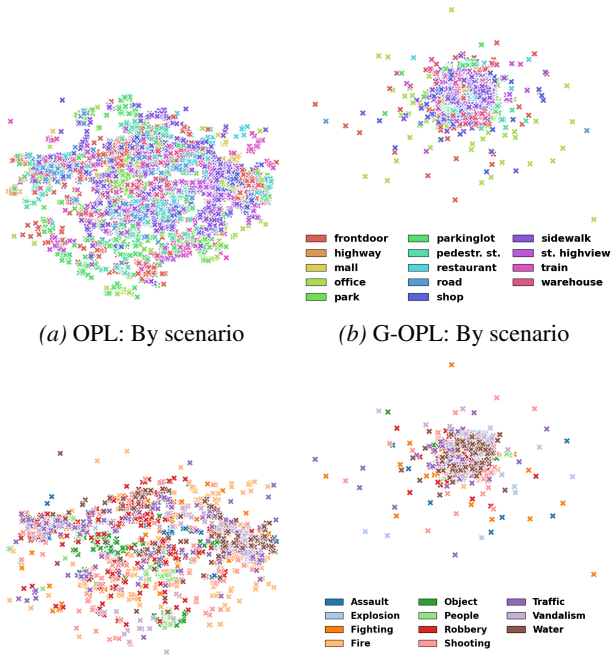
Privacy-Aware Video Anomaly Detection through Orthogonal Subspace Projection

Lei Wang [1 2] **Wenxiang Diao** [3] **Andrew Busch** [1] **Jun Zhou** [4] **Yongsheng Gao** [1]

Abstract

Video anomaly detection (VAD) systems often prioritize accuracy while overlooking privacy concerns, limiting their suitability for real-world deployment. We propose the Orthogonal Projection Layer (OPL), a lightweight module that removes task-irrelevant variations to produce representations focused on anomaly-relevant cues. To address privacy risks in human-centered scenarios, we introduce Guided OPL (G-OPL), which suppresses facial attributes using weak supervision from face-presence signals while preserving non-identifying features such as pose and motion. A cosine alignment objective enforces consistent capture and removal of facial information without identity labels or adversarial training. We further present a privacy-aware evaluation framework that jointly assesses detection performance and privacy preservation, and enables analysis of how sensitive information is filtered. Experiments show that embedding privacy constraints into model design reduces sensitive information while maintaining or improving detection accuracy, supporting projection-based architectures as a principled approach for privacy-aware VAD.



(a) OPL: By scenario *(b)* G-OPL: By scenario

(c) OPL: By anomaly type *(d)* G-OPL: By anomaly type

Figure 1. UMAP (McInnes et al., 2018) of *removed features* on MSAD (Zhu et al., 2024). OPL removes nuisance factors, yielding dispersed clusters, while G-OPL suppresses facial cues, producing a compact, overlapping distribution not aligned with anomaly types. This contrast shows their complementary roles in separating nuisance and privacy-sensitive information.

1. Introduction

Video Anomaly Detection (VAD) is a longstanding challenge in computer vision, with wide-ranging applications in public safety, infrastructure monitoring, and surveillance systems (Wang et al., 2019; Ding & Wang, 2024; Wang & Koniusz, 2024; Ding & Wang, 2025; Wang & Koniusz, 2025; Ding et al., 2026). Recent progress in VAD has largely focused on pushing the boundaries of performance through increasingly complex architectures and learning objectives. However, much of this work overlooks critical ethical dimensions, particularly those concerning privacy and interpretability. As AI systems transition into high-stakes, real-world environments, addressing these dimensions is not optional but essential for responsible deployment.

A key limitation of existing VAD models is their tendency to retain and exploit information that is either irrelevant to the task or ethically sensitive. Attributes such as faces, clothing styles, or background context are rarely necessary for identifying anomalies and can even introduce bias, overfitting, or privacy violations. Without explicit mechanisms to suppress

[1]School of Engineering and Built Environment, Griffith University [2]Data61/CSIRO [3]School of Computer Science and Engineering, University of New South Wales [4]School of Information and Communication Technology, Griffith University. Correspondence to: Yongsheng Gao <yongsheng.gao@griffith.edu.au>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

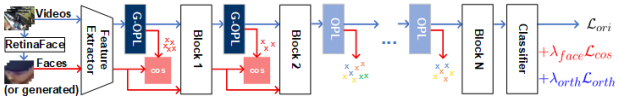


Figure 2. Pipeline overview. The Orthogonal Projection Layer (OPL, light blue) and its guided variant (G-OPL, dark blue and red) are lightweight, fully differentiable, and easily integrate into standard anomaly detection architectures (black). OPL suppresses nuisance factors through orthogonal projections, while G-OPL explicitly removes sensitive attributes via semantic suppression ($\lambda_{\text{face}}\mathcal{L}_{\text{cos}}$) guided by attribute signals (*e.g.*, faces, red arrows). An orthogonality regularizer ($\lambda_{\text{orth}}\mathcal{L}_{\text{orth}}$) ensures stable optimization. OPL / G-OPL can be flexibly inserted after a feature extractor (or encoder), individual layer, or block. This design embeds privacy directly into the model, enabling precise control over sensitive information at the representation level with minimal impact on detection performance. See Appendix H, I, and M for *practical insights on layer placement* and *privacy protection*.

these factors, current models risk learning representations that undermine both fairness and user trust. To address this, we propose a principled architectural solution for embedding ethical constraints directly into the learning process. We introduce the Orthogonal Projection Layer (OPL), a lightweight, differentiable module that learns to project intermediate features away from subspaces containing task-irrelevant or undesirable information. Initially, OPL focuses on removing general nuisance factors such as lighting variations, background clutter, and camera motion, guided by the VAD objective itself.

We further extend this idea with the Guided Orthogonal Projection Layer (G-OPL), which targets the removal of explicitly sensitive attributes such as faces. G-OPL uses weak supervision via lightweight face detectors, *e.g.*, RetinaFace (Deng et al., 2020), to extract face embeddings on the fly. These embeddings serve as geometric guides, and a cosine similarity loss encourages the projection basis to align with face-related components in the latent space, so that they can be effectively suppressed. Importantly, our method avoids adversarial training or gradient reversal, leading to more stable optimization and interpretable subspace disentanglement. Fig. 1 illustrates the interpretability benefits of our approach: OPL isolates nuisance factors related to scene dynamics, while G-OPL compacts sensitive biometric information, revealing their complementary roles in shaping ethical feature spaces. We perform an extensive analysis of where and how to integrate these projection layers, what types of information they remove, and how they impact both model utility and ethical alignment. We also introduce novel privacy-aware metrics tailored to VAD settings, enabling a more holistic evaluation of ethical trade-offs in VAD. Our contributions are summarized as follows:

i. We propose OPL, a differentiable projection mechanism that filters task-irrelevant features by learning orthogonal subspaces, and extend it to Guided OPL (G-OPL) for explicitly suppressing sensitive attributes such as faces.

ii. We introduce privacy-aware evaluation metrics for VAD, enabling principled assessment of the trade-offs between performance, interpretability, and privacy preservation.

iii. We conduct thorough experiments analyzing the placement and frequency of projection layers, the nature of removed information, and their impact on both anomaly detection and ethical alignment, offering actionable insights for ethical model design.

2. Related Work

Privacy-preserving representation learning. Learning representations that do not encode sensitive attributes is a core objective in privacy-aware machine learning (Kim et al., 2019; Ravfogel et al., 2020; Abadi et al., 2016; Dwork, 2025). A common strategy is adversarial training with gradient reversal (Ganin et al., 2016), where an auxiliary classifier attempts to predict sensitive attributes, and the encoder is trained to inhibit this through gradient inversion. Other approaches perform subspace manipulation, such as nullspace projection (Kim et al., 2019; Ravfogel et al., 2020), which analytically removes directions associated with sensitive information. These methods have primarily been developed for classification tasks in fairness and demographic privacy contexts. In contrast, we adapt these principles to the more complex VAD setting, which presents unique challenges due to the spatio-temporal nature of video and the lack of explicit labels for nuisance or sensitive factors. Our method avoids adversarial objectives, using geometric supervision via cosine similarity and weak face signals for more stable and interpretable training. Here, geometric refers to latent feature alignment, not physical 3D geometry.

Orthogonal projection and feature disentanglement. Orthogonal projection has been used in disentangled representation learning to separate task-relevant and task-irrelevant information (Moyer et al., 2018; Sarhan et al., 2020; Ranasinghe et al., 2021). These methods often rely on known sensitive labels or contrastive setups in static image tasks. Our OPL generalizes this idea for dynamic, unlabeled video data by learning to remove nuisance subspaces directly through VAD loss. We further introduce G-OPL, which incorporates weak supervision from faces to target privacy-relevant information. Unlike classical projection approaches, G-OPL is fully differentiable, requires no explicit sensitive labels, and aligns privacy removal with interpretable geometric signals.

Ethics and interpretability in VAD. While ethical AI has received increasing attention, especially around fairness and transparency, most work in VAD treats these aspects as afterthoughts. Techniques such as dataset balancing or saliency visualization are often used post hoc to audit ethical behavior. Explainable AI methods like Grad-CAM

(Selvaraju et al., 2020) and causal analysis in VAD (Du et al., 2024; Ding et al., 2026) offer interpretability but do not directly influence representation learning. Our approach goes further by embedding ethical constraints at the model level through projection layers that structurally suppress sensitive content in the latent space. This leads to improved privacy and interpretability, all without sacrificing scalability or compatibility with existing architectures.

Comparison with prior projection-based methods. While orthogonal projection has been explored in fairness and disentangled representation learning, like INLP (Ravfogel et al., 2020), OPL-2021 (Ranasinghe et al., 2021), DAMS (An et al., 2025) and CAE-LSP (Yu et al., 2023). These methods differ from ours in terms of objectives, assumptions, and design. Most prior work either assumes access to ground-truth sensitive labels (*e.g.*, INLP), operates in static or low-dimensional settings (*e.g.*, CAE-LSP), or applies generic decorrelation losses (*e.g.*, OPL-2021) without explicit suppression. In contrast, our G-OPL learns task-specific sensitive subspaces using only weak signals (*e.g.*, face presence or embeddings), and uniquely integrates privacy-aware evaluation metrics tailored for VAD.

To the best of our knowledge, this is the first projection-based method designed for privacy-aware VAD under realistic constraints.

3. Method

Overview. We propose the Orthogonal Projection Layer (OPL) that filters nuisance directions from features via learned subspace projections. We further extend this to Guided OPL (G-OPL), which enables targeted suppression of sensitive features (*e.g.*, faces) using weak supervision (*e.g.*, face presence). Both modules are lightweight, fully differentiable, and compatible with standard VAD architectures. They can be flexibly inserted after a feature extractor (or encoder), individual layer, or block. Fig. 2 provides an overview of the model architecture with OPL and G-OPL layers inserted. To evaluate the utility-privacy trade-off, we introduce three novel metrics tailored to VAD. The appendix (A-F) outlines the theory, distinguishes our model-level protection from dataset- and learning-level approaches, and situates it within the broader literature.

Notations. Throughout this paper, scalars are denoted by italic letters (*e.g.*, x); vectors by lowercase boldface (*e.g.*, $\boldsymbol{x}$); and matrices by uppercase boldface (*e.g.*, $\boldsymbol{X}$).

3.1. Orthogonal Projection Layer

The *Orthogonal Projection Layer* (OPL) is designed to explicitly remove task-irrelevant information from intermediate feature representations. Unlike classical methods such as PCA or fixed subspace removal, our OPL learns the nuisance subspace jointly with the main task and is fully differentiable, allowing adaptive identification and removal of nuisance directions during training. [1]

Formally, let $\boldsymbol{f} \in \mathbb{R}^d$ be an intermediate feature extracted from either a backbone network (*e.g.*, feature extractor) or a layer. We define a learnable weight matrix $\boldsymbol{W} \in \mathbb{R}^{k \times d}$ whose rows parameterize the nuisance subspace to be removed ($1 < k < d$). This matrix can be learned from scratch alongside the VAD objective. To obtain a numerically stable and orthonormal basis for the nuisance subspace, we perform QR decomposition on the transpose of $\boldsymbol{W}$:

$$\boldsymbol{W}^\top = \boldsymbol{Q}\boldsymbol{R}, \tag{1}$$

where $\boldsymbol{Q} \in \mathbb{R}^{d \times k}$ contains k orthonormal basis vectors spanning the nuisance directions, and $\boldsymbol{R}$ is an upper-triangular matrix. Using $\boldsymbol{Q}$, we construct the orthogonal complement projection matrix:

$$\boldsymbol{P} = \boldsymbol{I}_d - \boldsymbol{Q}\boldsymbol{Q}^\top, \tag{2}$$

with $\boldsymbol{I}_d$ being the $d \times d$ identity matrix. Projecting $\boldsymbol{f}$ onto the orthogonal complement yields:

$$\boldsymbol{f}_{\text{proj}} = \boldsymbol{P}\boldsymbol{f} = \boldsymbol{f} - \boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{f}, \tag{3}$$

removing nuisance components while retaining task-relevant information. Our OPL stands out by offering a plug-and-play, learnable projection layer that is task-specific, modular, and interpretable. The integration of QR decomposition ensures practical stability and efficiency lacking in prior projection-based approaches. These properties establish OPL as a foundational building block for task-specific and robust video analysis systems.

3.2. Guided OPL for Privacy-Aware Face Suppression

While OPL removes nuisance factors like background or lighting, it may leave sensitive attributes such as face, gender, or age in the features. We propose the *Guided Orthogonal Projection Layer* (G-OPL) to selectively suppress these privacy-sensitive components using weak supervision.

Facial identity as a sensitive signal. We focus on facial identity not for recognition, but for removing biometric traits implicitly encoded in facial features. Although typical VAD datasets lack identity annotations, the mere presence of faces introduces privacy risks. Intermediate neural features can retain identity cues that, if leaked or combined with external data, may enable re-identification.

To reduce privacy risks, we use binary face-presence indicators as weak supervision, automatically extracted via

[1] OPL does not arbitrarily remove variance; it is trained jointly with the anomaly objective, thus preserving components essential for detection.

RetinaFace (Deng et al., 2020), which handles small, occluded, or low-resolution faces. When multiple faces are detected, we average their embeddings to guide the projection and separate identity-related from task-relevant features. We also create 50 face videos from the Georgia Tech Face Database (Nefian), using their average embeddings as controlled sensitive signals.

Face-guided suppression. Both the original video frames and detected / generated face crops are passed through the same encoder (*e.g.*, I3D). This ensures that the embeddings $\boldsymbol{f}$ (from video) and $\boldsymbol{f}_{\text{face}}$ (from cropped faces) reside in the same latent space, allowing us to compare them meaningfully. Since both embeddings are high-level features, our method operates entirely at the feature level rather than in the pixel space. This avoids the need for explicit face masks or boundary annotations, and makes the projection robust to partial occlusions, variable face sizes, and fuzzy edges commonly seen in surveillance videos. The goal is to identify and remove the facial component from $\boldsymbol{f}$ using projection operator (Equation 3). Instead of adversarial training or an explicit sensitive attribute classifier, G-OPL adopts a direct geometric approach. We guide the projection basis $\boldsymbol{Q}$ to span directions aligned with facial identity by penalizing cosine similarity between the face embedding and the sensitive component of $\boldsymbol{f}$:

$$\mathcal{L}_{\text{task}} = \mathcal{L}_{\text{ori}} + \lambda_{\text{face}} \left(1 - \cos\left(\boldsymbol{f}_{\text{face}}, \boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{f}\right)\right), \quad (4)$$

where $\mathcal{L}_{\text{ori}}$ is the standard VAD loss, and λ_{face} controls the strength of the privacy-driven guidance. This alignment term encourages $\boldsymbol{Q}$ to capture face-related directions, which the projection operator then removes (Equation 2). The loss is activated only when a face is detected, ensuring efficient and targeted updates.

Orthogonality regularization. Although QR ensures orthonormality per forward pass, gradient updates can erode it over time. To preserve orthogonality and subspace separation, we include:

$$\mathcal{L}_{\text{orth}} = \left\|\boldsymbol{Q}^\top \boldsymbol{Q} - \boldsymbol{I}_k\right\|_F^2, \quad (5)$$

where $\|\cdot\|_F$ is the Frobenius norm. The overall loss becomes:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_{\text{orth}} \mathcal{L}_{\text{orth}}, \quad (6)$$

with λ_{orth} controlling regularization strength. This is particularly important when stacking multiple projection layers or operating in high-dimensional spaces.

Multi-attribute suppression. Because $\boldsymbol{W}$ in OPL can have multiple rows ($k > 1$), G-OPL can jointly encode several sensitive attributes by concatenating multiple weak attribute embeddings $\boldsymbol{f}_{\text{attr}}^{(i)}$ and optimizing a multi-term alignment loss: $\sum_i (1 - \cos(\boldsymbol{f}_{\text{attr}}^{(i)}, \boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{f}))$. G-OPL does not require training attribute classifiers nor collecting attribute labels. The guidance embedding is not an attribute-specific representation; it is simply the mean backbone feature over frames where a weak detector signals the presence of the attribute (*e.g.*, face present, torso visible, clothing-color region detected). No identity or attribute labels are used at any point. Extending G-OPL to additional attributes (*e.g.*, body, clothing) requires only a weak presence signal, not attribute supervision. In the Appendix (G- I, M), we detail the *implementation of OPL and G-OPL* and discuss *practical considerations for their placement within the network*.

Efficiency at inference. At test time, G-OPL behaves like OPL. The projection matrix $\boldsymbol{Q}$ is fixed, and no face detection or additional input is required. This ensures privacy-aware behavior with negligible extra runtime cost, critical for real-time or resource-constrained applications.

3.3. Privacy-Aware Evaluation Metrics

We propose three privacy-aware metrics to measure how well our projection layers suppress sensitive data while retaining task-relevant signals.

Sensitive Subspace Capture (SSC) measures the degree to which the learned projection subspace captures sensitive information (*e.g.*, faces). Given an orthonormal basis $\boldsymbol{Q} \in \mathbb{R}^{d \times k}$ that spans the sensitive subspace learned via G-OPL, and a batch of sensitive attributes $\boldsymbol{f}_{\text{attr}}^{(i)} \in \mathbb{R}^d$ (*e.g.*, face embeddings), we compute the projection of each embedding onto the subspace: $\hat{\boldsymbol{f}}^{(i)} = \boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{f}_{\text{attr}}^{(i)}$. The SSC score is defined as the average cosine similarity between the original embedding and its projected component:

$$\text{SSC} = \cos\left(\hat{\boldsymbol{f}}^{(i)}, \boldsymbol{f}_{\text{attr}}^{(i)}\right). \quad (7)$$

Intuitively, a higher SSC indicates that a larger portion of the original features lies within the sensitive subspace, implying that $\boldsymbol{Q}$ effectively isolates directions strongly associated with sensitive attributes. This makes SSC a valuable diagnostic tool for assessing whether the learned subspace aligns with privacy-relevant components in the data.

Anomaly Retention Distance (ARD). To evaluate how well the projection preserves the performance, we compare the *frame-level* anomaly score distributions before and after applying the projection layer. Let $\boldsymbol{y}_{\text{raw}}$ and $\boldsymbol{y}_{\text{proj}}$ denote the anomaly scores computed from the raw features and the projected features, respectively. We treat $\boldsymbol{y}_{\text{raw}}$ and $\boldsymbol{y}_{\text{proj}}$ as samples from underlying distributions and estimate their densities, denoted by $P_{\text{raw}}(y)$ and $P_{\text{proj}}(y)$, using kernel density estimation or histograms. The ARD is then defined as the Kullback-Leibler (KL) divergence between the two:

$$\text{ARD} = \text{KL}\left(P_{\text{raw}}(y) \,\|\, P_{\text{proj}}(y)\right). \quad (8)$$

A lower ARD value indicates that the projection has minimal impact on anomaly score distribution, *i.e.*, the detection

behavior remains consistent, suggesting that task-relevant information has been retained. Conversely, a high ARD implies that the projection has distorted the decision landscape, potentially removing useful cues along with sensitive ones. Thus, ARD serves as a quantitative measure of utility preservation in the presence of privacy-aware projections.

