LEARN LOCALLY, CALIBRATE GLOBALLY: CLIENT-LEVEL DIFFERENTIALLY PRIVATE VIDEO ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

An anomaly detector can rank moments within a video correctly yet misorder normal and abnormal frames across videos. We exploit this gap to rethink private collaboration: transfer knowledge for cross-video score alignment rather than an entire detection model. In this paper, we propose Public-Anchor Response Decoding for weakly supervised video anomaly detection (PARD-VAD), which learns locally and calibrates globally. Each client projects differences between matched private and public-only teacher responses onto a public function basis. Clipping, secure aggregation, and Gaussian perturbation produce a single two-coefficient differentially private (DP) release while teacher parameters remain local. Public decoding reconstructs temporal evidence, and public-score-weighted pooling converts it into one log-odds correction per video. This changes cross-video rankings while preserving within-video ordering, with inference performed as DP post-processing. For fixed public assets, sensitivity analysis and Rényi DP accounting establish client-level replace-one privacy. Using the same release and decoding rule, PARD-VAD achieves 89.19% AUC on UCF-Crime and 84.03% AP on XD-Violence on the complete test sets at a per-deployment budget of $\epsilon \leq 8$, $\delta = 10^{-5}$, exceeding the reported privacy-preserving baselines in our comparison. Matched public-only controls isolate gains of 0.81 and 0.54 percentage points, and an exhaustive frame-pair audit attributes these gains to cross-video ordering.

1 INTRODUCTION

Video anomaly detection (VAD) aims to locate rare events, such as violence, theft, and traffic accidents, in long surveillance videos. Weakly supervised learning makes this task practical by learning from video-level labels instead of densely annotated event boundaries (Sultani et al., 2018). However, the videos needed to cover diverse anomalies are distributed across institutions with different scenes and activity patterns. These collections also contain sensitive information about people and their routines, restricting their centralized use. Distributed VAD therefore faces a central tension: it must benefit from complementary local observations while controlling what collaboration reveals about each client's data.

Federated VAD methods address the data-sharing barrier by training locally and exchanging learned information (Qi et al., 2024; Wang et al., 2025). Yet keeping videos local does not itself bound information leakage: gradients, updates, and model parameters can expose training information through inversion or inference attacks (Zhu et al., 2019; Geiping et al., 2020; Shokri et al., 2017). Client-level differential privacy (DP) provides such a bound by limiting how much replacing an entire client's collection can change the distribution of released outputs. Preserving detection utility under this constraint calls for careful release design. Clipping limits each client's influence, and Gaussian noise perturbs the communicated representation; releasing a full model or dense responses subjects many coordinates to this perturbation, including behavior already available from public models. The question is therefore not only how to privatize communication, but *what private knowledge the detector needs to receive*.

Our starting point is that temporal localization and cross-video score comparability play distinct roles in VAD. A public detector can provide useful ordering of moments within a video while as-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

signing inconsistent score levels across videos: an abnormal moment in one scene may score below a normal moment in another. Because benchmark evaluation ranks frames across the complete test collection, these cross-video comparisons matter even when within-video ordering is informative. This suggests a complementary role for private collaboration: use local knowledge to align the scores of a public detector while retaining its temporal structure. Rather than requiring the private release to carry an entire detection function, this division of labor targets the corrections needed for global ranking. We call this principle *learn locally, calibrate globally*.

In this paper, we introduce Public-Anchor Response Decoding for VAD (PARD-VAD) to implement this principle. Each client trains a private teacher and evaluates it alongside a matched public-only teacher on shared public anchors. Subtracting their centered temporal responses expresses the client's contribution relative to a public reference. We then project this difference into a basis of public response functions, obtaining a compact coefficient vector that can be jointly clipped and aggregated. A trusted release service combines secure aggregation with Gaussian noise to publish the coefficients while teacher weights and dense responses remain local. The resulting release contains only two coordinates in our implementation, separating the capacity of local learning from the size of the communicated knowledge.

To turn these coefficients into useful corrections, the decoder evaluates the same public functions on query videos and reconstructs temporal evidence. After smoothing and scale calibration, it pools this evidence according to public anomaly scores, so the correction depends on how the released response interacts with the video's score profile. The pooled signal determines one log-odds shift per video. Different videos can therefore be adjusted upward or downward while each video's temporal ordering is preserved. Reconstruction and calibration use only the released vector, fixed public assets, and query features, making inference DP post-processing without further access to private teachers or post-release student training. This design connects local behavior learning to global score alignment through a compact private release. Building on this release-and-calibration design, our contributions are threefold.

