-	88.2	88.8	90.0	90.1	**90.5**	90.3	88.8
STEP	128	✓	-	-	87.5	87.9	88.8	**89.2**	**89.2**	88.4

trends are fundamental to the framework. A complete ablation matrix detailing every T and K combination is provided in the supplementary material.

First, we observe that our σ-modulated Residual MLP consistently outperforms the standard unmodulated MLP used in previous EBM frameworks by roughly 1%, suggesting the network requires stronger architectural conditioning to utilize the multi-scale noise. Furthermore, standardizing (whitening) the PC-space is critical; without it, the isotropic noise $\mathcal{N}(0, \sigma^2 I)$ disproportionately corrupts low-variance principal components while under-perturbing the dominant ones, causing performance to drop by 2.2% on UBnormal.

Next, we evaluate the role of pose estimator confidence. A common heuristic, *e.g.* used by SeeKer [3], is to drop poses below a certain confidence threshold (*e.g.*, < 0.4) during training. While applying a highly conservative filter (< 0.1) yields a marginal 0.1% improvement for our model, relying on such thresholds is inherently brittle and introduces another hyperparameter. Aggressive filtering (< 0.4) discards too much valid, albeit noisy, training data, causing a massive 4.5% performance collapse on UBnormal. By dropping these low-confidence poses, the model fails to learn the natural variance of fast or occluded normal motions, creating blind spots in the learned density. Instead, our soft, sequence-level confidence weighting dynamically scales the loss. This allows the model to leverage the full training set while naturally down-weighting tracking failures.

Finally, we evaluate the bounds of the noise scales ($\sigma_{low} = 0.1, \sigma_{high} = 1.0$) employed in the DSM objective. As shown in the table, the model is robust within a reasonable range (*e.g.*, lowering σ_{high} to 0.5 yields comparable results), but performance degrades at the extremes. Setting the minimum noise scale too low ($\sigma_{low} = 10^{-4}$) causes the network to become overly sensitive to the microscopic measurement noise of the pose estimator rather than evaluating the

Table 3: Component Ablation of STEP across both datasets at $T = 12, K = 48$. Removing our core architectural choices or aggressively dropping uncertain training poses results in consistent performance drops across both datasets. All AUROC (%) results are reported for a fixed seed.

Configuration	ShanghaiTech		UBnormal	
Full STEP (Ours)	**86.3**	Δ	**90.4**	Δ
Architecture & PC-space				
w/o $\sigma-$modulation Residual MLP (Vanilla MLP)	85.6	-0.7	89.3	-1.1
w/o Whitening	85.5	-0.8	88.2	-2.2
Training Confidence Handling				
Drop Pose (Confidence < 0.1)	86.4	+0.1	90.5	+0.1
Drop Pose (Confidence < 0.2)	86.0	-0.3	89.6	-0.8
Drop Pose (Confidence < 0.4)	84.3	-2.0	85.9	-4.5
Noise Boundaries (vs. $\sigma_{low} = 0.1, \sigma_{high} = 1.0$)				
Minimum Noise $\sigma_{low} = 10^{-4}$	86.1	-0.2	89.6	-0.8
Minimum Noise $\sigma_{low} = 0.5$	82.7	-3.6	86.9	-3.5
Maximum Noise $\sigma_{high} = 0.2$	86.4	+0.1	89.9	-0.5
Maximum Noise $\sigma_{high} = 0.5$	86.4	+0.1	90.6	+0.2

macroscopic validity of the pose itself. By maintaining a lower bound ($\sigma_{low} = 0.1$), the noise acts as an implicit kernel smoother in the PC space, absorbing minor coordinate fluctuations and forcing the network to model true kinematic structure rather than memorizing noisy training samples. Conversely, if σ_{low} is too large (0.5), the density estimate becomes over-smoothed, blurring the fine-grained, high-density basins of normal motion and dropping performance by 3.5% on UBnormal and 3.6% on ShanghaiTech. At the other end of the spectrum, an insufficient maximum noise scale ($\sigma_{high} = 0.2$) limits the spatial reach of the score estimator. While this narrower coverage is adequate for the subtle anomalies in ShanghaiTech, it fails to provide gradient signals in the sparse regions distant from the primary PC-space cluster. Consequently, the network remains untrained for the extreme anomalous poses found in UBnormal, resulting in less reliable energy scores and a slight performance drop (-0.5%). Our selected bounds provide adequate density coverage for the unit-variance PCA space across both benchmarks.