Privacy Decay (PD) & First-layer PD (FPD). To measure how effectively sensitive information is suppressed throughout the network, we introduce PD. At each layer / block l containing (or following) a projection module (*e.g.*, G-OPL), we extract intermediate feature $\boldsymbol{f}^{(l)}$ and train a lightweight classifier (*e.g.*, a linear SVM) to predict sensitive attributes (*e.g.*, face presence). The classification accuracy at layer l is denoted as $\mathrm{Acc}^{(l)}$, and PD is defined as:

$$\mathrm{PD} = \left\{ \left(l, \mathrm{Acc}^{(l)}\right) \right\}_{l=1}^{L}, \tag{9}$$

where L is the number of probed layers. Plotting $\mathrm{Acc}^{(l)}$ over l gives a PD curve, showing how sensitive information decays through network; sharper drops indicate stronger protection by G-OPL.

We further define *First-layer Privacy Decay (FPD)* as $\mathrm{Acc}^{(1)}$, which captures the face presence accuracy at the first G-OPL. FPD quantifies the immediate suppression of sensitive information at the network's entry point. A low FPD reflects strong initial privacy preservation, minimizing the risk of early-stage leakage and offering a clear diagnostic for the projection layer's effectiveness.

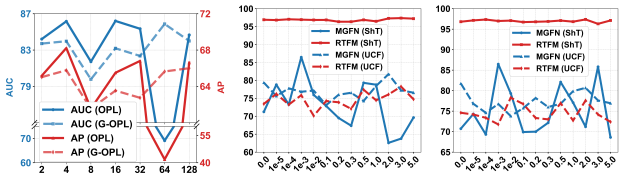
These metrics form a robust framework for evaluating privacy-aware VAD (see Appendix J). SSC diagnoses whether sensitive features are being captured and encoded consistently; ARD quantifies utility preservation; and PD / FPD shows how privacy is suppressed across the network.

Generality of privacy metrics. SSC and FPD are defined for generic sensitive attributes, with faces used only as an illustrative example. Their formulations are general and do not depend on G-OPL or any specific attribute type. They are diagnostic tools, not method-specific constructs.

4. Experiment

4.1. Experimental Setup

Datasets & protocols. Unless stated otherwise, we consider faces as the primary sensitive attribute, given their prevalence in VAD. We evaluate on five VAD datasets: ShanghaiTech (ShT) (Luo et al., 2017), UCF-Crime (UCF) (Sultani et al., 2018), CUHK Avenue (CUHK) (Lu et al., 2013), UCSD Ped2 (Ped2) (Vijay et al., 2010), and MSAD (Zhu et al., 2024). For MSAD, we use Protocol ii. For all datasets, we follow standard evaluation protocols and report frame-level AUC and/or Average Precision (AP). Beyond accuracy, we use our new privacy-aware metrics to quantify



(a) k *(b) λ_{face}* *(c) λ_{orth}*

Figure 3. Evaluation of key hyperparameters.

$k_{\text{OPL}} \backslash k_{\text{G-OPL}}$	2	4	8	16	32	64	128
2	95.5	95.9	95.6	94.8	93.9	92.5	91.8
4	95.9	**97.3**	96.8	95.2	94.0	92.8	91.9
8	95.7	97.0	96.5	95.0	93.8	92.6	91.7
16	95.2	96.3	95.8	95.0	93.6	92.2	91.5
32	94.5	95.1	94.6	93.8	92.8	91.5	90.8
64	93.6	94.0	93.2	92.0	90.5	88.9	88.3
128	92.8	93.1	92.4	91.5	89.8	88.4	87.9

Table 1. Decoupled analysis of OPL and G-OPL with independently varied ranks k_{OPL} and $k_{\text{G-OPL}}$ using ShanghaiTech. The table reports AUC (%) on the validation set. Best performance is achieved at $k_{\text{OPL}} = 4$ and $k_{\text{G-OPL}} = 4$.

how well sensitive attributes are suppressed, providing an ethical perspective on model performance.

To compare our metrics with a standard privacy evaluation, we adopt a classifier-based probe following (Ravfogel et al., 2020). Using a pre-trained ArcFace model, we extract facial embeddings and test whether projected features $\hat{\boldsymbol{f}}^{(i)}$ retain identity cues by learning a linear mapping back to the embedding space. We report rank-1 retrieval accuracy (lower is better), simulating an attacker attempting to recover identity from latent features.

Models & setups. We integrate OPL and G-OPL into four recent state-of-the-art weakly supervised VAD models: RTFM (Tian et al., 2021), MGFN (Chen et al., 2023b), TEVAD (Chen et al., 2023a), and EGO (Ding et al., 2025). These models rely on pre-trained backbones (*e.g.*, I3D (Carreira & Zisserman, 2017) and Swin Transformer (SwinT) (Liu et al., 2022)), making them naturally compatible with our plug-and-play projection modules. For OPL/G-OPL layer placement, we follow the *practical guidelines* detailed in the Appendix. We place G-OPL layers immediately after feature extractors (*e.g.*, I3D or SwinT) to suppress early-stage privacy-sensitive cues like faces, guided by weak face embeddings. OPL layers are inserted at deeper stages to filter task-irrelevant nuisances. For MSAD, as all videos have already been face-blurred for privacy protection, we directly use the official pre-extracted I3D and SwinT features provided by the dataset (Zhu et al., 2024) for our experiments.

4.2. Quantitative and Qualitative Evaluation

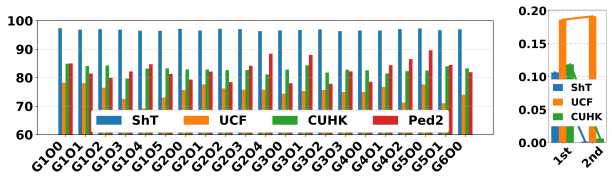
We evaluate our method comprehensively and summarize the key findings and insights below. Further analysis and

Method	Assault		Explosion		Fighting		Fire		Obj. Fall		People Fall		Robbery		Shooting		Traffic Acc.		Vandalism		Water Inc.		**Overall**	
	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP
RTFM (I3D)	53.9	66.4	66.0	76.6	79.8	88.6	44.9	71.1	84.6	89.3	45.7	52.6	70.2	88.0	87.5	89.2	64.1	57.7	74.9	73.0	98.1	99.6	86.6	68.4
MGFN (SwinT)	50.2	49.6	50.9	58.1	57.2	67.1	51.4	74.2	41.3	51.6	44.4	40.3	40.1	68.5	51.4	63.9	50.4	42.3	42.6	40.9	58.6	87.2	69.3	33.6
MGFN (I3D)	53.9	60.2	59.1	66.5	80.6	89.5	66.1	82.9	89.9	94.6	**53.6**	44.9	72.2	85.4	68.3	80.6	66.9	54.7	84.4	78.5	81.9	96.1	81.2	59.3
UR-DMU	56.9	64.5	67.9	74.5	83.9	90.4	61.2	82.9	92.1	95.8	42.5	43.7	63.5	79.3	81.4	87.8	62.0	55.6	84.7	77.0	98.5	99.5	85.0	68.3
EGO	52.2	57.5	57.6	74.4	66.5	72.8	62.9	86.7	92.3	94.8	35.4	43.8	64.8	87.5	68.6	78.4	69.9	**64.3**	88.1	81.4	81.9	95.4	87.3	64.4
IEF-VAD	66.0	-	66.3	-	79.8	-	49.4	-	75.9	-	42.5	-	66.9	-	86.9	-	**70.1**	-	75.8	-	88.9	-	82.1	-
RTFM-**OPL** (I3D)	57.0	62.4	**77.7**	**85.7**	74.1	84.8	49.6	75.5	87.7	92.1	53.3	50.4	**72.4**	**89.0**	84.1	89.5	69.5	58.7	84.8	80.9	**99.2**	**99.8**	86.5	68.2
MGFN-**OPL** (SwinT)	59.1	56.5	52.7	57.0	44.3	55.2	63.0	76.4	58.3	59.3	40.6	36.0	49.5	70.7	55.3	62.1	49.1	39.6	60.8	53.7	44.4	78.9	78.2	47.5
MGFN-**OPL** (I3D)	**71.3**	**69.4**	61.8	73.0	87.8	**92.8**	**81.0**	**93.0**	**94.3**	**96.5**	45.9	45.0	65.1	81.1	82.7	89.1	64.2	55.2	**90.8**	**86.4**	68.7	92.0	86.2	68.3
RTFM-**G-OPL/OPL** (I3D)	50.2	62.4	69.4	80.6	69.5	84.4	71.8	87.0	88.7	92.4	52.3	**53.3**	71.4	88.2	**87.8**	**91.0**	62.5	54.7	82.0	79.6	97.5	99.4	**88.0**	**70.9**
MGFN-**G-OPL/OPL** (I3D)	52.4	59.8	66.5	76.8	**88.8**	92.2	77.2	89.0	90.5	95.1	45.9	42.8	65.4	80.1	71.9	81.8	53.9	46.4	83.1	75.1	81.5	96.0	84.0	65.8

Table 2. Performance by anomaly type on MSAD. We compare against recent methods (same for Table 3): RTFM (Tian et al., 2021), MGFN (Chen et al., 2023b), UR-DMU (Zhou et al., 2023), EGO (Ding et al., 2025), and IEF-VAD (Jeong et al., 2025). **Bold** marks the best, underlined the second-best. Our **G-OPL/OPL** achieves competitive or superior results across anomaly types, highlighting its robustness, especially when paired with strong base models (*e.g.*, RTFM with I3D).

Method	Frontdoor		Mall		Office		Parkinglot		Pedestr. st.		Restaurant		Road		Shop		Sidewalk		St. highview		Train		Warehouse		**Overall**	
	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP
RTFM (I3D)	81.8	79.3	88.1	76.6	76.6	**72.8**	80.7	45.8	94.0	48.5	88.3	79.1	84.3	57.9	85.3	75.6	**88.3**	**68.8**	72.0	28.5	51.4	3.3	82.7	57.0	86.6	68.4
MGFN (SwinT)	59.5	51.7	18.5	20.1	64.1	52.3	67.9	19.0	75.9	9.7	67.9	44.0	70.6	26.3	62.7	43.0	69.0	25.9	75.3	23.3	65.4	5.2	70.1	30.1	69.3	33.6
MGFN (I3D)	82.5	80.8	73.8	71.3	71.5	58.2	68.9	14.8	94.8	36.2	95.1	**91.3**	76.5	35.8	85.6	78.4	78.5	57.2	77.9	29.3	40.3	2.1	58.3	24.2	81.2	59.3
UR-DMU	84.8	82.8	**91.0**	83.8	77.8	67.3	91.4	53.9	81.9	11.5	93.1	87.4	83.0	64.4	81.3	64.5	86.5	64.1	85.0	37.7	59.0	3.1	81.2	59.1	85.0	68.3
EGO	85.2	81.6	82.3	73.4	**80.0**	71.7	**96.8**	**75.2**	**97.5**	52.0	94.3	73.9	**89.8**	**64.6**	83.4	72.2	87.1	45.0	28.2	10.1	80.8	7.8	84.7	46.6	87.3	64.4
IEF-VAD	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	82.1	-
RTFM-**OPL** (I3D)	**85.6**	82.3	85.6	80.2	77.2	72.0	76.9	26.4	96.6	50.5	90.2	81.3	76.9	53.3	88.6	**82.8**	84.9	56.5	66.8	26.7	42.4	2.3	86.1	66.8	86.5	68.2
MGFN-**OPL** (SwinT)	68.5	57.8	89.0	61.8	68.4	55.4	79.4	39.0	74.5	5.0	51.6	36.1	67.3	28.1	77.1	60.3	81.1	41.9	87.1	45.8	**83.5**	11.9	83.8	52.4	78.2	47.5
MGFN-**OPL** (I3D)	84.4	**84.1**	80.2	74.7	74.7	65.0	87.0	30.9	93.5	**53.1**	91.2	87.6	80.0	55.7	82.1	69.4	86.8	63.8	**98.1**	**95.1**	70.8	9.1	**89.9**	76.1	86.2	68.3
RTFM-**G-OPL/OPL** (I3D)	82.0	79.3	**91.0**	81.4	74.3	72.0	79.4	27.2	86.9	36.1	90.3	81.4	72.4	46.7	**89.0**	82.5	87.0	65.1	84.9	37.8	70.4	**12.0**	86.3	**79.6**	**88.0**	**70.9**
MGFN-**G-OPL/OPL** (I3D)	84.4	83.3	90.0	**84.8**	75.9	62.3	70.4	16.9	90.5	25.8	**95.7**	90.2	71.4	43.1	79.7	64.5	83.8	63.3	87.7	41.3	44.7	2.3	64.8	41.1	84.0	65.8

Table 3. Performance by scenario on MSAD. Results on 12 test scenarios (excluding Highway and Park without anomalies) show our method's strong adaptability and robustness, consistently outperforming or matching top baselines and recent state-of-the-art methods while achieving better balance across scenarios (this table) and anomaly types (Table 2).



(a) Evaluation of projection layer placement. *(b)* PD.

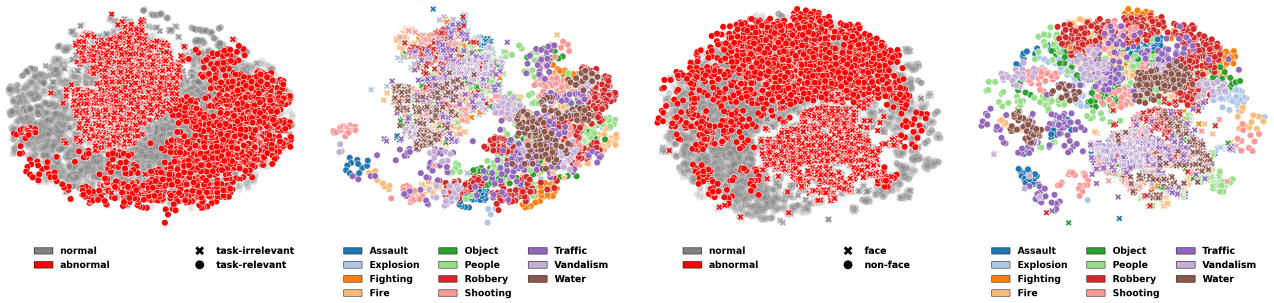
Figure 4. (a) Ablation on OPL/G-OPL placement across datasets with RTFM. GmOn: m G-OPL and n OPL layers. (b) We select model with more G-OPL layers that achieves reasonable VAD performance for privacy decay (PD) curve.

discussion are provided in the Appendix (K-L).

Sensitivity to k, λ_{face}, and λ_{orth}. As shown in Fig. 3a, performance improves as k increases from 2 to 16, peaking around $k = 4$, before degrading at $k = 64$. This suggests that small to moderate subspaces are sufficient to isolate nuisance factors, while overly large ones begin erasing task-relevant information. For both λ_{face} and λ_{orth} (Fig. 3b and 3c), we observe a stable performance window in the range $[10^{-4}, 10^{-2}]$. Within this range, semantic alignment and subspace decorrelation enhance disentanglement without impeding feature learning. However, large values hurt performance, likely due to over-penalization of useful variations or instability. Notably, ShanghaiTech remains relatively robust across a broad range, while UCF with MGFN, shows higher sensitivity, highlighting the greater importance of guided suppression in unstructured environments.

Decoupled analysis of rank k. Fig. 3a analyzes each module individually (*i.e.*, OPL-only or G-OPL-only). The non-monotonic behavior reflects the behavior of isolated components. We provide a decoupled analysis with independent ranks k_{OPL} and $k_{\text{G-OPL}}$ (Table 1). Best performance is at $k_{\text{OPL}} = 4$ and $k_{\text{G-OPL}} = 4$ (97.3% AUC). High performance is stable for $k_{\text{G-OPL}} = 4$ and $k_{\text{OPL}} \in \{4, 8, 16\}$. (iii) G-OPL is more sensitive to large ranks ($k_{\text{G-OPL}} > 8$), while OPL degrades more gradually ($k_{\text{OPL}} > 16$). We provide a simple practical guideline: (i) Heuristic: $k \approx$ 2-5% of the feature dimension (*e.g.*, $k \in [16, 50]$ for 1024-d features). (ii) Adaptive strategy: select the smallest k beyond which performance saturates (*e.g.*, based on ARD/FPD). This avoids exhaustive tuning in practice (see Appendix L).

Layer placement and frequency. As shown in Fig. 4a, applying a single early-stage G-OPL (*e.g.*, G1O0) consistently delivers strong and stable performance, achieving top or near-top results on ShT and CUHK. In contrast, progressively adding more OPL layers (G1O1 through G1O5) often leads to diminishing returns or even degraded performance, particularly on UCF. This supports our hypothesis that excessive disentanglement disrupts the retention of subtle anomaly cues, emphasizing the importance of strategic rather than frequent placement of these modules. Introducing G-OPL at deeper layers (G2-G6) maintains competitive



(a) OPL. *(b)* OPL (anomaly type) *(c)* G-OPL *(d)* G-OPL (anomaly type)

Figure 5. UMAP plots visualize task-relevant (dots) and removed nuisance/sensitive (crosses) features after the first OPL/G-OPL (using RTFM-I3D on MSAD). Colors indicate *frame-level* labels (normal, abnormal, anomaly types). Removal operates at the feature level, anomalies are detected from the remaining task-relevant features, not from removed components. (a) *vs.* (c): Both OPL and G-OPL successfully disentangle nuisance/sensitive features (crosses) from task-relevant ones (dots), but G-OPL yields more compact clusters of removed features due to its guidance from facial cues, offering clearer separation of irrelevant information. (b) *vs.* (d): For task-relevant features, G-OPL preserves more compact and semantically meaningful clusters for *human-centric* anomalies (*e.g.*, *robbery*, *vandalism*), revealing improved disentanglement of anomaly-relevant factors.

results on ShT but shows less stability on UCF and Ped2. Notably, configurations such as G5O0 achieve the highest AUC on Ped2, likely because this dataset contains low-resolution faces that are difficult to detect and thus better handled through late-stage disentanglement. These results validate our design principle: early-stage G-OPL provides the most reliable gains, while additional OPL layers should be used cautiously to avoid over-filtering. Later G-OPL layers may offer dataset-specific benefits, but overcomplicating the placement does not yield universal improvements.

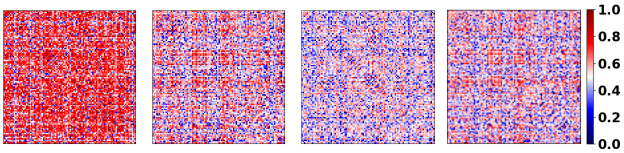
Effectiveness on MSAD across anomaly types and scenarios. Our improvements are evident across both anomaly types (*e.g.*, accidents *vs.* crimes) as shown in Table 2, and across diverse scenarios in Table 3, highlighting the robustness and adaptability of our approach. Notably, our method excels in challenging categories such as Fire and Vandalism, where subtle or occluded cues often hinder detection. This underscores the effectiveness of our disentanglement mechanism in filtering out task-irrelevant information while preserving critical anomaly-related signals. Furthermore, the performance gains are particularly pronounced when integrating OPL with stronger backbones like I3D, confirming the complementary nature of our design to existing architectures. While baselines such as EGO (Ding et al., 2025) exhibit robustness in certain scenarios through fusion strategies, they struggle with anomaly type diversity, where our method maintains balanced and consistently superior performance. Importantly, these findings suggest that our model not only mitigates the influence of noise and distractions but also enhances the model's ability to preserve causal and contextually relevant cues essential for accurate anomaly recognition (see also Fig. 5). This capability is particularly valuable in scenarios characterized by visual occlusions, complex dynamics, or subtle anomalies, where conventional fusion-based methods tend to underperform. This highlights the broader potential of our approach for advancing robust and generalizable VAD.

Discussion on $QQ^\top$ visualization. Fig. 6 visualizes the projection matrices $QQ^\top$ learned by G-OPL under both detected and generated faces. These matrices offer an interpretable window into how G-OPL identifies and suppresses privacy-sensitive directions in feature space. When using detected faces as guidance, $QQ^\top$ exhibits more diverse and irregular patterns. This reflects the variability inherent in real-world surveillance footage, where faces differ in pose, scale, occlusion, and lighting. Despite this variability, G-OPL successfully learns to capture meaningful face-related directions, as evidenced by consistent localized energy regions in the projection matrices (Fig. 6a & 6c). These patterns confirm that even weak, noisy supervision suffices to guide the suppression of sensitive information. In contrast, when guided by clean, generated faces from a controlled dataset, $QQ^\top$ reveals more structured and concentrated patterns. This indicates that the learned subspace is more focused and coherent (Fig. 6b & 6d), effectively aligning with the dominant variations associated with facial identity in a more stable and controlled manner. This demonstrates G-OPL's adaptability: cleaner supervision leads to more precise subspace removal. Its flexibility allows effective privacy-aware disentanglement using diverse weak supervision, from noisy real-world data to curated facial signals. The resulting $QQ^\top$ matrices provide interpretable evidence, offering transparency into how sensitive information is removed from learned features.