- We formulate private knowledge transfer in distributed weakly supervised VAD around matched temporal-response innovations, using a public function basis to communicate changes relative to a public reference.
- We develop a public-score-weighted decoder that converts this compact release into video-dependent calibration and makes its effect on cross-video ranking explicit.
- We establish the release's client-level privacy bound through sensitivity analysis and Rényi DP accounting, and evaluate the shared mechanism through complete-test comparisons, matched private-contribution ablations, budget sensitivity experiments, and an exact ranking audit.

2 Related Work

2.1 Weakly Supervised Video Anomaly Detection

MIL ranking established video-level supervision as a practical alternative to dense temporal annotation (Sultani et al., 2018). Subsequent methods improve pseudo-labeling, representation learning, feature magnitude modeling, and temporal reasoning (Feng et al., 2021; Chen et al., 2023). CLIP (Radford et al., 2021) enables semantic transfer through pretrained visual-language features. These advances inform our public reference and local teachers; we study their use under client-level privacy. We distinguish cross-video score calibration from within-video localization because both affect frame-level benchmark ranking.

2.2 Federated and Privacy-Preserving VAD

CLAP, FedVAD, Fed-WSVAD, and Fed-L2G Experts learn across distributed anomaly-detection clients (Al-lahham et al., 2024; Qi et al., 2024; Wang et al., 2025; Su et al., 2025). Fed-WSVAD communicates multimodal prompt learning information, whereas our clients release coefficients of response differences on fixed public inputs. PPVAD combines federated learning with vision-language representations (Li et al., 2026); FedVAR aligns normality prototypes for federated anomaly recognition and also reports binary detection results (Haider et al., 2026). Their privacy settings are

not interchangeable with our client-level DP guarantee. TeD-SPAD, PRIME, and SPLAVU instead study anonymization or motion-based representations (Fioresi et al., 2023; Su et al., 2023; Fioresi et al., 2026), addressing disclosure in visual inputs rather than the same released-model threat model.

2.3 Private Knowledge Transfer

PATE transfers knowledge from private teachers through noisy aggregation (Papernot et al., 2017; 2018); FedMD uses distillation to support heterogeneous federated models (Li & Wang, 2019). Keeping teachers private and using public queries is therefore an established principle. Our focus is a matched temporal-response representation and its public-score-weighted readout for cross-video ranking. We use standard Gaussian DP and RDP composition (Dwork et al., 2006; Abadi et al., 2016; Mironov, 2017). Secure aggregation (Bonawitz et al., 2017) conceals individual contributions under its system assumptions; the noise mechanism supplies the formal privacy bound.

3 Methodology

3.1 Problem Setup and Preliminaries

Problem Setup. The federated dataset is $D = (D_1, \ldots, D_N)$, where N is a fixed number of clients and $D_i = \{(X_{ij}, y_{ij})\}_{j=1}^{n_i}$ is client i's local collection. A video $X = (x_1, \ldots, x_L)$ contains L feature vectors $x_\tau \in \mathbb{R}^d$; $y \in \{0, 1\}$ denotes its normal/abnormal label. We seek a scoring function producing an anomaly probability at each temporal position τ. A labeled public anchor pool $\mathcal{A} = \{(A_k, y_k)\}_{k=1}^{K}$, containing K source videos, is fixed outside the protected dataset.

Privacy definition. Two datasets $D \sim D'$ are replace-one client adjacent if they differ only in one complete collection D_c, for a client index $c \in \{1, \ldots, N\}$. The public pool, client identities, and all public design choices are held fixed. A randomized release mechanism M, with output space Ω, is (ϵ, δ)-DP if

$$\Pr[M(D) \in E] \leq e^{\epsilon} \Pr[M(D') \in E] + \delta \quad \text{for every } D \sim D' \text{ and measurable } E \subseteq \Omega. \tag{1}$$

The probabilities are over the mechanism's internal randomness. The parameter ϵ bounds multiplicative distinguishability, and δ is additive slack, not an individual's probability of disclosure. Smaller values give stronger bounds. Our central guarantee covers outputs leaving a trusted release service; it does not protect that service's raw aggregate or the inference video. The Gaussian randomness is secret and is never published as a reproducible seed.