4.4 Computational Efficiency

Our lightweight architecture is highly efficient. Timing the complete pipeline, from the already-tracked pose sequences through PCA projection, whitening, and final anomaly scoring, reveals that latency is practically invariant to T. On a NVIDIA GTX 1080, this entire scoring process for 50 persons per frame takes < 1 ms ($\sim$1026 FPS) with a 32.6 MB memory footprint. Scaling to 100 persons takes just 1.62 ms ($\sim$618 FPS, 44.9 MB). Thus, STEP provides real-time, state-of-the-art anomaly detection at a fractional cost to the underlying pose extraction.

5 Limitations

While our proposed STEP framework sets a new benchmark for skeleton-based VAD, we acknowledge several important limitations regarding our design choices.

Pose Estimator: STEP relies on the upstream pose estimators. While total occlusion inevitably interrupts data extraction, the temporal nature of our sequence-level scoring ensures that the irregular kinematics leading up to an anomaly (*e.g.*, falling) are typically flagged before the subject is fully obscured.

Contextual and Multi-Person Anomalies: By design, our framework evaluates subjects independently and strips away pixel-level appearance to ensure robustness against background clutter and to preserve privacy. However, a fundamental trade-off of this pure pose formulation is that it naturally struggles with anomalies defined entirely by multi-person interactions (*e.g.*, illegal exchange of goods) or object manipulation (*e.g.*, abandoning a bag) if the underlying posture remains typical. We argue that while reintroducing visual appearance features could easily capture this missing context, doing so fundamentally compromises the privacy-preserving properties that motivate skeleton-based anomaly detection in the first place. Instead, explicitly modeling multi-person relational dynamics using purely pose features remains a key direction for future work.

6 Conclusion

We presented STEP, a robust and highly efficient score-based framework for skeleton-based Video Anomaly Detection. To prevent the structural collapse that occurs when applying Denoising Score Matching directly to raw temporal coordinates, we project the pose sequences onto a compact, whitened PC-space. This low-dimensional representation, coupled with a soft sequence-level confidence weighting, allows our lightweight model to robustly score anomalies while naturally mitigating upstream tracking artifacts. Consequently, our architecture is efficient, capable of processing dense crowds with minimal computational overhead. Extensive experiments confirm the effectiveness of our approach, with STEP setting a new state-of-the-art on UBnormal and matching the performance on ShanghaiTech.

Acknowledgements

This work received funding from the BRIDGE programme of the Republic of Austria, Federal Ministry of Innovation, Mobility and Infrastructure (BMIMI) under the project ReCoDi (914991). The Austrian Research Promotion Agency (FFG) has been authorised for the programme management.