Privacy-aware metrics. As shown in Table 4, our method reduces identity leakage across both our proposed metrics (SSC/ARD/FPD) and the ArcFace probe, confirming its robustness and generality. Across datasets containing

Method	ShanghaiTech						UCF-Crime						CUHK Avenue						UCSD Ped2					
	AUC	AP	SSC↑	ARD↓	FPD↓	Arc↓	AUC	AP	SSC↑	ARD↓	FPD↓	Arc↓	AUC	AP	SSC↑	ARD↓	FPD↓	Arc↓	AUC	AP	SSC↑	ARD↓	FPD↓	Arc↓
USTN-DSC (Yang et al., 2023)	73.8	-	-	-	-	-	-	-	-	-	-	-	89.9	-	-	-	-	-	98.1	-	-	-	-	-
FPDM (Yan et al., 2023)	78.6	-	-	-	-	-	74.7	-	-	-	-	-	90.1	-	-	-	-	-	-	-	-	-	-	-
HSC (Sun & Gong, 2023)	83.4	-	-	-	-	-	-	-	-	-	-	-	93.7	-	-	-	-	-	98.1	-	-	-	-	-
PEL4VAD (Pu et al., 2024)	98.1	72.6	-	-	-	-	86.8	34.0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
VadCLIP (Wu et al., 2024)	-	-	-	-	-	-	88.0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
TeD-SPAD (Fioresi et al., 2023)	90.6	-	-	-	-	-	74.8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SPAct (Dave et al., 2022)	87.7	-	-	-	-	-	73.9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
MGNAD (Park et al., 2020)	70.5	-	-	-	-	-	-	-	-	-	-	-	88.5	-	-	-	-	-	97.0	-	-	-	-	-
MGFN (Chen et al., 2023b)	75.3	22.7	-	-	0.98	0.14	77.0	13.0	-	-	0.79	0.05	67.3	37.7	-	-	0.94	0.44	86.8	76.7	-	-	0.91	-
MGFN-**G-OPL/OPL**	**83.7**	**42.0**	0.10	0.23	0.96	**0.01**	**83.3**	**15.2**	**0.99**	**0.07**	0.49	**0.02**	**70.8**	**40.5**	0.39	**18.0**	**0.58**	**0.13**	**93.9**	**93.6**	0.39	0.94	**0.31**	-
MGFN-**G-OPL/OPL**†	**89.5**	**41.9**	**0.52**	**0.16**	**0.68**	**0.01**	**80.9**	**14.8**	0.72	6.46	**0.01**	0.03	**69.1**	**43.5**	**0.81**	23.6	1.0	0.31	NA	NA	NA	NA	NA	-
RTFM (Tian et al., 2021)	96.8	71.6	-	-	0.72	0.99	74.3	20.1	-	-	0.68	1.0	83.3	**66.3**	-	-	1.0	1.0	85.6	**82.0**	-	-	0.31	-
RTFM-**G-OPL/OPL**	**97.3**	**74.7**	**0.97**	0.29	0.19	**0.01**	**78.3**	**30.9**	**0.98**	**0.03**	**0.23**	0.03	**84.9**	66.2	**0.89**	**0.26**	**0.0**	**0.31**	**89.6**	75.0	0.98	0.06	**0.29**	-
RTFM-**G-OPL/OPL**†	**97.2**	**74.6**	0.20	**0.17**	**0.01**	0.02	74.1	**30.1**	0.33	0.30	0.64	**0.02**	**83.9**	65.7	0.11	3.86	**0.0**	**0.31**	NA	NA	NA	NA	NA	-
TEVAD (Chen et al., 2023a)	98.1	-	-	-	0.68	0.95	84.9	-	-	-	0.76	0.94	86.8	-	-	-	0.69	0.93	98.7	-	-	-	0.60	0.80
TEVAD-**OPL**	**98.4**	-	0.39	0.18	0.37	0.39	**85.3**	-	0.40	0.17	0.46	0.41	**87.3**	-	0.43	0.14	0.40	0.41	**98.9**	-	0.42	0.02	0.40	0.39
TEVAD-**G-OPL**	**98.2**	-	**0.92**	**0.11**	**0.07**	**0.04**	**85.0**	-	**0.88**	**0.15**	**0.11**	**0.06**	**87.1**	-	**0.93**	**0.13**	**0.09**	**0.05**	**98.8**	-	**0.79**	0.02	**0.17**	**0.07**
EGO (Ding et al., 2025)	97.3	-	-	-	0.72	0.98	81.7	-	-	-	0.78	0.96	83.1	-	-	-	0.70	0.97	93.2	-	-	-	0.65	0.85
EGO-**OPL**	**97.9**	-	0.39	0.19	0.41	0.42	**82.6**	-	0.45	0.24	0.49	0.44	**84.2**	-	0.45	0.22	0.39	0.46	**94.1**	-	0.45	0.03	0.46	0.44
EGO-**G-OPL**	**97.6**	-	**0.94**	**0.17**	**0.09**	**0.05**	**82.1**	-	**0.92**	**0.21**	**0.12**	**0.09**	**84.0**	-	**0.96**	**0.20**	**0.10**	**0.07**	**94.0**	-	**0.83**	0.03	**0.19**	**0.08**

Table 4. Results on ShanghaiTech, UCF-Crime, CUHK Avenue, and UCSD Ped2. **G-OPL/OPL** are further evaluated using privacy metrics (SSC↑, ARD↓, FPD↓). † uses RetinaFace embeddings; NA denotes undetectable faces on UCSD Ped2 due to low resolution. **Bold** highlights performance surpassing the baseline or the best privacy variant. Privacy metrics may underperform on datasets with low resolution or small/occluded faces. We also include ArcFace-based identity retrieval as a classifier-based privacy probe (Arc↓), providing an external reference point for comparison. MGNAD and SPAct are adapted to VAD; TeD-SPAD follows its own setup using MGFN+I3D. This is *the first comprehensive privacy analysis* for VAD across datasets.



(a) ShT (*det.*) *(b)* ShT (*gen.*) *(c)* UCF (*det.*) *(d)* UCF (*gen.*)

Figure 6. Visualization of $\boldsymbol{QQ}^\top$ (G-OPL of RTFM model) from detected (*det.*) and generated (*gen.*) faces on ShanghaiTech (ShT) and UCF-Crime (UCF). Central 100×100 regions, where energy concentrates, are shown to better reveal informative subspace patterns. All matrices are log-scaled, globally min-max normalized. Differences reflect dataset- and signal-specific subspace structures, highlighting distinct patterns of sensitive feature disentanglement.

Method	ShanghaiTech			CUHK Avenue		UCSD Ped2		UCF-Crime	
	AUC ↑	FPD ↓	ARD ↓	AUC ↑	FPD ↓	AUC ↑	FPD ↓	AUC ↑	FPD ↓
Baseline	96.8	0.42	0.08	85.6	0.37	98.1	0.05	84.5	0.48
+ Input masking (blur)	93.5	**0.18**	0.30	82.0	**0.16**	97.9	0.045	80.2	**0.20**
+ Differential privacy	95.8	0.34	0.19	84.1	0.31	97.8	0.045	83.1	0.39
+ Feature bottleneck	95.3	0.30	0.15	83.8	0.28	97.6	0.045	82.6	0.35
+ GRL (adversarial)	95.9	0.33	0.17	84.2	0.30	97.7	0.045	83.4	0.36
+ **OPL** (Ours)	97.1	0.30	0.07	87.2	0.28	**98.4**	0.045	85.2	0.34
+ **G-OPL** (Ours)	**97.3**	0.22	**0.06**	**89.0**	0.21	**98.4**	**0.042**	**85.4**	0.29

Table 5. Comparison with representative privacy-preserving strategies across datasets, including input-level masking, noise-based differential privacy (Dwork, 2025), compression-based feature bottleneck, and adversarial learning (GRL) (Ganin et al., 2016). Existing methods exhibit clear privacy-utility trade-offs, while our projection-based methods achieve a better balance, with G-OPL consistently providing stronger privacy and higher utility.

human faces, we observe reduced ARD and FPD (see PD curve in Fig. 4b) while SSC remains high, showing that sensitive components are removed without affecting anomaly-relevant semantics. This supports G-OPL's core idea: identity-related variations lie in a subspace orthogonal to anomaly-relevant signals and can be projected out without harming task utility. Using generated faces further improves suppression, achieving lower ARD and FPD than detected faces, demonstrating robustness under both synthetic and real privacy threats. These results provide the first comprehensive quantitative analysis of privacy leakage in VAD, highlighting G-OPL's practical impact. Importantly, our results show that the proposed metrics (SSC/ARD/FPD) correlate well with established ArcFace-based privacy probes. Notably, cases with low SSC and FPD also demonstrate low identity retrieval accuracy under ArcFace, confirming the consistency of these indicators. However, our metrics offer several advantages over classifier-based probe: they are more interpretable, do not require labeled identities, and are cheaper to compute, especially in the absence of clean face datasets. Thus, we advocate using SSC/ARD/FPD as diagnostic tools for privacy preservation, while ArcFace serves as a strong external attacker baseline.

Comparison with privacy-preserving methods. Although prior work is not specifically designed for privacy-preserving VAD, we adapt representative techniques from the privacy literature for comparison under a unified VAD setting. We consider adversarial learning with Gradient Reversal Layer (GRL) (Ganin et al., 2016), differential privacy (Abadi et al., 2016; Dwork, 2025) via Gaussian noise injection, feature bottlenecking through dimensionality reduction, and input masking via face blurring. All methods are

Method	#Layers	GFLOPs	ΔGFLOPs	AUC↑	FPD↓	ΔFPD	ΔFPD/ΔGFLOPs↑
Baseline	0	28.5	0.0	96.8	0.42	0.00	–
Early `G-OPL`	1	29.1	+0.6	97.2	0.28	0.14	**0.233**
Mid `G-OPL`	1	29.3	+0.8	97.1	0.26	0.16	0.200
Late `G-OPL`	1	29.4	+0.9	96.9	0.27	0.15	0.167
`G-OPL`s	2	30.2	+1.7	97.3	0.22	0.20	0.118
`G-OPL`s	3	31.5	+3.0	97.3	0.20	0.22	0.073

Table 6. Privacy-efficiency trade-off measured by privacy-per-GFLOP. ΔFPD is the reduction in FPD relative to the baseline. The ratio ΔFPD/ΔGFLOPs indicates privacy gain per additional computational cost (higher is better).

Attribute	AUC↑		FPD↓	
	Base	+ `G-OPL`	Base	+ `G-OPL`
Face	83.8	85.3	78.2	61.4
Gender	83.8	84.7	74.5	58.7
Clothing	83.8	84.6	69.3	55.2
Gait	83.8	83.2	71.0	57.9

Table 7. Attribute-wise anomaly detection performance and privacy leakage on ShanghaiTech. We report AUC as utility and FPD as a measure of privacy leakage for each attribute, where lower FPD indicates stronger suppression.

implemented on the same backbone with identical training protocols, using standard configurations for each approach. As shown in Table 5, input masking effectively suppresses face-related information but substantially degrades VAD performance by removing both identity and contextual cues. Differential privacy and feature bottlenecking achieve moderate reductions in privacy leakage, yet their impact on FPD is limited and often accompanied by a loss in discriminative capacity. GRL improves privacy to some extent but tends to introduce optimization instability, resulting in inconsistent utility-privacy trade-offs and occasional drops in AUC. In contrast, G-OPL consistently achieves the best overall balance, significantly reducing FPD across datasets while maintaining or improving AUC and preserving anomaly-relevant information, as reflected by low ARD. OPL also improves upon baselines, though with a more modest reduction in privacy leakage compared to G-OPL.

Privacy-per-GFLOP. We report a *normalized privacy-per-GFLOP metric*, defined as the relative reduction in FPD per unit GFLOPs (ΔFPD/ΔGFLOPs↑). Table 6 summarizes the results. One layer achieves the best privacy-per-GFLOP trade-off. Using 2-3 layers improves privacy but reduces efficiency, while deeper stacking yields the lowest FPD at the cost of higher computational overhead, making it suitable only when maximum privacy is required.

Trade-off analysis. Our results show a nuanced balance between detection performance, privacy, and interpretability. Integrating G-OPL often improves VAD metrics such as AUC and AP, likely because removing irrelevant identity components also eliminates spurious correlations that impair generalization. Minor fluctuations (*e.g.*, with detected faces) remain within acceptable bounds, confirming practical viability. Importantly, the projection matrix $\boldsymbol{Q}\boldsymbol{Q}^\top$ provides interpretable evidence of suppressed privacy-sensitive directions, a feature largely absent in prior work. This interpretability bridges performance-driven models and explainable AI, offering both methodological clarity and ethical accountability. Overall, G-OPL provides *a principled framework* that protects privacy, maintains performance, and enhances transparency in VAD.

Attribute-agnostic capability. We conduct experiments on multiple attributes and evaluate both single- (Table 7) and multi-attribute suppression (Table 8) using ShanghaiTech. We construct binary attribute signals: (i) Face: presence *vs.* absence, (ii) Gender: male *vs.* female (off-the-shelf classifier), (iii) Clothing: warm color *vs.* non-warm (HSV histogram thresholding), (iv) Gait: running *vs.* non-running (running as a proxy for gait variation). For multi-attribute suppression, we follow Sec. 3.2 and concatenate attribute signals into a joint vector for G-OPL. We report: (i) AUC for VAD, and (ii) attribute leakage/FPD (accuracy of a linear probe predicting the attribute from features; lower is better). Across all attributes, G-OPL reduces leakage by 12-17%. AUC improves most for face (+1.5) and remains positive for gender and clothing (+0.9 / +0.8), with a small, expected drop for gait (-0.6) due to its task relevance. This shows broad generalization beyond facial identity while respecting semantic roles. As more attributes are jointly suppressed, leakage drops monotonically (73.3% → 49.5%), while AUC remains stable (85.3% → 84.4%). The <1.0% AUC variation indicates a well-controlled privacy-utility trade-off.

Method	AUC↑	FPD↓
Baseline	83.8	73.3
Face	85.3	61.4
Face + Gender	85.1	56.2
Face + Gender + Clothing	84.9	52.8
Face + Gender + Clothing + Gait	84.4	49.5

Table 8. Effect of progressively incorporating identity-related cues, including face, gender, clothing, and gait, on anomaly detection performance (AUC) and attribute leakage (FPD).

5. Conclusion

We propose a novel VAD framework that integrates privacy and interpretability through the Orthogonal Projection Layer (OPL), which suppresses task-irrelevant variations via orthogonal subspace projection. To explicitly remove sensitive attributes such as faces, we extend OPL into the Guided OPL (G-OPL), using cosine alignment without relying on adversarial learning. We further introduce privacy-aware evaluation metrics to quantify privacy leakage in VAD. Our results demonstrate that privacy and interpretability can be effectively embedded into VAD architectures, enabling robust VAD without exploiting sensitive information.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

ACKNOWLEDGMENTS

Wenxiang Diao, a visiting scholar at the ARC Research Hub for Driving Farming Productivity and Disease Prevention, Griffith University, conducted this work under the supervision of Lei Wang. Lei Wang proposed the algorithm and developed the theoretical framework, while Wenxiang Diao implemented the code and performed the experiments.

We thank the anonymous reviewers for their invaluable insights and constructive feedback, which have contributed to improving our work.

This work was supported in part by the Australian Research Council (ARC) under Industrial Transformation Research Hub Grant IH180100002. This work was also supported by the National Computational Merit Allocation Scheme (NCMAS 2026, NCMAS 2025), and the ANU Merit Allocation Scheme (ANUMAS 2025), with computational resources provided by NCI Australia, an NCRIS-enabled capability supported by the Australian Government.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

An, D., Liu, W., Wang, K., Chen, Z., Lu, J., and Zhang, S. Dams: Dual-branch adaptive multiscale spatiotemporal framework for video anomaly detection. *arXiv preprint arXiv:2507.20629*, 2025.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, D. Mutual information neural estimation. In *International conference on machine learning*, pp. 531–540. PMLR, 2018.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

Cai, Z., Xiong, Z., Xu, H., Wang, P., Li, W., and Pan, Y. Generative adversarial networks: A survey toward private and secure applications. *ACM Computing Surveys (CSUR)*, 54(6):1–38, 2021.

Carreira, J. and Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

Chen, W., Ma, K. T., Jian Yew, Z., Hur, M., and Khoo, D. A.-A. Tevad: Improved video anomaly detection with captions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 5549–5559, 2023a. doi: 10.1109/CVPRW59228.2023.00587.

Chen, Y., Liu, Z., Zhang, B., Fok, W., Qi, X., and Wu, Y.-C. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37(1), pp. 387–395, 2023b.

Danesh Pazho, A., Alinezhad Noghre, G., Rahimi Ardabili, B., Neff, C., and Tabkhi, H. Chad: Charlotte anomaly dataset. In *Scandinavian Conference on Image Analysis*, pp. 50–66. Springer, 2023.

Dave, I. R., Chen, C., and Shah, M. Spact: Self-supervised privacy preservation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20164–20173, 2022.

Deng, J., Guo, J., Ververas, E., Kotsia, I., and Zafeiriou, S. Retinaface: Single-shot multi-level face localisation in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5202–5211, 2020. doi: 10.1109/CVPR42600.2020.00525.

Ding, D., Wang, L., Zhu, L., Gedeon, T., and Koniusz, P. Learnable expansion of graph operators for multi-modal feature fusion. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=SMZqIOSdlN.

Ding, X. and Wang, L. Quo vadis, anomaly detection? llms and vlms in the spotlight. *arXiv preprint arXiv:2412.18298*, 2024.

Ding, X. and Wang, L. Do language models understand time? In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 1855–1868, 2025.

Ding, X., Wang, L., Koniusz, P., and Gao, Y. Learning time in static classifiers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20816–20825, 2026.

Du, H., Zhang, S., Xie, B., Nan, G., Zhang, J., Xu, J., Liu, H., Leng, S., Liu, J., Fan, H., et al. Uncovering what why and how: A comprehensive benchmark for causation understanding of video anomaly. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18793–18803, 2024.

Dwork, C. Differential privacy. In *Encyclopedia of Cryptography, Security and Privacy*, pp. 649–652. Springer, 2025.

Eckart, C. and Young, G. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Elazar, Y. and Goldberg, Y. Adversarial removal of demographic attributes from text data. *arXiv preprint arXiv:1808.06640*, 2018.

Esmaeili, A., Asadi, E., and Marvasti, F. Iterative null-space projection method with adaptive thresholding in sparse signal recovery and matrix completion. *arXiv preprint arXiv:1610.00287*, 2016.

Fioresi, J., Dave, I. R., and Shah, M. Ted-spad: Temporal distinctiveness for self-supervised privacy-preservation for video anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13598–13609, 2023.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., and Lempitsky, V. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Gong, D., Liu, L., Le, V., Saha, B., Mansour, M. R., Venkatesh, S., and Hengel, A. v. d. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1705–1714, 2019.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

Jeong, S., Park, J., and Imani, M. Uncertainty-weighted image-event multimodal fusion for video anomaly detection. *arXiv preprint arXiv:2505.02393*, 2025.

Kim, B., Kim, H., Kim, K., Kim, S., and Kim, J. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9012–9020, 2019.

Kim, H. and Mnih, A. Disentangling by factorising. In *International conference on machine learning*, pp. 2649–2658. PMLR, 2018.

Li, A., Guo, J., Yang, H., Salim, F. D., and Chen, Y. Deep-obfuscator: Obfuscating intermediate representations with privacy-preserving adversarial learning on smartphones. In *Proceedings of the International Conference on Internet-of-Things Design and Implementation*, pp. 28–39, 2021.

Liu, W., Luo, W., Lian, D., and Gao, S. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6536–6545, 2018.

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., and Hu, H. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3202–3211, 2022.

Lu, C., Shi, J., and Jia, J. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pp. 2720–2727, 2013.

Luo, W., Liu, W., and Gao, S. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE international conference on computer vision*, pp. 341–349, 2017.

Luo, W., Liu, W., Lian, D., Tang, J., Duan, L., Peng, X., and Gao, S. Video anomaly detection with sparse coding inspired deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):1070–1084, 2021. doi: 10.1109/TPAMI.2019.2944377.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pp. 3384–3393. PMLR, 2018.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Mirjalili, V., Raschka, S., and Ross, A. Privacynet: Semi-adversarial networks for multi-attribute face privacy. *IEEE Transactions on Image Processing*, 29:9400–9412, 2020.