3.2 Overall Pipeline

Figure 1 summarizes PARD-VAD. Clients project and clip matched teacher-response differences for secure aggregation and a single Gaussian DP release. Public functions decode the released coefficients into temporal evidence. Temporal calibration and public-score-weighted pooling then yield video-level odds corrections, aligning cross-video scores while preserving within-video ordering.

3.3 Matched Teacher Responses

Let $\ell_0(X)_\tau$ be a public reference logit. A teacher f_i predicts

$$\ell_i(X)_\tau = \ell_0(X)_\tau + b_i(X) + h_i(X)_\tau, \qquad \sum_{\tau=1}^{L} h_i(X)_\tau = 0. \tag{2}$$

Here $b_i(X)$ is a scalar video-level calibration and $h_i(X)_\tau$ is the centered output of a temporal adapter. The video head takes pooled CLIP and public semantic features; the temporal head takes their aligned sequence. Both participate in local optimization, but only the temporal response is encoded for release. The teacher architecture need not have the same capacity as its communicated representation.

Each private teacher is trained on local video bags and public replay. The loss combines weak MIL classification, a positive–negative bag ranking term, a video-head binary classification term, sparsity,

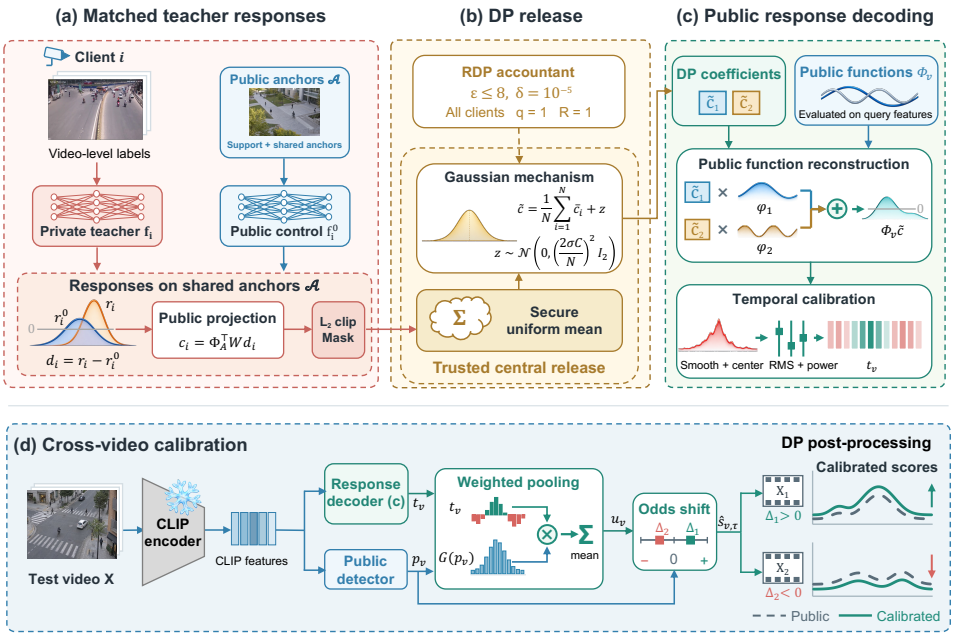

Figure 1: Overview of PARD-VAD. (a) Clients project and clip matched teacher-response differences on public anchors. (b) Secure aggregation and Gaussian noise produce DP coefficients. (c) Public functions decode temporal evidence. (d) Public-score-weighted pooling yields video-level odds shifts, aligning cross-video scores while preserving within-video ordering.

temporal smoothness, and squared residual penalties. All sampled windows from one video share a single bag label; an abnormal label is not assigned independently to every window. A matched teacher f_i^0, shown as the public control in Figure 1(a), uses public-only support with the same architecture, initialization seed, replay pool, and optimization schedule. Both teachers are evaluated on the same shared anchors. We encode each teacher using

$$r_i(X)_\tau = \mathcal{C}_X[\tanh(h_i(X)/a)]_\tau\,, \qquad d_i(X)_\tau = r_i(X)_\tau - r_i^0(X)_\tau, \tag{3}$$

where $a > 0$ is a public response scale, $\mathcal{C}_X$ subtracts a sequence's temporal mean, and r_i^0 is the corresponding response of f_i^0. Anchor subsampling is followed by recentering on the sampled positions. We compute d_i before projection and clipping, using the matched control to express each client's response relative to shared public training conditions.