References

1. Acsintoae, A., Florescu, A., Georgescu, M., Mare, T., Sumedrea, P., Ionescu, R.T., Khan, F.S., Shah, M.: Ubnormal: New benchmark for supervised open-set video anomaly detection. In: CVPR (June 2022)
2. Barbalau, A., Ionescu, R.T., Georgescu, M.I., Dueholm, J., Ramachandra, B., Nasrollahi, K., Khan, F.S., Moeslund, T.B., Shah, M.: Ssmtl++: Revisiting self-supervised multi-task learning for video anomaly detection. Computer Vision and Image Understanding **229** (2023)
3. Delić, A., Grčić, M., Šegvić, S.: Sequential keypoint density estimator: an overlooked baseline of skeleton-based video anomaly detection. In: ICCV (2025)
4. Fang, H.S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., Li, Y.L., Lu, C.: Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. IEEE TPAMI (2022)
5. Flaborea, A., Collorone, L., di Melendugno, G.M.D., D'Arrigo, S., Prenkaj, B., Galasso, F.: Multimodal motion conditioned diffusion model for skeleton-based video anomaly detection. In: ICCV (2023)
6. Flaborea, A., di Melendugno, G.M.D., D'Arrigo, S., Sterpa, M.A., Sampieri, A., Galasso, F.: Contracting skeletal kinematics for human-related video anomaly detection. PR (2024)
7. Georgescu, M.I., Barbalau, A., Ionescu, R.T., Khan, F.S., Popescu, M., Shah, M.: Anomaly detection in video via self-supervised and multi-task learning. In: CVPR (2021)
8. Georgescu, M.I., Ionescu, R.T., Khan, F.S., Popescu, M., Shah, M.: A background-agnostic framework with adversarial training for abnormal event detection in video. IEEE TPAMI **44**(9) (2021)
9. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S.: Learning temporal regularity in video sequences. In: CVPR (2016)
10. Hirschorn, O., Avidan, S.: Normalizing flows for human pose anomaly detection. In: ICCV (2023)
11. Hyvärinen, A.: Estimation of non-normalized statistical models by score matching. JMLR **6**(24), 695–709 (2005)
12. Ionescu, R.T., Khan, F.S., Georgescu, M.I., Shao, L.: Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In: CVPR (2019)
13. Jain, Y., Sharma, A.K., Velmurugan, R., Banerjee, B.: Posecvae: Anomalous human activity detection. In: ICPR (2020)
14. Kanu-Asiegbu, A.M., Vasudevan, R., Du, X.: Bipoco: Bi-directional trajectory prediction with pose constraints for pedestrian anomaly detection. arXiv (2022)
15. Li, N., Chang, F., Liu, C.: Human-related anomalous event detection via spatial-temporal graph convolutional autoencoder with embedded long short-term memory network. Neurocomputing (2021)
16. Liu, W., Luo, W., Lian, D., Gao, S.: Future frame prediction for anomaly detection–a new baseline. In: CVPR (2018)
17. Luo, W., Liu, W., Gao, S.: A revisit of sparse coding based anomaly detection in stacked rnn framework. In: ICCV (2017)
18. Luo, W., Liu, W., Gao, S.: Normal graph: Spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection. Neurocomputing (2020)
19. Markovitz, A., Sharir, G., Friedman, I., Zelnik-Manor, L., Avidan, S.: Graph embedded pose clustering for anomaly detection. In: CVPR (2020)

20. Micorek, J., Possegger, H., Narnhofer, D., Bischof, H., Koziński, M.: MULDE: Multiscale Log-Density Estimation via Denoising Score Matching for Video Anomaly Detection. In: CVPR. pp. 18868–18877 (June 2024)
21. de Morais, R.F.A.B., Le, V., Tran, T., Saha, B., Mansour, M.R., Venkatesh, S.: Learning regularity in skeleton trajectories for anomaly detection in videos. In: CVPR (2019)
22. Nguyen, T.N., Meunier, J.: Anomaly detection in video sequence with appearance-motion correspondence. In: ICCV (2019)
23. Park, H., Noh, J., Ham, B.: Learning memory-guided normality for anomaly detection. In: CVPR (2020)
24. Rodrigues, R., Bhargava, N., Velmurugan, R., Chaudhuri, S.: Multi-timescale trajectory prediction for abnormal human activity detection. In: WACV (2020)
25. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. In: NeurIPS (2019)
26. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: CVPR (2018)
27. Vincent, P.: A connection between score matching and denoising autoencoders. Neural Computation **23**(7), 1661–1674 (2011)
28. Wang, G., Wang, Y., Qin, J., Zhang, D., Bao, X., Huang, D.: Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In: ECCV (2022)
29. Wang, J., Cherian, A.: Gods: Generalized one-class discriminative subspaces for anomaly detection. In: ICCV (2019)
30. Yan, C., Zhang, S., Liu, Y., Pang, G., Wang, W.: Feature prediction diffusion model for video anomaly detection. In: ICCV (2023)
31. Yu, G., Wang, S., Cai, Z., Zhu, E., Xu, C., Yin, J., Kloft, M.: Cloze test helps: Effective video anomaly detection via learning to complete video events. In: ACM MM (2020)
32. Zaheer, M.Z., Mahmood, A., Khan, M.H., Segu, M., Yu, F., Lee, S.I.: Generative cooperative learning for unsupervised video anomaly detection. In: CVPR (2022)
33. Zhu, L., Wang, L., Raj, A., Gedeon, T., Chen, C.: Advancing video anomaly detection: A concise review and a new dataset. In: NeurIPS (2024)