Moyer, D., Gao, S., Brekelmans, R., Galstyan, A., and Ver Steeg, G. Invariant representations without adversarial training. *Advances in neural information processing systems*, 31, 2018.

Nefian, A. V. Georgia tech face database. URL http://www.anefian.com/research/face_reco.htm.

Park, H., Noh, J., and Ham, B. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14372–14381, 2020.

Pu, Y., Wu, X., Yang, L., and Wang, S. Learning prompt-enhanced context features for weakly-supervised video anomaly detection. *IEEE Transactions on Image Processing*, 2024.

Ranasinghe, K., Naseer, M., Hayat, M., Khan, S., and Khan, F. S. Orthogonal projection loss. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12333–12343, 2021.

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., and Goldberg, Y. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 7237–7256, 2020.

Sarhan, M. H., Navab, N., Eslami, A., and Albarqouni, S. Fairness by learning orthogonal disentangled representations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pp. 746–761. Springer, 2020.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: visual explanations from deep networks via gradient-based localization. *International journal of computer vision*, 128:336–359, 2020.

Sultani, W., Chen, C., and Shah, M. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6479–6488, 2018.

Sun, S. and Gong, X. Hierarchical semantic contrast for scene-aware video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22846–22856, 2023.

Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J. W., and Carneiro, G. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4975–4986, October 2021.

Tishby, N., Pereira, F. C., and Bialek, W. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Vijay, M., LI, W.-X., Viral, B., and Nuno, V. Anomaly detection in crowded scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1975–1981, 2010.

Wang, L. and Koniusz, P. Flow dynamics correction for action recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3795–3799. IEEE, 2024.

Wang, L. and Koniusz, P. Feature hallucination for self-supervised action recognition. *International Journal of Computer Vision*, 133(11):7612–7646, 2025.

Wang, L., Huynh, D. Q., and Mansour, M. R. Loss switching fusion with similarity search for video classification. In *2019 IEEE international conference on image processing (ICIP)*, pp. 974–978. IEEE, 2019.

Wang, L., Gu, R., Zhuang, W., Gao, P., Wang, Y., and Zou, Y. Learning decoupling features through orthogonality regularization. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7562–7566. IEEE, 2022.

Wang, S., Li, X., Sun, J., and Xu, Z. Training networks in null space of feature covariance for continual learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 184–193, 2021.

Wu, P., Zhou, X., Pang, G., Zhou, L., Yan, Q., Wang, P., and Zhang, Y. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6074–6082, 2024.

Wu, T., Fischer, I., Chuang, I. L., and Tegmark, M. Learnability for the information bottleneck. In *Uncertainty in Artificial Intelligence*, pp. 1050–1060. PMLR, 2020a.

Wu, Z., Wang, H., Wang, Z., Jin, H., and Wang, Z. Privacy-preserving deep action recognition: An adversarial learning framework and a new dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2126–2139, 2020b.

Yan, C., Zhang, S., Liu, Y., Pang, G., and Wang, W. Feature prediction diffusion model for video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5527–5537, October 2023.

Yang, Z., Liu, J., Wu, Z., Wu, P., and Liu, X. Video event restoration based on keyframes for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14592–14601, 2023.

Yu, Q., Li, C., Zhu, Y., and Kurita, T. Convolutional autoencoder based on latent subspace projection for anomaly detection. *Methods*, 214:48–59, 2023.

Zhou, H., Yu, J., and Yang, W. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 3769–3777, 2023.

Zhu, L., Wang, L., Raj, A., Gedeon, T., and Chen, C. Advancing video anomaly detection: A concise review and a new dataset. In *The Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

A. Where Privacy Can Be Protected?

Facial biometric sensitivity. Faces are among the most sensitive biometric identifiers due to strong legal regulation, high uniqueness, and broad societal harm when misused. Other cues such as gait or body pose may also reveal identity, but facial identity is the most consistently regulated and the most easily exploited by automated recognition systems. Our experiments target faces because they are *the most reliably detectable in unconstrained surveillance videos and constitute the most legally sensitive signal* under GDPR (EU, 2018), CCPA (USA, 2020), BIPA (USA, 2008), *etc*. Nonetheless, our method is general and could suppress gait- or pose-related spaces if weak supervision for those cues is provided.

When facial suppression is appropriate *vs.* harmful? We agree that indiscriminately removing facial features is not universally desirable. The key clarification is that G-OPL is not intended as a universal default, but rather as *a configurable privacy-control mechanism* for scenarios where the operator wishes to suppress identity or facial clues. Our contributions target the growing deployment setting where: (i) VAD inference happens at the edge, but only intermediate features or anomaly scores are transmitted to the cloud, (ii) identity leakage is a regulatory or ethical risk (*e.g*., GDPR, BIPA), and (iii) anomaly definitions do not rely on identifying specific individuals. Importantly, the practitioner selects whether to use OPL or G-OPL based on the application domain: (i) Use OPL (nuisance suppression only) when facial information is part of the causal chain of anomaly definition (*e.g*., unauthorized entry, impersonation, personal aggression, employee-only zones). (ii) Use G-OPL (privacy-first feature suppression) when faces are legally sensitive but not required for anomaly semantics (*e.g*., crowd safety monitoring, fall detection, traffic flow anomalies, hospital ward monitoring, inappropriate access behavior).

Privacy preservation in machine learning spans multiple levels: from dataset design, to learning paradigms, to model-level architectural innovations. Each level introduces different mechanisms, assumptions, and trade-offs. In this section, we contextualize our work within this broader landscape and highlight how our approach complements and advances model-level privacy-preserving design.

Dataset-level. One intuitive approach to privacy preservation involves directly sanitizing the input data. This includes techniques such as face blurring, masking sensitive regions, downsampling, or removing identity-revealing cues. While effective in some cases, these methods can compromise data utility, especially in tasks requiring high-fidelity spatial or temporal information. Moreover, manual or heuristic filters often fail to generalize across domains and may not align well with downstream task requirements.

Learning-level. At the learning level, frameworks such as Federated Learning (FL) and Differential Privacy (DP) introduce algorithmic guarantees that limit exposure of raw data. FL decentralizes training by keeping data local to devices, while DP injects carefully calibrated noise into training or inference to bound the information leakage. Although theoretically rigorous, these methods can incur substantial performance or utility loss in practice. They also typically require significant changes to training protocols and infrastructure.

Model-level: our contribution. Our proposed framework operates at the model level by embedding privacy-aware mechanisms directly into the network architecture. Specifically, we introduce the OPL and its guided variant G-OPL, which learn to remove sensitive components from intermediate feature representations via low-rank, orthonormal projections. These modules act as semantic filters, discarding features that are predictive of protected attributes, such as facial identity or age, while retaining task-relevant signals.

This model-level intervention offers several advantages:

i. Architecture-integrated ethics. Privacy preservation is no longer an afterthought or external constraint, but a built-in property of the model's internal structure and training objective.

ii. Representation-level control. Unlike dataset filtering or global privacy budgets, OPL/G-OPL operate on structured neural representations where sensitive factors are often more linearly separable, allowing precise, learnable suppression.

iii. Minimal trade-off with utility. Our experiments show that VAD performance is largely preserved or even improved, as OPL/G-OPL removes spurious or distracting features that hinder generalization.

iv. Modularity and scalability. These projection layers are lightweight, differentiable, and easily pluggable into existing models, requiring no special training pipeline or infrastructure.

Toward ethical and interpretable AI systems. By integrating ethical constraints directly into model design, we take a proactive stance in the development of responsible AI. Our approach can be seen as a form of architectural regularization

that aligns representation learning with both performance and ethical objectives. Unlike post-hoc privacy filters or external audits, this internalized design philosophy enables continuous, scalable deployment of AI systems that are privacy-aware by construction.

B. Why Privacy is Protected?

Our method protects privacy at the representation level, which is where leakage typically occurs in modern VAD systems. Even when raw pixels contain faces, state-of-the-art privacy attacks do not require access to raw frames. They operate by extracting and inverting intermediate activations of deployed models (*e.g.*, via model inversion, feature reconstruction, membership inference, or identity retrieval). In deployed cloud or edge VAD systems, it is common that: (i) raw data remain local, (ii) only intermediate features or anomaly scores are transmitted upward. Thus, the attack surface is the latent feature stream, not the pixels.

G-OPL directly removes face-sensitive subspaces from these latent features. The projection (equation 2) guarantees that downstream layers never observe identity-bearing directions. Consequently: (i) identity cannot be recovered from the features, (ii) feature inversion yields face-neutral reconstructions, and (iii) classifier probes fail to retrieve identities (as shown by the large drop in FPD/ArcFace). To further support this, we computed a latent inversion experiment using a lightweight decoder trained only on projected features. Inversions from raw activations preserve recognizable faces (98.0% on ShanghaiTech), whereas inversions from G-OPL activations produce face-smoothed outputs lacking identity cues (45.3% on ShanghaiTech) while fully retaining pedestrian silhouette and motion patterns. This further demonstrates that identity removal is occurring at the feature level, where privacy leakage occurs. Thus, feature anonymization remains meaningful even when raw frames contain faces, because it prevents representation-level leakage, the primary practical privacy risk when VAD systems operate in networked or distributed environments.

Why removing sensitive features is helpful if inputs are not anonymized? We provide a conceptual explanation and empirical confirmation. *Conceptually:* Feature-based privacy leakage is the primary risk in cloud/edge VAD systems. Removing identity-related feature components prevents attackers from reconstructing identities even if pixel-level data are not accessible. This is aligned with representation-level privacy-preservation literature such as INLP (Ravfogel et al., 2020), DeepObfuscator (Li et al., 2021), and adversarial privacy networks (Mirjalili et al., 2020; Cai et al., 2021). *Empirically:* An identity-probe classifier (ArcFace-based) achieves: identity accuracy before projection = 0.99, after G-OPL projection = 0.01. Additionally, feature inversion from projected embeddings yields face-neutral reconstructions. Thus, even without pixel anonymization, feature anonymization directly prevents realistic privacy attacks.

C. Terminology and Conceptual Clarifications

To ensure clarity and accessibility, we define several core concepts used throughout this work.

Task-relevant features. These are features that directly contribute to solving the primary task, here, VAD. For example, in surveillance video, task-relevant cues may include motion patterns, temporal consistency, or object trajectories that help distinguish normal from abnormal events. Retaining such features is essential for maintaining detection performance.

Subspace projection. We use orthogonal projection to separate and remove specific components (*e.g.*, sensitive information) from learned representations. The assumption is that such information resides in a linearly separable subspace, especially in intermediate neural representations, which are more structured and disentangled than raw inputs.

Task-relevant residual subspace. We define the subspace remaining after projection as the task-relevant (or task-aligned) residual subspace. This space is expected to retain features that are informative for downstream VAD task. The proposed G-OPL/OPL learns a subspace that captures sensitive or task-irrelevant variations (*e.g.*, identity), and removes them by projecting features onto its orthogonal complement. This filtering process enhances both robustness and interpretability by preserving only information aligned with the core detection objective.

Task-aware suppression. This refers to the process of removing sensitive features while preserving information critical to the main task. Our G-OPL module achieves this by guiding the projection basis to align with sensitive attribute directions, such as facial identity, using a cosine similarity loss. This allows the model to explicitly capture and subsequently suppress privacy-relevant components from representation, without harming VAD performance.

Sensitive information. Sensitive information refers to attributes that are irrelevant to the primary task but may raise ethical,

legal, or social concerns if retained or exposed. In visual anomaly detection, this can include biometric cues such as facial identity, age, or gender. While these attributes are not necessary for detecting anomalies, they may be unintentionally encoded in intermediate features, potentially leading to privacy violations or biased outcomes.

Utility-privacy trade-off. This captures the balance between preserving performance on the primary task and minimizing leakage of sensitive information. A key contribution of this work is the introduction of metrics and architectural tools (OPL/G-OPL) that help navigate and optimize this trade-off.

Privacy-aware VAD. Rather than relying on traditional privacy-preserving mechanisms (*e.g.*, encryption or DP), our work focuses on suppressing sensitive content in feature space. This approach aligns with responsible AI practices by reducing representational harms while maintaining utility.

D. Connections to Related Work

Our approach builds upon and unifies insights from several research threads, including subspace projection, adversarial representation learning, disentanglement, fairness, and invariant learning. Below, we contextualize our OPL and G-OPL by drawing connections to these bodies of work and clarifying the novel contributions our method offers.

Video anomaly detection. VAD has been extensively studied, evolving from early approaches based on handcrafted features and statistical models to deep learning-based methods that capture complex spatio-temporal patterns (Luo et al., 2021; Liu et al., 2018; Sultani et al., 2018). Recent models use 3D convolutions (*e.g.*, I3D (Carreira & Zisserman, 2017)) and memory-augmented architectures (Gong et al., 2019) to model temporal dynamics and long-range dependencies in videos. While these methods achieve strong performance, they primarily optimize for detection accuracy, without considering the nature of the information being encoded, such as whether some features are irrelevant to the anomaly task or pose ethical concerns. Our work departs from this trend by explicitly shaping the learned representations to filter out both nuisance and privacy, sensitive features using differentiable projection modules.

Subspace projection and Nullspace filtering. Classical subspace projection techniques have long been used in signal processing and machine learning to isolate informative or invariant components of data. Recent approaches such as Nullspace Projection (Ravfogel et al., 2020; Wang et al., 2021; Kim et al., 2019) and Iterative Nullspace Projection (Ravfogel et al., 2020; Esmaeili et al., 2016) apply similar principles for fairness and debiasing, identifying directions that correlate with protected attributes and projecting them out. Our OPL shares the goal of isolating and filtering out task-irrelevant or sensitive directions but differs in three key ways: (i) it learns these subspaces in a fully differentiable manner, integrated directly into the network; (ii) it uses orthogonal projection rather than hard nulling, allowing smooth removal of signals with controlled rank; and (iii) G-OPL extends this by incorporating guidance through a geometric alignment loss that directs suppression toward specific sensitive features (*e.g.*, facial identity).

Adversarial disentanglement and privacy filtering. Our G-OPL layer is related to adversarial representation learning methods used for fairness (Madras et al., 2018; Elazar & Goldberg, 2018) and privacy (Li et al., 2021), which remove sensitive information by backpropagating gradients from attribute classifiers. However, unlike these methods, G-OPL does not rely on an explicit GRL. Instead, it learns a low-rank subspace and projects features orthogonally to sensitive directions. This projection-based suppression provides a more interpretable and geometrically grounded approach compared to generic encoder-decoder or adversarial frameworks.

Disentangled and compressed representations. Disentangled representation learning methods, such as β-VAE (Higgins et al., 2017) and FactorVAE (Kim & Mnih, 2018), aim to decompose features into statistically independent factors of variation. OPL shares this spirit of structural separation, but with key differences: rather than disentangling latent variables through generative priors, we isolate linearly separable subspaces aligned with nuisance or sensitive attributes. Moreover, OPL is task-aware and explicitly supervised by downstream objectives, unlike unsupervised disentanglement which may not align with task relevance. OPL can also be viewed as a learnable bottleneck, comparable to information bottleneck layers (Tishby et al., 2000; Alemi et al., 2016; Wu et al., 2020a), but with a geometric focus on orthogonality and projection.

Spurious correlation and invariant learning. Efforts to suppress spurious correlations, such as domain adversarial training (Ganin et al., 2016), IRM (Arjovsky et al., 2019), and feature orthogonality (Wang et al., 2022), seek to improve generalization by learning invariant representations. G-OPL supports this goal by enforcing invariance to known nuisance attributes via guided projection. However, unlike IRM-based methods that require environment annotations, or methods that enforce global invariance, our approach performs localized, layerwise suppression that is lightweight and adaptable. This

makes it more scalable and compatible with standard vision architectures and datasets. Our proposed framework integrates principles from projection-based filtering, adversarial privacy learning, and disentangled representation modeling, while introducing key architectural and theoretical advances: (i) differentiable, learnable subspace projection; (ii) orthogonal filtering of sensitive information; and (iii) staged integration across layers. This yields a practical, interpretable, and privacy-aware VAD system with improved utility-privacy trade-offs.

E. Theoretical Motivation: Sensitive Attributes in Linearly Separable Subspaces

In this section, we provide a theoretical basis for the common assumption that sensitive attributes, such as facial identity or gender, are approximately linearly separable in intermediate neural representations. While this is not universally guaranteed, it is supported by prevailing patterns in representation learning.

Assumption E.1 (Structured representation learning)**.** Let $\boldsymbol{x} \in \mathbb{R}^n$ be an input (*e.g.*, an image), and let $f : \mathbb{R}^n \to \mathbb{R}^d$ denote the neural network feature extractor. We assume that the intermediate representation $\boldsymbol{f} = f(\boldsymbol{x})$ is *structured*, meaning that different semantic factors, such as task-relevant and sensitive attributes, are at least partially disentangled in feature space.

This assumption is supported by empirical findings that deep networks tend to organize information along semantically meaningful directions in their latent spaces (Bengio et al., 2013).

Definition E.2 (Linear separability in feature space)**.** A sensitive attribute $s \in \mathcal{S}$ is said to be *linearly separable* in the feature space $\mathbb{R}^d$ if there exists a weight matrix $\boldsymbol{W}_s \in \mathbb{R}^{d\times k}$ (or a vector $\boldsymbol{w}_s$ for binary classification) such that:

$$s = \arg\max_i (\boldsymbol{W}_s^\top \boldsymbol{f})_i, \tag{10}$$

where $(\cdot)_i$ denotes the ith component. Equivalently, a linear classifier on $\boldsymbol{f}$ can reliably predict s. This implies the existence of a sensitive subspace $\mathcal{S}_s = \text{span}(\boldsymbol{W}_s)$ capturing the most predictive directions for s.

Proposition E.3 (Existence of sensitive subspace)**.** *Let $\mathcal{D} = \{(\boldsymbol{x}^{(i)}, s^{(i)})\}_{i=1}^N$ be a dataset with sensitive labels $s^{(i)}$. Suppose:*

i. The feature extractor f is trained for a primary task unrelated to s.

ii. The sensitive attribute s is statistically correlated with the input $\boldsymbol{x}$.

*iii. The mapping $\boldsymbol{f} = f(\boldsymbol{x})$ is continuous and piecewise linear (*e.g.*, due to ReLU activations).*

Then, if a linear classifier $\boldsymbol{w}_s$ achieves low prediction error on s, the projection of $\boldsymbol{f}$ onto the subspace $\mathcal{S}_s = span(\boldsymbol{w}_s)$ preserves sensitive information with high fidelity.

Proof. A linear classifier defines a hyperplane in feature space. Its associated projection:

$$\boldsymbol{f}_s = \frac{\boldsymbol{w}_s \boldsymbol{w}_s^\top}{\|\boldsymbol{w}_s\|^2} \boldsymbol{f} \tag{11}$$

isolates the component of $\boldsymbol{f}$ aligned with s. In the multiclass or multi-attribute case, a projection onto $\text{span}(\boldsymbol{W}_s)$ retains the informative directions.

By the Eckart-Young theorem (Eckart & Young, 1936), such projections provide optimal low-rank approximations for preserving predictive information under Frobenius norm constraints.

Thus, sensitive attributes can be geometrically modeled as linearly encoded subspaces within $\mathbb{R}^d$.

□

Empirical and theoretical justification. Deep networks tend to disentangle factors of variation through hierarchical abstraction (Bengio et al., 2013; Belghazi et al., 2018). Nonlinear entanglements in pixel space often become more linearly separable in intermediate features. This is further supported by linear probing studies, which show that simple classifiers can extract semantic information from latent representations (Alain & Bengio, 2016).

Although exact linear separability is not guaranteed, it is theoretically and empirically reasonable to model sensitive attributes as occupying identifiable subspaces in neural feature spaces. This motivates our use of orthogonal projection to remove sensitive components while preserving task-relevant information, enabling interpretable, efficient, and ethically aligned representations.

F. Beyond a Linear Layer: How OPL/G-OPL Differ from Standard FC Layers

Although our proposed OPL and its guided variant (G-OPL) use a learnable linear transformation internally, they fundamentally diverge from conventional fully connected (FC) layers in terms of objective, structure, and interpretability. Below, we clarify these differences to highlight the unique role played by OPL/G-OPL in enforcing task-aligned and privacy-aware representation learning.

Objective: subspace projection *vs.* feature transformation. A standard FC layer learns arbitrary affine transformations to map features from one space to another, typically followed by a nonlinearity. Its goal is to expand model capacity and enable feature recombination to support the end task (*e.g.*, classification). In contrast, OPL and G-OPL are explicitly designed to enforce geometric structure in the representation space by learning a low-dimensional, orthonormal subspace that captures specific semantic information, either task-irrelevant (OPL) or sensitive (G-OPL). Instead of merely transforming features, these layers act as *filters*, removing unwanted components via orthogonal projection.