3.4 Client-Level Private Public-Anchor Release

Public function basis. We train paired public surrogate teachers on disjoint public supports, with common public replay and matched initialization within each pair. Let $Y \in \mathbb{R}^{M\times J}$ collect their centered temporal-response differences, where J is the number of fitting pairs and M the total number of sampled anchor positions. A diagonal matrix $W \in \mathbb{R}^{M\times M}$ assigns equal mass to each anchor video and equal mass to its sampled positions, with $\mathrm{tr}(W) = 1$. The public second moment defines

$$Y^\top W Y = V\Lambda V^\top, \qquad T = V_{:,1:p}\,\mathrm{diag}\Big(\max(\lambda_j,\eta)^{-1/2}\Big)_{j=1}^{p}, \qquad \Phi_A = YT. \tag{4}$$

Here V contains eigenvectors, λ_j are the eigenvalues in descending order, $\Lambda = \mathrm{diag}(\lambda_1,\dots,\lambda_J)$, p is the retained rank, and $\eta > 0$ is a public eigenvalue floor. The columns of Φ_A represent nonlinear response functions constructed from public surrogates. Evaluating the same surrogate pairs on a query video X_v gives $Y_v \in \mathbb{R}^{L_v\times J}$ and $\Phi_v = Y_v T$, where L_v is its feature-sequence length.

One clipped release. Stack client i's temporal innovations over the sampled anchors into $d_i \in \mathbb{R}^M$. Project and clip the complete coefficient vector:

$$c_i = \Phi_A^\top W d_i \in \mathbb{R}^p, \qquad \bar{c}_i = c_i \min\left(1, \frac{C}{\|c_i\|_2}\right). \tag{5}$$

The zero vector is left unchanged. The radius $C > 0$ is selected using public surrogate reconstruction, independently of private update norms. All N clients participate once and contribute equally. The trusted aggregation-and-noise service in Figure 1(b) releases

$$\tilde{c} = \frac{1}{N}\sum_{i=1}^{N}\bar{c}_i + z, \qquad z \sim \mathcal{N}(0, s^2 I_p), \qquad s = \sigma\frac{2C}{N}, \tag{6}$$

where I_p is the $p \times p$ identity matrix, $\sigma > 0$ the noise multiplier, and s the standard deviation per coordinate. Client masking conceals individual vectors within the secure-aggregation abstraction, implemented here by simulated pairwise-mask cancellation. Raw responses, teacher weights, and the unperturbed mean remain inside the trusted boundary. The inference server receives only $\tilde{c}$ and public assets.

3.5 Privacy Accounting

Proposition 1 (conditional client-level privacy). With fixed public assets and $\|\bar{c}_i\|_2 \le C$, Eq. 6 has replace-one sensitivity at most $2C/N$. For each order $\alpha > 1$, it satisfies $(\alpha, \alpha/(2\sigma^2))$-RDP. For R separately released vectors with multipliers σ_r, $r \in \{1, \dots, R\}$, a valid bound is

$$\epsilon(\delta) = \min_{\alpha\in\Gamma}\left\{\sum_{r=1}^{R}\frac{\alpha}{2\sigma_r^2} + \frac{\log(1/\delta)}{\alpha - 1}\right\}, \tag{7}$$

where Γ is the chosen finite set of RDP orders, all greater than one. The proof is given in Appendix A. The final deployment uses $R = 1$ and full participation, applying the Gaussian bound directly. Budget ablations are separate releases and compose if disclosed together.