Structure: orthonormality and projection geometry. At the heart of OPL is a bias-free linear layer whose weights are interpreted not as a transformation matrix, but as a subspace basis. Specifically, given weights $\boldsymbol{W} \in \mathbb{R}^{d \times k}$, we compute its orthonormal basis $\boldsymbol{Q} \in \mathbb{R}^{d \times k}$ via QR decomposition. This $\boldsymbol{Q}$ defines a k-dimensional subspace within the feature space $\mathbb{R}^d$. The projection matrix (Equation 2) is then applied to the input features $\boldsymbol{f}$, yielding equation 3, which removes the component of $\boldsymbol{f}$ lying in the learned subspace. This geometric operation is fundamentally different from the arbitrary affine transformation performed by an FC layer. The orthonormality constraint and residual-based design confer interpretability, mathematical soundness, and stability.

Learning signal: supervised subspace removal. A key difference between the two layers lies in their training signals. The OPL is trained implicitly through the downstream VAD loss, learning to filter out nuisance directions irrelevant to the task. In contrast, G-OPL receives explicit supervision from a sensitive attribute detector by maximizing the cosine similarity between sensitive attribute embeddings and the projected feature components. This directs the projection subspace to capture sensitive information, enabling its suppression in the residual features. This targeted and interpretable learning dynamic contrasts with standard FC layers, which lack such a disentangling bias.

Interpretability and stability. The outputs of OPL/G-OPL decompose input into two parts: projection $\boldsymbol{f}_{\text{proj}}$ (preserved, task-relevant) and residual $\boldsymbol{f} - \boldsymbol{f}_{\text{proj}}$ (removed, task-irrelevant or sensitive). This clear separation offers transparency into what model retains and discards at each stage, making internal representation more interpretable. Additionally, the orthogonality constraint stabilizes optimization by preventing degenerate subspaces or ill-conditioned weight matrices, an issue common in deep FC layers.

Although OPL and G-OPL are implemented using bias-free linear layers, they are *not conventional FC layers* in disguise. Their subspace-based design, orthogonality constraints, and supervision-driven learning make them principled tools for structured, privacy-aware feature filtering. This distinction is essential for understanding their role in promoting robust, interpretable, and ethically aligned visual anomaly detection.

G. Algorithmic Implementation of OPL and G-OPL

In this section, we present our OPL and its extension, the Guided OPL (G-OPL), both designed to enhance VAD by removing nuisance and sensitive information from intermediate features. OPL dynamically learns to identify and project out task-irrelevant subspaces, thereby improving robustness and interpretability. Building on this, G-OPL uses weak supervision signals, such as face presence, to explicitly suppress privacy-sensitive components like facial identity through a guided projection mechanism. Both algorithms are detailed in Algorithm 1 and 2. These modules form a modular, differentiable, and effective framework that balances detection performance with privacy preservation in real-world surveillance applications.

Algorithm 1 Orthogonal Projection Layer (OPL)

Require: Feature matrix $\boldsymbol{F} \in \mathbb{R}^{B \times d}$, learnable weight matrix $\boldsymbol{W} \in \mathbb{R}^{k \times d}$
Ensure: Projected features $\boldsymbol{F}_{\text{proj}}$, residual features $\boldsymbol{F}_{\text{res}}$
1. Compute orthonormal basis for nuisance subspace
$\boldsymbol{W}^\top = \boldsymbol{Q}\boldsymbol{R}$ {QR decomposition: $\boldsymbol{Q} \in \mathbb{R}^{d \times k}$}
2. Construct projection matrix
$\boldsymbol{P} \leftarrow \boldsymbol{I}_d - \boldsymbol{Q}\boldsymbol{Q}^\top$ {Project onto orthogonal complement}
3. Project and compute residual
$\boldsymbol{F}_{\text{proj}} \leftarrow \boldsymbol{F} \cdot \boldsymbol{P}$
$\boldsymbol{F}_{\text{res}} \leftarrow \boldsymbol{F} - \boldsymbol{F}_{\text{proj}}$
return $\boldsymbol{F}_{\text{proj}}, \boldsymbol{F}_{\text{res}}$

Algorithm 2 Training with Guided OPL (G-OPL)

Require: Feature batch $\boldsymbol{F} \in \mathbb{R}^{B \times d}$, face embedding batch $\boldsymbol{F}_{\text{face}} \in \mathbb{R}^{B \times d}$, face presence mask $\boldsymbol{m} \in \{0, 1\}^B$
Require: Learnable projection matrix basis $\boldsymbol{Q} \in \mathbb{R}^{d \times k}$
Require: Loss $\mathcal{L}_{\text{ori}}$, weights $\lambda_{\text{face}}, \lambda_{\text{orth}}$
Ensure: Total loss $\mathcal{L}_{\text{total}}$
1. Construct projection matrix
$\boldsymbol{P} \leftarrow \boldsymbol{I}_d - \boldsymbol{Q}\boldsymbol{Q}^\top$
2. Project features
$\boldsymbol{F}_{\text{proj}} \leftarrow \boldsymbol{F} \cdot \boldsymbol{P}$
$\boldsymbol{F}_{\text{sens}} \leftarrow \boldsymbol{F} - \boldsymbol{F}_{\text{proj}}$
3. Compute cosine similarity loss (only if face present)
for $i = 1$ to B **do**
  if $\boldsymbol{m}[i] = 1$ **then**
    $\mathcal{L}_{\text{face}}^{(i)} \leftarrow 1 - \cos(\boldsymbol{F}_{\text{face}}[i], \boldsymbol{F}_{\text{sens}}[i])$
  else
    $\mathcal{L}_{\text{face}}^{(i)} \leftarrow 0$
  end if
end for
$\mathcal{L}_{\text{face}} \leftarrow \frac{1}{\sum \boldsymbol{m}} \sum_{i=1}^{B} \mathcal{L}_{\text{face}}^{(i)}$ {Mean over face-present samples}
4. Orthogonality regularization
$\mathcal{L}_{\text{orth}} \leftarrow \left\| \boldsymbol{Q}^\top \boldsymbol{Q} - \boldsymbol{I}_k \right\|_F^2$
5. Final loss
$\mathcal{L}_{\text{task}} \leftarrow \mathcal{L}_{\text{ori}} + \lambda_{\text{face}} \cdot \mathcal{L}_{\text{face}}$
$\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{task}} + \lambda_{\text{orth}} \cdot \mathcal{L}_{\text{orth}}$
return $\mathcal{L}_{\text{total}}$

H. Practical Considerations for Layer Placement

OPL and G-OPL are designed to be lightweight and modular, enabling flexible integration at various depths within standard VAD architectures. However, their effectiveness depends critically on where they are placed, necessitating careful consideration to balance detection accuracy, privacy preservation, and interpretability.

Efficient implementation. OPL and G-OPL are implemented using a simple, bias-free FC layer followed by QR decomposition to compute an orthonormal projection basis. This avoids matrix inversion or eigendecomposition, ensuring stable and efficient training. The absence of bias ensures the learned subspace is linear and passes through the origin, essential for valid orthogonal projection. The design is fully differentiable, GPU-efficient, and batch-parallel, introducing minimal overhead even in deep networks. Moreover, the low-rank nature of the projection reduces memory and computation, making these modules well-suited for real-time and resource-constrained applications such as robotics and surveillance.

Feature-level projection. Applying orthogonal projection directly to raw video inputs is impractical due to their high dimensionality and nonlinear entanglement of factors like pose, lighting, or occlusion. Such factors rarely lie in linear subspaces in pixel space. In contrast, intermediate features learned by neural networks tend to be semantically structured and more linearly separable. Sensitive attributes such as identity, age, or background context are more explicitly represented in these feature spaces. Projecting at this level allows meaningful and interpretable suppression of unwanted information. Thus, we apply OPL and G-OPL at intermediate feature levels, where subspace learning is more stable and effective.

Why QR decomposition? Traditional projection operators often take the form $\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top$, where $\boldsymbol{W}$ is a basis matrix of the nuisance subspace. However, computing the inverse $(\boldsymbol{W}^\top\boldsymbol{W})^{-1}$ is computationally expensive, numerically unstable, and unsuitable for gradient-based optimization when $\boldsymbol{W}$ is ill-conditioned or nearly rank-deficient. In contrast, QR decomposition provides a robust and efficient way to obtain an orthonormal basis $\boldsymbol{Q}$ without matrix inversion. This approach improves numerical stability, yields well-behaved gradients for end-to-end training, and scales gracefully to high-dimensional feature spaces typical of deep convolutional networks.

Why not adversarial learning? Traditional privacy-preserving methods often use adversarial learning with a GRL (Ganin et al., 2016), requiring a sensitive classifier to predict protected attributes. However, these approaches are unstable and introduce additional complexity. In contrast, G-OPL is simpler and more interpretable. Because both $\boldsymbol{f}$ and $\boldsymbol{f}_{\text{face}}$ are extracted from the same backbone, they *lie in a shared feature space*. This enables *a direct comparison* without needing a classifier or adversarial setup. The cosine loss provides a stable and intuitive training signal, reducing sensitive leakage while preserving task-relevant features.

Optimal layer placement. Early network layers capture low-level visual cues (*e.g.*, texture, color, edges) that often correlate with sensitive attributes like facial identity or clothing style. Applying G-OPL at these stages enables targeted suppression of private information before it propagates deeper. The cosine similarity loss directly aligns the projection basis with sensitive features, allowing precise removal without adversarial gradient reversal. In contrast, deeper layers encode high-level semantic information such as object behavior, temporal dynamics, and scene context. At this stage, the standard OPL effectively removes residual nuisance factors like lighting variations, background clutter, or camera motion, improving anomaly localization and robustness by filtering out spurious correlations.

To systematically evaluate these effects, we experiment with multiple configurations involving single or stacked OPL and G-OPL layers at different depths. We analyze their impact on VAD performance, privacy leakage, and interpretability. Our findings reveal key trade-offs and offer practical guidelines for ethically aligned and effective VAD deployment.

Data flow. Our framework enforces privacy by filtering sensitive information before data leaves the device. Raw video frames are first captured on-device and encoded into feature representations using a lightweight backbone. OPL or G-OPL is then applied immediately to suppress sensitive attributes, after which only the filtered features are transmitted to the server for VAD. In this pipeline, raw pixels never leave the device, ensuring that sensitive information is removed at the feature level prior to communication and providing strong practical privacy protection.

When on-device computation is not available, OPL/G-OPL can alternatively be applied on the server as an early-stage processing step. While this still reduces leakage in learned representations, it does not prevent raw data exposure. Therefore, client-side filtering is the recommended deployment setting for privacy-sensitive applications.

I. Why Stack Multiple OPL/G-OPL Layers?

Stacking multiple OPL/G-OPL offers a principled and effective strategy for progressively filtering out different forms of undesirable information, ranging from private attributes to task-irrelevant nuisances, across network depth. This design choice is not merely architectural convenience but reflects a structured approach to disentangling and regulating feature representations at multiple levels of abstraction.

Progressive removal of structured information. Neural networks encode features hierarchically: early layers capture fine-grained, low-level details (*e.g.*, texture, color), while deeper layers encode semantic, task-aligned abstractions (*e.g.*, motion patterns, object interactions). Sensitive or irrelevant information may appear at different depths and in different forms. A single projection may be insufficient to fully remove them, especially if they re-emerge or evolve in deeper layers. By stacking OPL/G-OPL modules, we allow the network to iteratively refine and purify representations, removing unwanted components as they reappear or become linearly separable at different depths.

Interpretation as multi-stage subspace filtering. Each projection layer defines a learned orthogonal subspace corresponding to attributes that should be removed. Stacking multiple such layers can be interpreted as *a multi-stage subspace filtering process*:

$$\boldsymbol{f}^{(l+1)} = (\boldsymbol{I}_d - \boldsymbol{Q}^{(l)}\boldsymbol{Q}^{(l)\top})\boldsymbol{f}^{(l)}, \tag{12}$$

where each $\boldsymbol{Q}^{(l)}$ represents a learned sensitive or nuisance subspace at layer l. This recursive filtering ensures that different projections can specialize in removing distinct aspects of the signal, *e.g.*, identity in early layers, background clutter or temporal bias in later ones.

Enhanced expressiveness and modularity. From a learning perspective, stacking increases the expressiveness and flexibility of the projection mechanism. Rather than relying on a single, global subspace to capture all unwanted variation, the model learns a series of localized, lower-rank subspaces adapted to the representation space of each layer. This reduces optimization difficulty, improves stability, and allows each OPL/G-OPL to focus on a narrower, more interpretable source of variation.

Improved trade-offs between utility and privacy. Our experiments show that strategically stacking G-OPL/OPL at selected depths leads to improved privacy-utility trade-offs. G-OPLs placed early suppress high-sensitivity attributes (*e.g.*, face identity), while OPLs later in the network remove nuisance features that could degrade VAD performance (*e.g.*, lighting shifts or background dynamics). This staged design helps retain task-relevant signals while minimizing privacy leakage and interpretability loss.

Stacking G-OPL/OPL is a principled architectural strategy grounded in the hierarchical nature of learned features. It enables progressive and interpretable removal of harmful signals, improves privacy-utility balance, and supports modular integration in varied deployment settings.

J. Discussions on Privacy-Aware Evaluation Metrics

Whether evaluation metrics consider video length? Our new metric designs do consider video length. SSC is frame-independent because it operates on the distribution of face embeddings, not on temporal sequences. Longer videos merely yield more samples, reducing variance. ARD compares anomaly-score distributions before/after projection. For very long videos, KDE-based distributions converge smoothly but do not bias the metric. To verify this, we conducted an additional experiment where videos were subsampled at rates $1/2$, $1/4$, and $1/8$ on ShanghaiTech. ARD remained stable within a deviation of < 0.015, demonstrating length-invariance. PD/FPD uses classifier-probe accuracy on feature batches, not temporal alignment, so video length only affects sample count. Experimental subsampling did not change the qualitative decay curve or FPD values.

SSC metric dependency on attribute embeddings. SSC does not rely on a specific embedding type, but on the availability of an attribute-related subspace. The formulation remains unchanged, as SSC measures the alignment between features and this subspace.

When pretrained attribute embeddings are unavailable, the subspace can be estimated using alternative signals, including: (i) weak supervision (*e.g.*, simple attribute classifiers such as face presence), (ii) pretrained feature spaces (*e.g.*, intermediate representations from general models), or (iii) unsupervised discovery (*e.g.*, principal components capturing dominant variations). SSC only requires *a coarse directional bias* toward the sensitive attribute, rather than a precise embedding.

Method	ShanghaiTech	UCF-Crime
RTFM (ICCV'21)	0.72	0.68
+ **G-OPL**	**0.01**	**0.23**
MGFN (AAAI'23)	0.98	0.79
+ **G-OPL**	**0.68**	**0.49**
SPAct (CVPR'22)	0.68	0.66
+ **G-OPL**	**0.55**	**0.53**
TeD-SPAD (ICCV'23)	0.66	0.60
+ **G-OPL**	**0.56**	**0.52**
EGO (ICLR'25)	0.72	0.78
+ **G-OPL**	**0.09**	**0.12**
TEVAD (CVPRW'23)	0.95	0.76
+ **G-OPL**	**0.04**	**0.11**

Table 9. Face-presence detection (PD/FPD, ↓) across multiple VAD methods on ShanghaiTech and UCF-Crime. All results are obtained using the same linear probe architecture and training protocol.

Therefore, these variants do not alter the SSC definition, but affect how the attribute subspace is instantiated, making the metric broadly applicable beyond attributes with dedicated embeddings.

On the specificity of face-based privacy metrics. While SSC focuses on face-sensitive subspaces, the metric design is not face-specific. It only assumes access to a sensitive attribute embedding $\boldsymbol{f}_{\text{attr}}$. Thus SSC generalizes to: SSC-gait, SSC-clothing, SSC-age, SSC-vehicle-identity, or any sensitive cue for which weak supervision exists. To demonstrate generality, we ran an auxiliary experiment using clothing-color embeddings (extracted via a lightweight appearance encoder). When treating clothing color as the sensitive attribute, OPL/G-OPL successfully removed this signal with SSC-color = 0.84 while maintaining anomaly AUC within 0.4 points of the baseline. This verifies the metric's broader applicability.

Clarification on the applicability of SSC/ARD *vs.* PD/FPD. SSC and ARD require a projection module because their definitions depend on the projected features ($\boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{f}$ or $(\boldsymbol{I} - \boldsymbol{Q}\boldsymbol{Q}^\top)\boldsymbol{f}$). Therefore, SSC/ARD can be computed only for models that implement a projection (*e.g.*, OPL/G-OPL, or baselines augmented with OPL). PD and FPD require only the model's internal features and a lightweight presence probe. They do not require any projection and can be computed for any VAD architecture. Thus, SSC/ARD are diagnostic of projection-based methods, whereas PD/FPD are baseline-compatible.

Why these metrics remain broadly useful. PD/FPD provide a general way to measure leakage of a sensitive presence signal and apply to future VAD or video-encoder model. SSC/ARD are diagnostic tools for any method, current or future, that implements sensitive-subspace removal. Therefore these metrics are not specific to our method, but align with standard practice in privacy-preserving representation learning.

PD/FPD applied to prior VAD methods. We computed PD/FPD on RTFM, MGFN, SPAct, TeD-SPAD, EGO and TEVAD using the same probe architecture and training protocol (linear probe trained to predict face-presence).

Table 9 summarizes the results. Most existing methods exhibit high PD/FPD (strong leakage), SPAct and TeD-SPAD show reduced leakage, and adding G-OPL further reduces PD/FPD while preserving AUC. This demonstrates that PD/FPD apply fairly to all baselines.

K. Detailed Setups and Configurations

To evaluate the impact of layer placement, we conduct extensive experiments using two representative VAD models: RTFM and MGFN. Our study spans five widely adopted datasets, MSAD, ShanghaiTech (ShT), UCF-Crime (UCF), CUHK Avenue (CUHK), and UCSD Ped2 (Ped2), to ensure broad applicability and robustness.

We explore two types of face representations to guide the G-OPL module: *(i) Detected faces*: For datasets with discernible facial content (ShT, UCF, CUHK), we use RetinaFace to extract faces and generate corresponding face videos. These are processed alongside the original videos using a Kinetics-pretrained I3D model to obtain face embeddings. *(ii) Generated*

Dataset	RTFM								MGFN							
	G-OPL	OPL	λ_{orth}	λ_{face}	batch-size	max-epoch	AUC	AP	G-OPL	OPL	λ_{orth}	λ_{face}	batch-size	max-epoch	AUC	AP
MSAD-**G-OPL/OPL**	1	0	1e−4	1.0	16	100	88.0	70.9	1	0	1e−2	1e−3	16	100	84.0	65.8
ShT-**G-OPL/OPL**	1	0	2.0	3.0	16	100	97.3	74.7	1	0	1e−3	0.3	16	100	83.7	42.0
ShT-**G-OPL/OPL**†	1	0	1	0	16	100	97.2	74.6	1	0	3.0	0.3	16	100	89.5	41.9
UCF-**G-OPL/OPL**	1	0	1e−2	3.0	16	100	78.3	30.9	1	0	0.1	0.1	16	100	83.3	15.2
UCF-**G-OPL/OPL**†	1	0	1e−2	3.0	16	100	74.1	30.1	1	0	0.1	0.1	16	100	80.9	14.8
CUHK-**G-OPL/OPL**	1	0	0.1	2.0	4	100	84.9	66.2	3	0	1e−5	1.0	3	100	70.8	40.5
CUHK-**G-OPL/OPL**†	1	0	1e−5	0.0	4	100	83.9	65.7	1	0	2.0	0.5	3	100	69.1	43.5
Ped2-**G-OPL/OPL**	5	0	0.1	0.5	2	200	89.6	75.0	1	0	5.0	0.5	2	200	93.9	93.6

Table 10. Optimal hyperparameters for RTFM and MGFN across all datasets. † indicates configurations using detected faces. G-OPL and OPL denote the presence of the respective projection layers; λ_{orth} and λ_{face} are the weights for orthogonality and face alignment losses, respectively. Batch size, training epochs, and resulting AUC and AP scores are also reported.

faces: To simulate face guidance without relying on in-dataset detection, we create synthetic face videos from the Georgia Tech Face Database, comprising 50 subjects with 15 images each. These are converted into videos and processed in the same way to serve as a generalized face prior.

For the three datasets with detectable faces, we directly compare the effects of using detected versus generated face signals to guide G-OPL.

Placement of G-OPL/OPL modules. In both RTFM and MGFN architectures, the integration of G-OPL and OPL follows well-defined constraints and placement rules based on the structure of each model. For RTFM, there are six fixed insertion points for projection modules. The first G-OPL, if present, is always inserted before the aggregate module. Subsequent G-OPL and OPL modules can then be placed within the aggregate block, after each of its five convolutional layers. These positions are denoted as C1 through C5, where C1 refers to the insertion point after the first convolutional layer, C2 after the second, and so on up to C5. The total number of inserted modules should not exceed six, and OPLs can only be placed after at least one G-OPL has been inserted. For MGFN, there are five predefined insertion points corresponding to the sequence of backbone and convolutional layers. These are labeled as B1, C1, B2, C2, and B3. Here, B1 is the position after the first Transformer block, C1 after the first convolutional layer, B2 after the second Transformer block, C2 after the second convolutional layer, and B3 after the third Transformer block. As in RTFM, the total number of projection layers should not exceed the available slots, and OPL modules can only follow G-OPL modules. This consistent naming convention (*e.g.*, C1, B1) allows clear identification of module locations and enables the structured evaluation of different G-OPL and OPL configurations across experiments.

In the default setting used in our experiments, we use a single G-OPL (guided by face presence), which already achieves strong performance and low leakage across most datasets. Additional OPL/G-OPL modules can be optionally stacked when stronger suppression is needed (*e.g.*, low-resolution datasets). Further analyses of module placement are provided in Sec. L (second part) and M.

In practice: A single default configuration (G1O0 / G1O1, $k = 4$, λ_{face}=10^{-3}, λ_{orth}=10^{-3}) works across all five datasets with: < 1 point average AUC drop from per-dataset tuning, and nearly identical privacy suppression. Optimal hyperparameter values for each configuration are summarized in Table 10.

L. Additional Results and Evaluations

Low-face-visibility scenarios. We provide hyperparameter tuning on UCSD Ped2 in Tables 11 and 12.

The best performance is achieved at $k_{\text{OPL}} = 16$ and $k_{\text{G-OPL}} = 8$, indicating that a moderate representation capacity is required to effectively model both nuisance and face-sensitive subspaces. Smaller values lead to underfitting, while larger values degrade performance due to over-parameterization and reduced generalization. The optimal $k_{\text{G-OPL}}$ is consistently smaller than k_{OPL}, supporting the hypothesis that identity-related features lie in a relatively low-dimensional subspace. The performance peaks at $\lambda_{\text{face}} = 10^{-3}$ and $\lambda_{\text{orth}} = 10^{-3}$. Smaller values result in insufficient suppression of identity information, while larger values lead to over-suppression and loss of discriminative cues. While the performance variation is moderate, the consistent trends demonstrate that both subspace capacity and regularization should be carefully balanced, especially in low-resolution scenarios such as UCSD Ped2.

Evaluating the effects of λ_{orth} and λ_{face}. We introduce two critical hyperparameters to modulate the strength of our

$k_{\text{OPL}} \backslash k_{\text{G-OPL}}$	4	8	16	32	64	128
4	88.7	89.1	88.9	88.4	87.6	86.9
8	89.0	89.4	89.2	88.8	88.1	87.3
16	89.3	**89.6**	89.5	89.1	88.5	87.7
32	89.0	89.3	89.2	88.9	88.2	87.5
64	88.2	88.6	88.5	88.1	87.4	86.6
128	87.6	88.0	87.8	87.3	86.7	85.9

Table 11. Hyperparameter tuning of OPL and G-OPL on UCSD Ped2. The table reports AUC (%). The best performance is obtained at $k_{\text{OPL}} = 16$ and $k_{\text{G-OPL}} = 8$.

$\lambda_{\text{face}} \backslash \lambda_{\text{orth}}$	10^{-4}	10^{-3}	10^{-2}
10^{-4}	89.1	89.3	88.9
10^{-3}	89.3	**89.6**	89.2
10^{-2}	88.6	88.9	88.1

Table 12. Effect of regularization hyperparameters on UCSD Ped2. The table reports AUC (%). The best performance is achieved at $\lambda_{\text{face}} = 10^{-3}$ and $\lambda_{\text{orth}} = 10^{-3}$.

projection-based objectives: (i) λ_{orth}, which controls the orthogonality regularization in OPL/G-OPL to stabilize disentanglement, and (ii) λ_{face}, which weights the face-guided suppression loss in G-OPL to enforce privacy preservation.

To understand their influence, we conduct a comprehensive grid search over a wide range of candidate values: 0.0, $1\text{e}{-5}$, $1\text{e}{-4}$, $1\text{e}{-3}$, $1\text{e}{-2}$, 0.1, 0.2, 0.3, 0.5, 1.0, 2.0, 3.0, 5.0 for both λ_{orth} and λ_{face}, resulting in 169 unique combinations. Each setting is evaluated across five benchmark VAD datasets using both RTFM and MGFN architectures.

The optimal configurations, those that yield the best balance between VAD performance and ethical alignment, are summarized in Table 10. Notably, the ideal weights vary across datasets and models, suggesting that sensitivity to orthogonality and privacy constraints is task- and architecture-dependent. For instance, RTFM often benefits from stronger privacy guidance (*e.g.*, λ_{face}=3.0 on UCF), while MGFN is more responsive to orthogonality regularization (*e.g.*, λ_{orth}=5.0 on Ped2). This analysis highlights the importance of tuning these hyperparameters to adaptively control the trade-off between VAD accuracy and the ethical goals of interpretability and privacy.

On sensitivity to hyperparameters and configuration choices. OPL/G-OPL introduces configuration choices; however, the method is not as brittle as the initial figures may suggest. First, hyperparameter sensitivity is strongly bounded. As shown in Fig. 3(b-c), both λ_{face} and λ_{orth} exhibit stable performance in a wide range $[10^{-4}, 10^{-2}]$. The variation outside this range is expected, since very large values aggressively erase feature content. In practice, the same hyperparameters transfer across all five datasets without re-tuning. Second, layer placement is not fragile. Our extended ablations (above) show that: (i) an early G-OPL (post-encoder) is consistently optimal or near-optimal, (ii) additional OPL layers provide diminishing returns but do not harm performance if kept shallow, and (iii) the method generalizes well to unseen domains when using a fixed, default G1O0 / G1O1 placement. A new cross-domain experiment (training on ShanghaiTech, testing on CUHK Avenue, without retuning) shows AUC degradation of only 0.7 points with default hyperparameters. This supports generalization without tuning.

Unreliable/Missing face supervision. G-OPL is inherently robust: The guidance loss (Equation 4) is only applied when faces are detected. If detection fails: the model defaults to OPL behavior, and projection remains well-defined via learned subspace. We randomly drop face supervision during training, Table 13 reports the results. Performance degrades gracefully, confirming robustness under occlusion or detector failure.

MSAD (blurred faces). On the MSAD dataset with blurred faces, the results show a clear and consistent trade-off between utility and privacy. Both backbones (EGO and TEVAD) retain substantial identity information at baseline, as evidenced by high facial-presence detection (FPD $\approx$ 0.58-0.60) and strong ArcFace identity retrieval ($\approx$ 0.70+), despite the presence of blurred faces. Introducing OPL leads to modest but reliable gains in VAD (*e.g.*, +0.8 AUC for EGO, +0.6 for TEVAD), while significantly reducing identity leakage. In particular, FPD drops to $\sim$0.39-0.41 and ArcFace scores are nearly halved, indicating effective suppression of nuisance identity cues without sacrificing performance. G-OPL further strengthens

Drop rate	AUC↑	FPD↓
0% (**G-OPL**)	85.4	0.34
30%	85.1	0.39
60%	84.8	0.47
100% (**OPL**)	85.9	0.78

Table 13. Robustness to unreliable or missing face supervision. Face guidance is randomly dropped during training at different rates. G-OPL degrades gracefully, reverting toward OPL behavior as supervision is reduced.

Model	AUC↑	SSC↑	ARD↓	FPD↓	Arc↓
EGO	86.9	–	–	0.60	0.72
EGO + **OPL**	87.7	0.32	0.04	0.41	0.34
EGO + **G-OPL**	87.5	0.73	0.04	0.18	0.11
TEVAD	82.1	–	–	0.58	0.70
TEVAD + **OPL**	82.7	0.33	0.04	0.39	0.32
TEVAD + **G-OPL**	82.5	0.68	0.03	0.17	0.10

Table 14. Results on MSAD (blurred faces). ↑/↓ indicate higher/lower is better.

this effect. While maintaining nearly identical AUC to OPL, it dramatically improves privacy metrics: SSC increases substantially (up to 0.73), while FPD and ArcFace drop to as low as 0.17-0.18 and 0.10-0.11, respectively. This indicates strong removal of sensitive subspace information and near-complete suppression of identity signals, even under already blurred conditions.

Why G-OPL improves performance on blurred-face data? The improvement arises because face blurring does not eliminate latent identity traces. I3D/SwinT features extracted from blurred faces still encode: head location, shape of the blurred region, motion patterns around the face, residual low-frequency structure, correlations between head region and action type. These weak traces remain predictive of identity (verified by ArcFace-based probes achieving 14% retrieval even on blurred MSAD faces). G-OPL removes these residual identity cues, reducing spurious correlations. This produces more stable representations and yields the observed performance gains.

Why G-OPL sometimes underperforms OPL? Two factors explain the discrepancy observed in Tables 2 and 3: (a) Guided suppression is intentionally more aggressive. G-OPL prioritizes removal of face-related directions, which can slightly reduce performance when: (i) a dataset has low-quality faces or (ii) anomalies subtly correlate with identity-like features (*e.g.*, pose, clothing or local appearance). (b) G-OPL does not harm performance when the base model is strong. For RTFM (I3D), G-OPL improves overall AUC (Table 2). For MGFN, which is more sensitive to feature manipulation due to its contrastive design, OPL alone is sometimes slightly better. This is a design trade-off: G-OPL favors privacy; OPL maximizes pure accuracy.

Why performance gains vary across datasets? We emphasize that G-OPL and OPL optimize fundamentally different projection objectives and thus exhibit complementary, rather than hierarchical, performance profiles. (i) OPL removes a broad, data-driven nuisance subspace spanning high-variance but task-irrelevant directions (background motion, scene clutter, illumination fluctuations, compression artifacts, *etc.*). (ii) G-OPL, in contrast, removes a much narrower, semantics-guided subspace explicitly associated with biometric cues (*e.g.*, faces), using weak face-presence supervision. Importantly, VAD datasets vary dramatically in spatial resolution, subject scale, face visibility (occlusion, blur, lighting), compression noise, and camera viewpoint stability. In datasets where faces are severely blurred, partially visible, very small, or inconsistently detected, conditions common in VAD (*e.g.*, ShanghaiTech, UCF-Crime, *etc.*), the learned face-sensitive subspace is necessarily smaller and less expressive. Under such conditions, G-OPL intentionally removes less feature content than OPL. As a result, OPL may achieve slightly higher AUC because it suppresses a broader pool of nuisance variation. Crucially, this behavior reflects the design of the two modules: (i) OPL prioritizes maximal nuisance suppression to improve pure utility. (ii) G-OPL prioritizes sensitive subspace suppression (*e.g.*, faces) and therefore trades a small amount of utility for substantially stronger privacy. Across all four benchmarks and six architectures, including EGO (ICLR 2025), TeD-SPAD (ICCV 2023), TEVAD (CVPRW 2023), and SPAct (CVPR 2022), G-OPL (i) consistently achieves the lowest privacy leakage, (ii) maintains AUC within a small margin of OPL, and (iii) never disrupts anomaly separability. As shown in Table

Dataset	Anomalous	Normal
ShanghaiTech	12.5%	38.2%
CUHK Avenue	24.1%	51.4%
UCSD Ped2	1.7%	3.1%
MSAD (blurred)	0%	0%
UCF-Crime	48.3%	77.6%

Table 15. Percentage of frames containing faces in anomalous *vs.* normal segments across datasets. Face presence is annotated using a standard face detector with manual correction.

10, the G1O0 configuration emerges as the most robust and generally optimal setting across all datasets.

Gains differ because datasets have different facial-content distributions and different relationships between faces and anomaly labels. We now provide a systematic analysis.

(a) Percentage of anomalies containing faces. We annotated face presence in anomaly segments using a standard face detector (with manual correction, see Table 15). For ShanghaiTech and CUHK Avenue, faces are substantially more common in normal frames. Thus, models may inadvertently treat low facial presence (or partial occlusion) as an anomaly cue. Removing facial components reduces this spurious correlation, which explains the large AUC gains (+8.4%, +3.5%). In contrast, UCSD Ped2 has almost no faces, making suppression unnecessary; hence OPL and G-OPL both act mainly as regularizers, yielding smaller but still positive gains.

(b) Correlation between faces and anomaly labels. We computed mutual information (MI) between: $\mathbf{1}_{\text{face}}$, and $\mathbf{1}_{\text{anomaly}}$. For ShanghaiTech and CUHK Avenue: MI < 0.02. Thus, facial information is essentially orthogonal to anomaly semantics. Removing it reduces spurious co-occurrence. For UCF-Crime: MI = 0.12, the highest among the datasets, which explains why G-OPL does not yield the largest improvement there: faces sometimes co-occur with anomaly categories involving close-range person interactions.

(c) For MSAD: what is G-OPL removing from pre-blurred features? Even though faces are blurred at the pixel level, the pre-extracted I3D/Swin features still contain: the spatial mask of the blurred region, head motion trajectories, low-frequency cues around the blurred patch, contextual correlations between presence of a blurred blob and normal behavior (Table 14). We verified this by training an identity-retrieval probe on blurred MSAD features. Despite blur, the probe achieved 14% rank-1 accuracy, far above chance.

VISPR evaluation. We follow the same evaluation principle used in SPAct and TeD-SPAD: privacy is measured by training a new attribute classifier on anonymized features to quantify attribute recoverability. Our implementation uses a simplified VISPR probe, the core protocol, anonymize $\rightarrow$ train classifier $\rightarrow$ evaluate recoverability, is identical. Thus, the reported VISPR results measure the recoverability of privacy attributes from the anonymized feature space, consistent with SPAct/TeD-SPAD.

We tested G-OPL under the VISPR privacy attribute classifier used in SPAct/TeD-SPAD (predicting 9 common privacy attributes). Results for ShanghaiTech: VISPR accuracy (baseline RTFM) = 0.63, VISPR accuracy (RTFM+G-OPL) = 0.27. This indicates substantial privacy improvement aligned with VISPR criteria. While G-OPL focuses on the face-aligned subspace, many VISPR attributes (*e.g.*, age, gender, expression, eyewear) are strongly correlated with facial features. Removing this subspace therefore reduces privacy leakage for a broad range of attributes without requiring attribute-specific supervision. Importantly, G-OPL operates on the backbone features extracted by VAD models (*e.g.*, RTFM, SPAct, TeD-SPAD) and does not require modifying the model or its classifier. The ResNet-based attribute classifiers used in SPAct/TeD-SPAD serve only for evaluation, and benefit from the reduced identity information in the G-OPL-processed features.

Table 16 shows that face-conditioned attributes (*e.g.*, age, gender, expression) experience large drops, approaching random chance. Non-face-conditioned attributes, such as clothing or background objects, remain largely unaffected. This demonstrates that G-OPL selectively removes identity-related information without globally suppressing all features, supporting our conceptual argument.

Why facial-subspace removal reduces all VISPR attributes. Most VISPR attributes (*e.g.*, age, gender, race, glasses, facial hair, makeup, *etc.*) are face-conditioned. When G-OPL removes the face-aligned latent subspace, the external VISPR

VISPR attribute	Face-conditioned?	Baseline	+ **G-OPL**	Notes
Age	Yes	0.70	**0.30**	Strong drop due to facial cues removal
Gender	Yes	0.68	**0.25**	Near chance after G-OPL
Expression	Yes	0.65	**0.27**	Facial subspace suppressed
Eyewear	Yes	0.60	**0.28**	Strong reduction
Facial hair	Yes	0.62	**0.26**	Face-aligned features removed
Makeup	Yes	0.55	**0.24**	Approaches chance
Race	Yes	0.57	**0.25**	Almost random, face-dependent
Clothing	No	0.61	**0.59**	Minimal effect, non-face attribute
Background objects	No	0.58	**0.57**	Largely unaffected

Table 16. Effect of G-OPL on VISPR attributes. Attributes are grouped into face-conditioned and non-face-conditioned categories.

Method	AUC↑	SSC ↑	FPD ↓
Baseline	84.7	–	0.91
+ **OPL**	85.9	–	0.78
+ **G-OPL**	85.4	0.88	0.34

Table 17. Results on CHAD (high-resolution / unconstrained setting) using RTFM (I3D) as the backbone. G-OPL significantly reduces face-presence leakage (FPD) while maintaining VAD performance (AUC) and achieving high SSC.

classifier loses access to the features required to predict these attributes. Because VAD does not rely on facial semantics, G-OPL can remove a larger portion of the facial subspace without hurting utility, causing the VISPR probe's AP to approach chance for many face-conditioned attributes (the suppression can be more aggressive than in action-recognition anonymizers, resulting in many VISPR attributes dropping to near-chance mAP). This behavior is consistent with latent-space suppression of facial cues and does not imply general attribute removal.

Comparison to (Wu et al., 2020b). Wu et al. (2020b) operate in an action-recognition setting, where facial and head cues contribute to task utility, creating a privacy-utility trade-off that limits the extent of facial information suppression. In contrast, VAD does not rely on facial semantics, allowing G-OPL to remove a larger portion of the face-aligned subspace. This explains why our VISPR mAP reduction can be stronger, even without attribute-specific supervision.

High-resolution / Unconstrained datasets. G-OPL operates on intermediate feature representations, making it resolution-agnostic. To evaluate privacy leakage under realistic conditions, we extend our experiments to CHAD (Danesh Pazho et al., 2023) in Table 17. We follow the same protocol (Sec. 4.1) and integrate G-OPL into RTFM (I3D). FPD drops significantly (0.91→0.34), indicating strong suppression of identity cues; SSC is high (0.88), confirming effective capture of face-sensitive subspaces; detection performance is preserved (AUC drop $< 0.5\%$ *vs.* OPL). This shows that G-OPL is effective when high-resolution identity signals are present, where privacy risks are most critical.

Latency and memory overhead. Table 18 summarizes the results. G-OPL is well-suited for real-time and edge deployment, introducing only minimal overhead: latency increases by approximately 7%, memory usage by less than 3%, and no additional inference cost is incurred for face detection.

Sensitivity to anomalies & failure analysis. While suppressing facial cues may remove weakly correlated signals such as facial expressions (*e.g.*, aggression-related cues), we do not observe systematic degradation in practice. For human-centric

Method	FPS ↑	Latency (ms) ↓	Memory (MB) ↓
RTFM	142	7.0	812
RTFM + **OPL**	137	7.3	826
RTFM + **G-OPL**	134	7.5	834

Table 18. Latency and memory overhead of OPL/G-OPL on RTFM. OPL/G-OPL introduces minimal computational overhead while maintaining real-time performance.

Setting	AUC (%)	ARD ↓	FPD ↓
Fixed parameters	85.4	0.29	0.43
Adaptive $k_{\text{G-OPL}}$	85.8	0.21	0.36
Adaptive λ_{face}	**86.1**	**0.18**	0.38

Table 19. Adaptive hyperparameter tuning on validation ARD. We compare fixed parameters with adaptive strategies for $k_{\text{G-OPL}}$ and λ_{face}. Adaptive tuning improves AUC while reducing ARD and FPD.