3.6 Public Response Decoding and Cross-Video Calibration

Response reconstruction. Figure 1(c) decodes the released vector as a weighted sum of fixed public response functions. The illustrated functions φ_1, φ_2 are the two retained basis functions; their values on query video X_v form the columns of Φ_v. Thus, Φ_A and Φ_v evaluate the same functions on anchors and query features, respectively. For query video v, define

$$h_v = \mathcal{C}_{X_v}\mathcal{S}_w[\Phi_v\tilde{c}], \qquad r_v = \left(L_v^{-1}\sum_{\tau=1}^{L_v}h_{v,\tau}^2\right)^{1/2}, \tag{8}$$

where $\mathcal{S}_w$ is a moving average of width w with replicated boundaries. To control amplitude differences, let r_* be the median of the positive r_v values in the unlabeled query collection. For $r_v > 0$, calibrate the response as

$$t_v = \mathcal{C}_{X_v}\left[\operatorname{sgn}(h_v)\left(\frac{|h_v|}{r_v}\right)^{\kappa} r_v^{1-\gamma} r_*^{\gamma}\right]. \tag{9}$$

The sign, absolute value, and power act elementwise; $\gamma \in [0, 1]$ controls amplitude sharing and $\kappa > 0$ controls response shape. The outer $\mathcal{C}_{X_v}$ recenters the response after the nonlinear power transform, ensuring that t_v has zero temporal mean. Zero responses remain zero. These operations constitute the temporal calibration in Figure 1(c), with reference scale r_* shared across the unlabeled query collection.

Public-score-weighted pooling. In Figure 1(d), frozen CLIP features feed both the response decoder from panel (c) and a fixed public detector B. Let $p_{v,\tau} = B(X_v)_\tau \in [0, 1]$ be its public-only recovery score. Here B combines public MIL scores with calibrated CLIP semantics, feature-derived temporal priors, affinity pooling, and public-only score calibration. Its dataset-specific components are summarized in Appendix C. This recovery branch is distinct from the teacher reference ℓ_0 and uses public assets and query features only. The identical B is retained in the matched no-private comparison. For nonnegative exponents α_g, β_g, define the normalized gate $G(p)$ and pooled evidence u_v:

$$G(p) = \frac{p^{\alpha_g}(1-p)^{\beta_g}}{\max_{x\in[0,1]} x^{\alpha_g}(1-x)^{\beta_g}}, \qquad u_v = \frac{1}{L_v}\sum_{\tau=1}^{L_v}G(p_{v,\tau})t_{v,\tau}. \tag{10}$$

An exponent of zero contributes a factor of one, including at the endpoints. The gate measures agreement between the decoded evidence and the public score profile. With a constant gate, the centered evidence averages to zero, so a nonconstant gate is necessary for this readout.

Final scoring. A pooling gain $\rho \geq 0$ and fusion strength $\lambda \geq 0$ yield one offset Δ_v for the entire video:

$$\Delta_v = a\lambda \, \mathrm{clip}(\rho u_v, -1, 1), \qquad \hat{s}_{v,\tau} = \frac{p_{v,\tau} \exp(\Delta_v)}{1 - p_{v,\tau} + p_{v,\tau} \exp(\Delta_v)}. \tag{11}$$

The scalar clip truncates to $[-1, 1]$. For probabilities strictly between zero and one, Eq. 11 multiplies the public odds by $\exp(\Delta_v)$: positive offsets raise scores and negative offsets lower them, as illustrated in Figure 1(d). The transformation preserves probabilities zero and one and exactly returns B when private evidence is disabled. Since Δ_v is constant within a video, it preserves temporal ordering while adjusting *cross-video ranking*. Temporal structure determines the correction through reconstruction and weighted pooling. Here, *calibration* means cross-video score alignment, not estimation of calibrated posterior probabilities. Define $\mathrm{logit}(p) = \log(p/(1-p))$ for $0 < p < 1$. For positions τ and τ' in videos v and v', respectively,

$$\mathrm{logit}(\hat{s}_{v,\tau}) - \mathrm{logit}(\hat{s}_{v',\tau'}) = \mathrm{logit}(p_{v,\tau}) - \mathrm{logit}(p_{v',\tau'}) + \Delta_v - \Delta_{v'}. \tag{12}$$

The offsets cancel within a video, but can change ordering across videos. A single common offset for every video would leave the complete-test ranking unchanged. Video-dependent offsets therefore provide the mechanism for global ranking correction. Both datasets use these same release and readout equations. Reference weights, vocabulary, public recovery configuration, and scalar hyperparameters can differ. Inference uses only the DP vector, fixed public assets, and the query collection, so it is post-processing with respect to training-client privacy.

4 Experiments

We organize the evidence around complete-test performance, the contribution of private knowledge and core components, privacy–utility sensitivity, and an exhaustive test of cross-video calibration. The main text reports these results under the stated deployment budget and illustrates the score adjustments with qualitative examples. Appendices provide detailed privacy records, implementation settings, alternative readouts, protocol checks, and supplementary ranking visualizations.