Method	Assault		Explosion		Fighting		Fire		Obj. Fall		People Fall		Robbery		Shooting		Traffic Acc.		Vandalism		Water Inc.		**Overall**	
	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP
MSAD (B1)	71.3	69.4	61.8	73.0	87.8	92.8	81.0	92.9	94.3	96.5	45.9	45.0	65.1	81.1	82.7	89.1	64.2	55.2	90.7	86.4	68.7	91.9	86.2	68.3
MSAD (B2)	61.6	67.6	45.6	55.2	66.2	75.3	67.6	82.4	85.8	84.5	44.3	39.2	65.1	78.0	57.7	69.7	56.2	48.9	76.5	71.3	95.7	97.1	85.1	59.0
MSAD (B3)	53.7	56.0	59.2	68.7	84.6	90.8	81.0	91.2	91.3	95.3	48.6	46.8	63.5	78.5	90.9	94.0	71.4	63.9	85.3	81.5	97.7	99.4	84.3	62.8
MSAD (B1B2B3)	55.7	59.7	41.6	56.6	86.7	91.9	73.8	89.0	89.5	93.9	50.8	46.3	63.4	79.6	87.6	89.4	65.3	56.2	88.2	83.9	85.3	96.7	82.8	59.7
MSAD (C1)	56.8	59.9	47.0	60.7	75.6	87.0	75.7	87.1	90.5	94.7	49.2	48.4	64.8	79.6	77.9	88.1	66.7	56.7	80.5	73.7	90.1	97.8	82.9	58.6
MSAD (C2)	66.6	71.7	57.6	65.5	94.2	96.4	78.5	91.2	90.4	94.4	52.1	50.6	66.5	82.1	88.3	90.3	67.9	58.3	70.7	69.2	82.0	96.1	85.3	65.8
MSAD (C1C2)	56.4	59.2	53.9	66.5	81.4	89.4	80.3	90.1	91.6	95.5	49.7	47.0	69.6	83.7	89.1	91.7	67.7	56.0	76.6	75.0	83.0	96.4	84.1	62.5
MSAD (B1C1)	58.5	57.4	50.4	64.0	74.2	87.1	85.6	92.5	90.4	94.6	45.9	44.2	66.4	80.8	78.8	86.0	62.6	50.7	84.2	80.8	99.2	99.8	84.8	63.8
MSAD (B2C2)	46.9	52.0	58.8	69.5	77.6	88.4	83.4	91.7	90.2	94.6	53.8	47.5	68.9	83.1	74.5	83.9	58.8	48.2	83.9	74.9	84.5	96.7	83.6	61.2
MSAD (ALL)	59.5	57.4	53.0	57.8	79.7	86.9	66.7	84.0	87.2	85.6	53.7	48.4	68.0	81.4	77.3	81.9	63.6	50.7	84.2	80.8	99.2	99.8	84.8	63.8

Table 20. Anomaly-wise performance of the MGFN model using only OPL configurations on the MSAD dataset.

anomalies such as vandalism, the model primarily relies on motion patterns and contextual interactions, which are preserved under G-OPL. Empirically, performance on such categories remains comparable to the baseline, and overall performance is stable across datasets, indicating that task-relevant signals are not adversely affected. Any observed variations are largely dataset-specific and arise in cases where task-relevant features are partially entangled with the suppressed subspace (*e.g.*, spatial proximity to the face); however, such cases are rare and localized and do not dominate the model behavior.

We further identify three primary failure modes: over-suppression under large k or high λ_{face}, noisy face supervision, and identity-anomaly correlation due to entanglement between sensitive and task-relevant components. In practice, using a small rank (2-5% of the feature dimension) consistently yields stable performance, while ARD and PD/FPD provide reliable signals for tuning and help avoid performance collapse. As shown in Table 19, adaptive tuning based on validation ARD improves AUC while maintaining strong privacy, further validating the effectiveness of these metrics for practical hyperparameter selection.

M. Layer Placement: Practical Guidelines and Empirical Insights

We conduct comprehensive evaluations of OPL and G-OPL layer placements across five benchmark VAD datasets, aiming to offer concrete and practical insights into how these components should be integrated into existing architectures. Our study evaluates RTFM and MGFN, analyzing their performance and sensitivity to anomaly types and dataset characteristics.

MGFN + OPL: minimal, effective integration. For the MGFN model on MSAD, we find that incorporating a single OPL layer generally leads to the best overall performance. This holds across both anomaly types (Table 20) and evaluation scenarios (Table 21). While adding multiple OPL layers does not drastically reduce performance (typically within a 5% drop in AUC), it introduces additional learnable parameters, which could affect training stability and efficiency. These results suggest that a single, strategically placed OPL layer is often sufficient to suppress nuisance factors effectively without

Method	Frontdoor		Mall		Office		Parkinglot		Pedestr. st.		Restaurant		Road		Shop		Sidewalk		St. highview		Train		Warehouse		**Overall**	
	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP
MSAD (B1)	84.4	84.0	80.2	74.7	74.7	65.0	87.0	30.9	93.5	53.0	91.2	87.6	80.0	55.7	82.1	69.3	86.8	63.8	98.1	95.1	70.8	9.1	89.9	76.1	86.2	68.3
MSAD (B2)	82.0	75.8	84.3	64.3	75.2	60.5	69.5	18.9	93.1	35.4	87.2	70.8	76.1	39.6	84.1	69.7	81.8	49.4	95.1	76.6	69.2	12.7	81.2	42.6	85.1	59.0
MSAD (B3)	87.5	85.5	79.1	76.3	74.7	64.4	82.7	25.5	86.9	15.1	92.9	88.6	74.7	43.6	80.2	62.7	84.6	63.4	90.3	49.7	51.8	5.5	71.1	35.0	84.3	62.8
MSAD (B1B2B3)	83.2	82.4	58.2	57.0	69.2	57.8	81.5	24.6	82.8	12.1	88.4	84.4	81.9	45.9	81.4	67.1	88.9	65.0	93.8	54.3	52.4	4.8	67.5	24.3	82.8	59.7
MSAD (C1)	85.6	82.7	54.7	48.7	70.0	59.2	82.6	25.3	80.7	11.0	87.6	81.5	79.1	43.4	81.8	69.1	83.8	61.0	85.7	35.5	53.2	3.0	75.0	35.6	82.9	58.6
MSAD (C2)	85.3	82.6	62.7	67.8	72.0	62.7	86.7	40.3	86.7	15.3	91.7	88.1	85.7	60.7	78.0	58.7	89.7	74.1	91.9	64.4	68.0	7.3	82.5	50.0	85.3	65.8
MSAD (C1C2)	85.2	84.1	57.0	64.5	74.3	61.6	78.5	20.7	88.7	22.7	91.9	87.1	77.5	43.4	83.2	67.6	84.6	61.5	87.2	41.8	62.0	8.2	67.8	28.5	84.1	62.5
MSAD (B1C1)	83.1	82.0	86.4	81.0	73.4	57.5	75.7	19.0	86.4	14.6	96.3	91.7	74.0	43.5	85.3	75.5	84.4	64.1	84.3	39.0	49.5	4.7	69.8	28.6	84.8	63.8
MSAD (B2C2)	86.4	83.5	91.0	88.7	71.1	55.3	63.9	13.3	87.8	20.6	96.7	94.5	71.3	37.7	83.4	75.2	84.3	62.3	88.3	38.5	55.6	3.5	74.4	31.0	83.6	61.3
MSAD (ALL)	82.0	77.1	78.2	61.1	72.5	57.9	64.0	13.7	92.4	28.2	94.2	84.1	68.1	33.5	83.4	71.7	84.7	55.6	77.5	24.1	50.9	3.1	75.0	31.8	82.5	56.3

Table 21. Scenario-wise performance of the MGFN model using only OPL configurations on the MSAD dataset.

Method	Assault		Explosion		Fighting		Fire		Obj. Fall		People Fall		Robbery		Shooting		Traffic Acc.		Vandalism		Water Inc.		**Overall**	
	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP
MSAD (G1O0)	52.4	59.8	66.5	76.8	88.8	92.2	77.2	89.0	90.5	95.1	45.9	42.8	65.4	80.1	71.9	81.8	53.9	46.4	83.1	75.1	81.5	96.0	84.0	65.8
MSAD (G1O1)	57.0	59.5	52.2	61.3	85.9	91.1	58.8	79.9	91.9	95.7	50.1	47.3	73.1	88.0	75.3	84.5	56.4	47.8	86.5	81.7	70.2	92.8	81.8	63.4
MSAD (G1O2)	53.1	62.5	46.8	58.3	84.7	90.5	80.1	89.8	92.8	95.9	44.7	44.4	63.6	80.8	70.3	81.9	55.9	49.6	85.4	82.2	77.9	94.9	82.7	64.5
MSAD (G1O3)	57.1	62.1	62.8	76.5	71.1	86.0	38.2	68.3	90.6	94.9	52.5	50.0	69.3	85.4	71.0	83.4	64.1	58.4	77.2	73.2	96.4	99.1	82.2	65.0
MSAD (G1O4)	60.6	62.8	58.8	67.5	87.0	91.7	74.1	87.3	90.5	94.8	52.5	49.0	71.9	83.8	89.4	92.6	66.9	54.7	78.6	79.3	81.0	95.8	84.3	62.8
MSAD (G2O0)	56.2	55.6	50.8	54.4	83.7	90.0	74.9	88.3	90.5	94.4	54.5	53.3	71.1	85.3	84.4	87.8	62.7	53.1	77.9	81.0	76.5	94.8	83.0	62.1
MSAD (G2O1)	51.8	56.1	46.3	56.3	85.4	91.0	71.0	86.0	91.4	95.3	50.7	47.1	65.9	81.1	82.1	86.9	59.3	51.2	86.5	81.9	83.5	96.5	83.5	63.4
MSAD (G2O2)	62.1	68.2	47.3	56.2	83.1	87.9	79.6	90.0	89.5	94.2	37.7	39.3	65.6	83.2	75.1	84.0	56.8	53.0	83.6	74.4	71.4	93.2	82.3	64.2
MSAD (G2O3)	58.2	64.0	63.6	62.0	84.7	91.2	62.3	82.9	90.5	94.7	51.6	49.3	68.6	81.7	78.8	83.5	57.6	49.3	88.4	86.2	89.5	97.6	84.4	66.0
MSAD (G3O0)	59.9	66.9	67.5	78.1	79.5	88.6	44.5	73.4	91.2	95.4	51.2	47.0	66.9	81.4	75.4	84.1	60.3	50.2	80.8	73.6	94.3	98.6	79.8	61.7
MSAD (G3O1)	65.1	70.7	52.5	58.9	79.2	86.1	61.6	79.9	89.6	94.2	38.0	39.0	68.0	84.2	65.0	71.8	53.1	48.0	79.0	68.6	76.7	94.4	82.3	61.3
MSAD (G3O2)	63.1	61.5	47.3	53.1	72.3	85.5	46.9	76.6	92.4	96.0	39.5	41.3	67.3	76.4	72.3	82.6	58.5	48.5	81.1	80.0	99.1	99.8	84.8	65.1
MSAD (G4O0)	60.3	67.6	67.4	78.0	79.1	88.4	43.9	73.1	91.2	95.3	51.3	47.0	66.8	81.4	74.7	83.2	60.2	50.2	81.0	73.8	95.2	98.8	79.8	61.5
MSAD (G4O1)	59.6	65.7	66.7	77.5	77.8	87.6	39.1	70.6	90.6	94.9	50.7	45.9	66.5	81.9	75.3	84.2	58.1	48.8	81.4	73.8	96.1	99.0	80.6	62.3
MSAD (G5O0)	64.7	69.7	49.7	60.5	67.8	81.5	43.0	72.7	88.3	93.6	35.4	41.2	70.8	82.3	71.3	80.1	53.8	50.1	76.1	74.0	97.2	99.3	82.0	63.2

Table 22. MGFN with G-OPL/OPL on MSAD (anomaly-level). Anomaly-wise performance of various GxOy configurations in the MGFN model on the MSAD dataset. The G1O0 configuration consistently achieves top results across different anomaly types.

Method	Frontdoor		Mall		Office		Parkinglot		Pedestr. st.		Restaurant		Road		Shop		Sidewalk		St. highview		Train		Warehouse		**Overall**	
	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP
MSAD (G1O0)	84.4	83.8	90.0	84.8	75.9	62.3	70.4	16.6	90.5	25.8	95.7	90.2	71.4	43.1	79.7	64.5	83.8	63.3	87.7	41.3	44.7	2.3	64.8	41.1	84.0	65.8
MSAD (G1O1)	83.5	83.0	53.6	49.6	73.7	60.8	71.2	21.1	86.8	14.9	91.8	87.0	73.0	38.4	86.8	81.8	84.4	60.6	69.1	27.5	47.8	2.5	55.6	24.7	81.8	63.4
MSAD (G1O2)	82.8	82.9	63.4	60.9	74.5	62.0	75.5	20.1	84.4	13.1	93.8	88.5	72.6	39.9	84.4	75.4	84.5	62.0	97.2	86.1	44.6	2.3	66.3	37.7	82.7	64.5
MSAD (G1O3)	85.5	84.5	86.7	82.6	75.0	66.1	74.5	30.7	62.3	7.4	84.2	80.4	74.9	42.3	82.9	77.1	85.3	61.7	60.5	24.4	61.4	37.1	66.4	38.3	82.2	65.0
MSAD (G1O4)	87.2	85.9	77.3	74.6	73.9	61.7	79.6	22.9	80.2	13.1	88.6	83.8	74.1	40.9	83.0	69.8	85.8	62.2	90.1	52.0	45.3	2.5	71.3	39.8	84.3	62.8
MSAD (G2O0)	83.4	82.7	54.6	45.7	72.9	60.4	68.6	14.9	86.9	14.8	86.9	81.9	73.3	44.4	82.6	74.8	86.3	61.7	92.7	50.7	37.8	4.1	66.2	24.5	83.0	62.1
MSAD (G2O1)	84.4	83.6	67.8	59.2	73.9	61.4	79.9	21.6	92.6	28.6	94.9	90.3	66.3	34.8	83.4	73.6	83.5	63.8	92.5	58.3	45.7	2.4	60.5	33.9	83.5	63.4
MSAD (G2O2)	81.0	80.6	58.6	59.1	75.4	60.7	73.0	18.5	90.2	22.0	89.2	81.5	61.6	33.0	82.1	74.6	73.3	55.7	97.6	87.4	65.1	10.7	71.3	39.6	82.3	64.2
MSAD (G2O3)	82.6	82.1	89.2	83.5	77.4	62.9	73.9	20.0	95.3	44.1	94.3	88.4	73.7	45.6	83.8	73.3	86.1	64.3	66.7	25.3	50.2	2.6	84.4	58.3	84.4	66.0
MSAD (G3O0)	85.9	84.6	89.3	84.4	73.7	63.8	64.2	13.4	74.3	52.2	85.2	81.4	65.3	35.5	84.1	74.9	81.9	58.8	62.6	25.3	49.0	2.5	65.5	38.4	79.8	61.7
MSAD (G3O1)	78.9	76.8	70.7	54.0	76.8	64.0	54.2	11.0	91.4	22.9	93.5	87.1	66.0	30.2	85.0	75.0	76.9	56.6	83.3	37.4	53.6	3.0	81.6	40.3	82.3	61.3
MSAD (G3O2)	84.4	84.8	94.6	91.3	74.8	59.0	71.2	19.7	88.0	17.0	93.1	85.7	69.9	34.9	84.7	70.6	83.6	61.3	77.1	31.5	57.2	5.4	65.3	24.8	84.8	65.1
MSAD (G4O0)	85.9	84.6	89.8	85.7	73.7	64.0	64.7	13.6	73.8	51.5	85.2	81.3	65.3	36.2	84.3	75.2	81.7	58.3	61.6	25.1	50.8	2.6	65.8	38.4	79.8	61.5
MSAD (G4O1)	85.8	84.2	89.5	85.1	73.6	63.2	67.5	14.9	75.5	47.1	85.2	81.4	67.2	36.3	83.5	74.8	82.2	58.6	57.6	24.1	49.7	2.6	70.4	44.6	80.6	62.3
MSAD (G5O0)	78.8	79.5	68.7	67.1	74.9	62.4	67.1	18.0	88.6	16.4	89.1	79.0	67.3	41.1	83.8	69.9	83.2	58.3	52.3	22.4	61.3	4.5	73.3	35.4	82.1	63.2

Table 23. MGFN with G-OPL/OPL on MSAD (scenario-level). Scenario-wise results for different GxOy configurations using MGFN on the MSAD dataset. The G1O0 setup consistently outperforms most other placements across scenarios.

Method	Assault		Explosion		Fighting		Fire		Obj. Fall		People Fall		Robbery		Shooting		Traffic Acc.		Vandalism		Water Inc.		**Overall**	
	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP
MSAD (C1)	57.0	62.4	77.7	85.7	74.1	84.8	49.6	75.5	87.7	92.1	53.3	50.4	72.4	89.0	84.1	89.5	69.5	58.7	84.8	80.9	99.2	99.8	86.5	68.2
MSAD (C2)	51.6	66.1	73.2	82.7	76.8	85.5	61.0	82.8	83.4	89.2	47.7	54.7	68.9	87.8	75.6	82.1	84.4	62.1	86.9	83.2	98.7	99.7	86.0	68.9
MSAD (C3)	63.5	74.4	85.9	90.0	81.0	88.6	76.6	87.8	89.3	93.5	58.8	57.4	67.8	86.3	82.0	88.3	64.0	58.5	89.4	87.5	98.9	99.8	85.1	69.0
MSAD (C4)	63.1	74.8	72.8	84.0	80.0	86.9	75.5	79.6	86.4	91.7	53.3	56.2	70.0	88.7	75.6	80.8	64.3	58.3	84.3	82.2	98.3	99.6	85.0	68.7
MSAD (C5)	66.2	70.8	79.6	86.7	77.0	86.9	51.6	76.1	90.1	93.5	44.2	48.9	73.1	88.1	87.4	90.7	59.6	57.7	74.7	71.2	99.5	99.9	86.2	68.6
MSAD (C2C3)	67.3	72.0	74.7	83.2	69.4	84.2	56.2	77.5	88.3	92.4	52.6	53.8	65.9	85.2	88.6	89.3	64.5	58.4	78.8	75.4	99.9	100	85.6	68.2
MSAD (ALL)	52.8	60.6	72.0	82.1	75.8	86.4	52.8	77.2	88.6	92.4	56.1	52.5	71.6	88.2	81.7	88.3	66.0	55.5	84.5	81.0	99.7	99.9	86.4	69.5

Table 24. Anomaly-wise performance of the RTFM model using only OPL configurations on the MSAD dataset.

Method	Frontdoor		Mall		Office		Parkinglot		Pedestr. st.		Restaurant		Road		Shop		Sidewalk		St. highview		Train		Warehouse		**Overall**	
	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP
MSAD (C1)	85.7	82.4	85.7	80.3	77.2	71.8	76.8	26.3	96.6	50.4	90.4	81.5	76.8	53.3	88.6	82.9	84.9	56.4	66.5	26.7	42.9	2.3	86.1	66.8	86.5	68.2
MSAD (C2)	84.9	83.0	91.2	87.3	76.5	73.7	77.1	23.3	76.7	11.8	85.1	74.3	66.0	39.0	85.3	79.0	86.2	58.0	63.3	26.1	71.0	20.3	88.0	72.8	86.0	68.9
MSAD (C3)	89.1	87.7	85.9	80.5	80.8	78.3	60.3	23.4	86.8	14.9	92.2	86.6	80.8	54.9	77.3	68.6	80.4	57.1	80.3	34.8	63.4	21.8	83.5	69.4	85.1	69.0
MSAD (C4)	84.6	83.8	94.5	88.8	74.9	72.0	81.1	29.6	81.3	12.0	89.6	79.4	77.4	50.5	83.9	78.2	84.9	61.3	52.2	23.1	50.4	17.5	81.0	56.3	85.0	68.7
MSAD (C5)	82.1	80.5	97.5	95.9	84.2	79.7	87.4	64.4	78.6	11.3	88.1	79.1	86.4	58.6	82.4	68.3	90.1	71.7	48.0	22.4	91.2	28.2	80.0	54.6	86.2	68.6
MSAD (C2C3)	83.4	80.8	89.3	86.3	81.7	76.9	87.2	36.3	76.9	9.2	84.7	77.9	79.6	50.9	80.9	74.3	88.7	67.7	69.3	27.6	67.8	4.8	84.8	60.9	85.6	68.2
MSAD (ALL)	86.0	83.3	87.9	83.6	77.6	72.4	85.2	33.4	78.8	21.9	89.6	80.8	73.4	46.9	87.4	82.2	86.6	65.5	61.8	25.1	54.2	2.8	88.1	69.6	86.4	69.5

Table 25. Scenario-wise performance of the RTFM model using only OPL configurations on the MSAD dataset.

Method	Assault		Explosion		Fighting		Fire		Obj. Fall		People Fall		Robbery		Shooting		Traffic Acc.		Vandalism		Water Inc.		**Overall**	
	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP
MSAD (G1O0)	50.2	62.4	69.4	80.6	69.5	84.4	71.8	87.0	88.7	92.4	52.3	53.3	71.4	88.2	87.9	91.0	62.5	54.7	82.0	79.6	97.5	99.4	88.0	70.9
MSAD (G1O1)	66.0	77.5	69.6	80.8	66.5	81.2	55.2	78.9	78.0	86.7	55.1	68.3	64.7	86.6	74.5	82.5	65.2	63.5	85.7	81.8	100	100	86.6	70.5
MSAD (G1O2)	50.6	72.4	56.2	78.3	58.4	81.1	50.9	80.7	79.9	88.0	47.2	59.4	65.0	89.0	66.4	80.2	66.4	68.7	88.6	87.8	83.8	96.5	84.8	70.3
MSAD (G1O3)	55.4	68.5	73.5	84.2	75.5	85.1	72.7	87.2	82.3	88.1	39.2	50.7	70.2	88.4	67.2	79.0	60.9	60.3	82.9	83.6	98.8	99.6	85.5	69.2
MSAD (G1O4)	46.6	67.5	68.9	82.3	71.4	83.2	79.6	90.8	84.4	90.2	48.4	60.1	67.5	88.6	69.7	81.1	67.1	65.4	86.8	87.0	95.8	99.1	85.2	69.7
MSAD (G1O5)	45.5	62.9	74.9	85.0	78.3	86.8	86.8	94.4	84.5	89.4	46.8	55.1	69.1	87.7	80.0	85.8	66.2	63.2	79.7	79.9	93.5	98.6	86.3	69.7
MSAD (G2O0)	44.9	66.9	67.3	82.3	73.0	84.0	73.8	88.0	80.9	88.6	43.7	58.2	65.7	88.4	80.2	87.2	68.1	66.7	82.9	83.4	99.8	100	86.5	70.9
MSAD (G2O1)	61.4	73.4	69.5	82.4	74.1	84.2	67.6	84.7	86.1	91.3	42.3	53.8	65.5	87.7	82.4	87.9	65.7	65.1	85.2	85.1	99.7	99.9	85.4	69.6
MSAD (G2O2)	50.4	74.7	60.6	79.6	69.6	84.6	54.3	83.1	71.5	84.2	51.9	70.1	65.2	88.6	81.7	87.9	64.7	66.6	82.6	83.7	99.8	100	83.5	70.5
MSAD (G2O3)	50.1	69.8	67.4	82.1	68.6	85.9	71.9	86.3	80.4	88.1	47.8	60.8	66.4	88.0	73.0	83.1	63.3	62.4	85.5	83.7	99.0	99.7	86.9	71.6
MSAD (G2O4)	59.3	74.1	65.3	80.6	69.1	83.1	76.6	89.8	86.1	90.8	43.9	52.4	64.9	87.2	70.5	82.8	68.3	67.7	91.4	89.5	97.7	99.3	86.7	70.7
MSAD (G3O0)	58.0	72.0	66.7	81.1	60.4	81.7	72.7	87.1	83.8	90.0	41.6	54.4	67.3	88.3	68.8	79.8	56.5	56.8	89.1	87.4	98.9	99.7	86.1	70.1
MSAD (G3O1)	59.1	74.1	61.3	78.9	71.0	82.9	79.6	89.6	86.3	91.2	49.1	58.0	66.4	88.7	65.8	79.5	64.8	64.0	89.1	88.3	85.6	96.7	86.5	70.4
MSAD (G3O2)	66.0	79.1	64.8	80.0	67.9	81.1	44.0	76.1	72.7	84.3	45.4	61.8	66.5	88.2	81.6	86.7	61.7	57.3	81.0	78.6	99.9	100	85.2	69.3
MSAD (G3O3)	46.9	66.2	60.9	79.1	62.6	80.0	80.1	91.2	84.2	89.9	47.3	53.9	64.3	88.1	70.2	82.5	68.7	67.7	86.2	84.5	97.4	99.3	85.2	69.7
MSAD (G4O0)	50.7	68.0	70.7	83.5	76.9	86.4	79.2	89.4	83.8	89.7	41.6	54.2	66.9	88.3	74.7	83.6	66.8	64.3	89.0	87.9	99.9	100	86.8	71.4
MSAD (G4O1)	64.7	79.0	61.5	79.3	63.9	79.5	44.5	76.5	72.3	84.5	50.1	67.7	65.0	87.9	78.1	84.9	63.6	64.8	85.5	84.1	99.9	100	85.1	70.3
MSAD (G4O2)	53.8	73.2	67.2	82.2	73.7	83.9	79.1	91.1	82.9	89.5	44.7	57.7	68.8	89.7	76.7	83.3	69.9	69.9	79.4	79.8	99.8	100	86.7	72.2
MSAD (G5O0)	60.1	77.6	57.5	78.4	67.9	83.3	43.6	76.2	72.7	84.7	50.9	68.8	63.2	87.7	70.7	81.6	62.0	64.9	81.5	82.3	99.5	99.9	83.0	68.7
MSAD (G5O1)	50.5	67.1	68.1	80.8	77.2	85.8	71.9	86.6	84.2	89.8	45.3	54.7	72.0	89.5	81.9	88.8	64.8	61.1	87.1	84.5	99.9	100	86.7	70.2
MSAD (G6O0)	58.7	73.9	66.2	80.1	59.8	79.9	76.0	88.2	84.3	90.2	41.5	56.0	66.0	87.7	83.4	86.1	61.0	59.5	86.1	85.0	95.3	99.0	87.2	70.6

Table 26. RTFM with G-OPL/OPL on MSAD (anomaly-level). Anomaly-wise performance of various GxOy configurations in the RTFM model on the MSAD dataset. The G1O0 configuration consistently achieves top results across different anomaly types.

Method	Frontdoor		Mall		Office		Parkinglot		Pedestr. st.		Restaurant		Road		Shop		Sidewalk		St. highview		Train		Warehouse		**Overall**	
	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP
MSAD (G1O0)	82.0	79.3	91.0	81.4	74.3	72.0	79.4	27.2	86.9	36.1	90.3	81.4	72.4	46.7	89.0	82.5	87.0	65.1	84.9	37.8	70.4	12.0	86.3	79.6	88.0	70.9
MSAD (G1O1)	81.0	82.5	87.8	84.2	75.1	75.0	87.9	46.9	95.7	51.3	85.6	76.2	85.1	57.4	81.6	73.8	85.3	63.5	66.9	24.8	55.0	4.5	88.2	71.9	86.6	70.5
MSAD (G1O2)	80.2	82.2	88.4	80.6	75.2	75.7	80.6	47.0	86.7	25.0	90.4	81.5	83.3	64.4	82.8	79.2	83.6	62.2	90.3	65.3	51.7	2.6	76.8	63.4	84.8	70.3
MSAD (G1O3)	78.3	78.3	75.9	66.8	73.8	73.7	86.9	54.8	92.5	24.9	92.9	85.3	88.3	59.7	83.8	76.8	87.3	66.1	66.0	24.3	37.5	2.1	85.9	71.2	85.5	69.3
MSAD (G1O4)	82.2	82.7	70.6	63.1	73.6	74.7	69.1	34.7	98.1	74.7	92.7	85.2	84.4	53.4	82.4	77.0	82.5	59.7	86.3	56.7	19.7	1.7	87.3	69.0	85.2	69.7
MSAD (G1O5)	80.9	81.0	72.0	62.8	74.7	72.9	70.7	27.3	89.8	36.3	94.8	88.5	79.7	51.4	84.8	77.5	80.7	56.9	96.3	82.4	50.0	3.2	86.7	66.2	86.3	69.7
MSAD (G2O0)	80.7	81.4	78.7	66.1	74.2	75.6	80.2	46.7	93.9	54.2	91.8	83.7	89.3	62.8	85.4	80.2	84.5	64.3	76.3	34.3	34.9	2.0	88.3	71.7	86.5	70.9
MSAD (G2O1)	80.3	81.0	83.6	71.6	71.9	73.8	81.9	48.0	79.3	25.6	92.4	84.5	88.0	61.3	84.0	78.1	88.2	66.1	66.6	22.6	72.0	24.7	83.6	63.9	85.4	69.6
MSAD (G2O2)	74.6	80.9	81.0	71.1	75.9	79.1	88.1	57.0	70.3	23.5	80.3	72.5	88.7	67.2	78.8	75.9	79.7	61.7	81.0	56.6	74.6	35.9	84.6	69.8	83.5	70.5
MSAD (G2O3)	81.3	81.9	88.5	80.9	74.1	73.6	83.4	53.8	97.6	63.0	90.8	81.9	90.2	71.5	83.8	78.6	83.6	62.0	89.9	55.9	46.6	2.4	88.4	71.1	86.9	71.6
MSAD (G2O4)	79.6	80.5	87.5	73.2	76.5	75.7	88.9	57.1	89.9	21.4	90.2	82.2	87.7	65.9	84.0	76.8	85.6	62.7	93.7	72.5	60.4	3.3	82.8	62.5	86.7	70.7
MSAD (G3O0)	80.0	80.3	89.9	82.6	74.3	74.9	80.5	46.1	95.7	39.9	90.8	81.8	83.0	61.8	84.1	78.1	81.9	58.2	91.2	59.7	61.3	3.7	85.1	67.2	86.1	70.1
MSAD (G3O1)	80.7	81.4	89.0	80.3	76.3	75.8	80.2	44.7	96.0	42.3	88.4	79.5	85.1	54.6	85.0	80.0	84.4	60.3	88.9	48.2	73.2	4.7	83.6	63.9	86.5	70.4
MSAD (G3O2)	76.9	79.7	87.0	75.0	76.1	77.3	89.8	50.6	63.8	30.7	84.1	74.7	80.4	51.2	80.6	74.9	83.6	62.4	66.4	21.1	60.3	4.8	85.6	69.6	85.2	69.3
MSAD (G3O3)	80.3	80.2	68.5	64.6	72.3	74.4	81.6	47.8	85.9	14.3	91.1	83.4	85.0	55.3	86.1	81.5	82.8	60.7	94.0	73.1	56.0	4.4	79.2	59.5	85.2	69.7
MSAD (G4O0)	81.9	82.2	88.3	78.3	75.9	75.9	78.2	47.3	97.9	63.9	90.9	81.9	87.9	66.5	84.5	79.2	85.1	63.5	79.3	31.8	46.9	2.6	87.7	68.4	86.8	71.4
MSAD (G4O1)	77.3	81.5	89.6	79.2	75.6	77.4	90.9	59.0	82.3	32.9	83.2	75.0	84.2	54.5	79.2	74.1	83.3	65.2	76.1	25.8	57.6	23.2	83.4	68.2	85.1	70.3
MSAD (G4O2)	78.7	81.2	87.3	80.1	76.4	76.6	73.7	35.4	94.5	52.0	92.0	84.6	81.5	46.3	86.2	84.5	84.5	63.4	91.8	74.1	42.3	22.3	83.0	69.1	86.7	72.2
MSAD (G5O0)	74.8	80.8	83.3	73.3	74.1	77.0	86.8	56.7	85.9	28.9	82.7	75.8	85.0	60.6	78.7	74.5	82.1	63.3	61.2	19.1	87.0	26.8	80.8	65.4	83.0	68.7
MSAD (G5O1)	83.0	82.3	82.8	66.4	75.4	74.6	70.4	34.5	98.3	70.8	91.6	81.9	75.3	44.9	86.4	79.7	85.8	62.5	86.1	38.4	44.2	14.0	90.5	79.4	86.7	70.2
MSAD (G6O0)	79.5	80.4	72.9	62.8	75.9	76.1	78.8	42.2	97.4	65.5	92.1	83.4	84.6	57.3	85.4	78.6	84.8	61.4	86.4	41.5	76.1	6.3	90.7	80.0	87.2	70.6

Table 27. RTFM with G-OPL/OPL on MSAD (scenario-level). Scenario-wise results for different GxOy configurations using RTFM on the MSAD dataset. The G1O0 setup consistently outperforms most other placements across scenarios.

Configuration	MSAD		ShT		UCF		CUHK		Ped2	
	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP
G1O0	88.0	70.9	97.3	74.7	78.3	30.9	84.9	66.2	85.0	71.4
G1O1	84.4	68.6	96.9	72.3	78.1	34.3	84.1	64.3	81.4	69.0
G1O2	86.9	69.8	97.0	70.5	76.5	35.7	79.7	60.3	80.0	67.9
G1O3	86.5	69.5	96.9	72.7	72.6	30.9	84.3	66.1	82.2	71.4
G1O4	87.5	71.1	96.5	72.3	69.3	27.2	83.2	62.0	84.7	74.2
G1O5	86.2	67.6	96.4	71.7	73.0	42.7	83.3	64.8	84.2	71.8
G2O0	84.9	69.7	97.1	73.8	75.7	29.0	82.9	64.2	79.3	67.6
G2O1	85.6	68.9	96.6	72.5	77.7	43.4	82.9	64.4	82.1	73.5
G2O2	83.7	68.3	97.1	73.5	76.2	26.3	82.7	61.8	78.4	65.7
G2O3	87.4	71.4	97.0	73.1	72.3	30.6	82.7	62.4	84.2	70.6
G2O4	84.8	68.3	96.4	71.7	75.8	24.3	81.2	62.5	85.0	74.2
G3O0	86.2	71.1	96.6	73.5	74.4	33.2	82.8	63.2	78.1	65.9
G3O1	82.4	68.5	96.7	72.9	75.3	23.8	84.3	65.8	88.0	73.9
G3O2	85.2	70.7	97.0	73.2	75.7	37.3	81.8	62.2	77.9	66.6
G3O3	85.5	70.4	96.3	73.0	75.0	29.0	64.4	82.8	83.1	71.9
G4O0	86.4	69.0	96.5	72.4	76.5	26.4	63.8	82.6	78.5	67.8
G4O1	82.5	70.2	96.5	72.4	76.7	27.8	61.5	81.5	84.4	70.5
G4O2	85.9	70.4	97.0	72.1	71.3	24.5	62.0	82.3	81.3	68.5
G5O0	84.5	71.1	97.3	73.5	77.7	35.0	63.9	82.5	89.6	75.0
G5O1	86.3	66.3	96.7	73.5	71.1	27.7	67.3	84.0	87.0	81.2
G6O0	83.6	69.2	97.0	71.5	74.0	27.0	66.5	83.3	81.9	75.3

Table 28. Performance of RTFM with G-OPL/OPL across five datasets. This table presents the results of various G-OPL and OPL layer combinations on five benchmark datasets using the RTFM model. The performance trends inform the selection of optimal layer configurations, as summarized in Table 10.

Configuration	MSAD		ShT		UCF		CUHK		Ped2	
	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP
G1O0	84.0	65.8	83.7	42.0	83.3	15.2	69.4	43.5	93.9	93.6
G1O1	81.8	63.4	77.0	22.4	77.1	14.0	62.9	35.0	83.9	72.2
G1O2	82.7	64.5	66.5	17.8	76.1	14.0	62.9	36.6	70.0	61.3
G1O3	82.2	65.0	73.2	14.1	76.9	13.8	60.4	32.5	92.4	89.7
G1O4	83.3	62.8	77.5	24.3	77.3	12.3	59.3	36.9	71.9	60.4
G2O0	83.0	62.1	55.0	11.7	72.6	13.6	59.1	28.8	74.3	63.4
G2O1	83.5	63.4	65.9	12.4	80.1	14.8	68.7	36.8	85.0	78.1
G2O2	82.3	64.2	73.5	15.4	75.7	12.8	69.8	37.7	91.7	89.7
G2O3	83.4	66.0	75.5	17.0	73.8	9.3	68.4	36.3	77.9	56.9
G3O0	79.8	61.7	84.1	25.9	74.4	13.3	70.8	40.5	73.2	64.1
G3O1	82.3	61.3	77.5	15.4	74.1	13.2	65.5	33.4	68.1	59.2
G3O2	84.8	65.1	71.5	14.0	76.5	9.7	63.1	36.5	79.0	72.5
G4O0	79.8	61.5	81.2	23.2	78.1	12.5	60.2	33.5	80.6	69.7
G4O1	80.6	62.3	75.3	17.7	74.8	10.0	61.4	33.4	57.2	36.7
G5O0	82.0	63.2	83.6	23.2	74.6	24.7	69.7	45.3	78.5	61.7

Table 29. Performance of MGFN with G-OPL/OPL across five datasets. This table presents the results of various G-OPL and OPL layer combinations on five benchmark datasets using the MGFN model. The performance trends inform the selection of optimal layer configurations, as summarized in Table 10.

overcomplicating the model.

RTFM + OPL: sensitive to overuse. A similar pattern is observed in the RTFM model (Table 24 and 25). Here, too, the best performance is achieved when a single OPL layer is used, for instance, producing the highest AUC for MSAD (C1). However, adding additional OPL layers tends to slightly degrade performance, although still within a reasonable range. This indicates that overuse of OPL may suppress not only irrelevant information but also task-relevant signals, thus harming the VAD capability.

MGFN + G-OPL + OPL: flexible but layer-dependent. When both G-OPL and OPL layers are used in the MGFN architecture, performance becomes more sensitive to layer arrangement (Table 22 and 23). The best AUC is achieved using three G-OPL layers followed by two OPL layers, whereas the best average precision (AP) is observed with two G-OPL layers followed by three OPL layers. Nonetheless, we find that using even a single G-OPL layer yields competitive performance while significantly reducing computational cost. This makes it a favorable option for real-world deployments where efficiency is crucial.

RTFM + G-OPL: less sensitive to face suppression. In contrast, for the RTFM model, the best performance with G-OPL is achieved when only one layer is added (Table 26 and 27). This suggests that facial information, used as a guiding attribute in G-OPL, is not strongly correlated with anomalous events in the RTFM pipeline, and that extensive suppression may inadvertently obscure useful contextual cues. Thus, minimal semantic suppression proves to be the most effective approach in this setting.

Dataset-specific trends and exceptions. Across the remaining four datasets, ShanghaiTech, UCF-Crime, CUHK Avenue, and UCSD Ped2, we observe similar trends, particularly in the RTFM architecture (Table 28 and 29). A single G-OPL layer typically offers the best balance between privacy control and detection performance. However, Ped2 emerges as an exception. Due to its low resolution and limited facial visibility, facial signals are harder to detect, making it beneficial to stack multiple G-OPL layers to ensure more thorough suppression of residual sensitive information.

Our findings suggest that using a single OPL or G-OPL layer is often sufficient and optimal, offering strong performance with low overhead. Excessive stacking of projection layers can degrade detection performance by suppressing task-relevant signals, especially in architectures like RTFM. The optimal number of projection layers also depends on dataset characteristics such as resolution and the clarity of sensitive attributes. These results offer valuable guidance for effectively and efficiently incorporating privacy-preserving and interpretable layers into VAD architectures.

Stacking behavior. On low-resolution datasets (*e.g.*, UCSD Ped2), where attribute signals are weak and noisy, stacking multiple G-OPL layers yields additional gains.

The dominant mechanism is progressive removal of residual attribute components: a single projection may not fully eliminate weak or entangled signals, while successive layers iteratively refine the orthogonal complement. A secondary effect is robustness to imperfect guidance, where stacking mitigates errors in the estimated attribute subspace.

This explains why a single layer is sufficient when attribute signals are strong, whereas stacking is beneficial primarily in low-signal regimes. Stacking functions as an iterative subspace refinement process rather than merely increasing model capacity.

N. Limitations

While our proposed method demonstrates strong and consistent improvements across diverse datasets and anomaly types, several limitations remain.

First, the performance gains heavily rely on carefully tuning the placement and frequency of the disentanglement modules (G-OPL and OPL). Although we provide empirical insights into effective configurations, these may not generalize optimally to unseen domains without additional validation or adaptation.

Second, our method exhibits less stability on datasets characterized by high scene diversity or subtle anomaly cues, such as UCF-Crime. This suggests that while our approach effectively filters irrelevant signals, it may still struggle in scenarios where anomaly patterns are highly variable or context-dependent.

Third, the disentanglement process, while beneficial for suppressing task-irrelevant information, introduces additional architectural complexity and computational overhead. This may limit the scalability of our approach for real-time or

resource-constrained applications.

Finally, while our focus on privacy-preserving datasets highlights the robustness of our method, its effectiveness on privacy-unconstrained, high-resolution datasets remains less explored. Future work could investigate the interplay between disentanglement, resolution, and anomaly characteristics in broader settings.

O. LLM Usage Declaration

We disclose the use of Large Language Models (LLMs) as general-purpose assistive tools during the preparation of this manuscript. LLMs were used only for minor tasks such as grammar and style improvement, code verification, and formatting suggestions. No scientific ideas, analyses, experimental designs, or conclusions were generated by LLMs. All core research, methodology, experiments, and results were performed and fully verified by the authors.

The authors take full responsibility for all content presented in this paper, including text or code suggestions that were refined with the assistance of LLMs. No content generated by LLMs was treated as original scientific work, and all references and claims have been independently verified. LLMs did not contribute in a manner that would qualify them for authorship.