4.1 Experimental Settings

Datasets and Evaluation metric. We use CLIP features on two large datasets, UCF-Crime (UCF) (Sultani et al., 2018) and XD-Violence (XD) (Wu et al., 2020). From training videos, 10% are designated public and 5% public validation. This benchmark simulation treats the designated videos and labels as public auxiliary data outside the protected client collections. Remaining videos form 13 event-based UCF clients and six random XD clients. Following previous works(Qi et al., 2024; Wang et al., 2025), UCF-Crime uses frame-level ROC-AUC as the main metric, while XD-Violence uses frame-level average precision (AP).

Implementation details. Both datasets use frozen 512-dimensional CLIP features and width-64 teachers trained for 32 epochs. Eight public surrogate pairs fit the response basis, and four additional pairs select clipping through public reconstruction tasks. The release contains two temporal coordinates and no video-head coordinates. All clients participate once, targeting $\epsilon \leq 8$ at $\delta = 10^{-5}$. Exact clipping, noise, and accounting records are in Appendix A (Table 4). Training objectives, optimizer settings, and decoder hyperparameters are in Appendix B.

Shared mechanism and public references. The matched coefficient release, response-scale calibration, weighted pooling, and odds correction are shared across datasets. The teacher reference ℓ_0 and the stronger public recovery branch B are distinct. Their dataset-specific compositions are summarized in Appendix C. Each private-contribution comparison retains the identical B; the recovery ablation replaces this whole branch. Reference RMS uses the full unlabeled query collection, so evaluation is batch inference, not streaming.

Selection and intervention protocol. Each reported result uses a single realized release. Decoder interventions reuse the main vector; raw-response and matched-PCA alternatives are separately clipped and independently noised. Budget interventions change only the calibrated noise

Table 1: Complete-test comparison (%). Privacy targets and evaluation protocols differ. Our client-level DP bound is conditional on fixed public assets and applies to one deployment at $\delta = 10^{-5}$. † From TeD-SPAD.‡ Federated CLIP implementations from Fed-WSVAD.

Method	Privacy approach	UCF AUC	XD AP
Downsample-2×†	Resolution reduction	76.09	62.11
Obf-Blurring†	Person blurring	75.69	59.36
SPAct (Dave et al., 2022)†	Anonymization	73.93	53.36
TeD-SPAD (Fioresi et al., 2023)	Temporal anonymization	74.81	60.32
SPLAVU (Fioresi et al., 2026)	Latent anonymization	85.08	–
PPVU (Doshi & Yilmaz, 2023)	Transformer FL	82.90	–
Temporal-CLIP (FL)‡	Temporal FL	83.89	72.07
FedCoOp‡	Prompt FL	85.36	73.37
CLAP (Al-lahham et al., 2024)‡	Collaborative FL	84.99	70.54
FedVAD (Qi et al., 2024)	Distillation in FL	85.13	66.78
FedVAD + DP (Qi et al., 2024)	FL + perturbation	84.68	66.21
Fed-WSVAD (Wang et al., 2025)‡	Prompt FL	85.07	75.99
PPVAD (Li et al., 2026)	Federated VLM	87.25	79.86
FedVAR (Random) (Haider et al., 2026)	Prototype FL	86.36	75.20
FedVAR (Event) (Haider et al., 2026)	Prototype FL	84.97	82.82
PARD-VAD (Ours)	**Client DP,$\epsilon \leq 8$**	**89.19**	**84.03**

multiplier, keeping actual teacher contributions and the decoder fixed, with one draw per additional budget. Retuned alternatives are distinguished from fixed-parameter ablations in Appendix D.1. Appendix A gives the composition and scope of these fixed-deployment privacy bounds.

4.2 Comparison with State-of-the-Art Methods

Table 1 compares complete-test detection results with visual anonymization and federated approaches. The shared PARD-VAD release and decoder obtain 89.19 AUC on UCF-Crime and 84.03 AP on XD-Violence with the fixed public branches described above, exceeding all listed values. Compared with PPVAD, the margins are 1.94 and 4.17 points. Our XD result also exceeds FedVAR's strongest reported binary-detection AP of 82.82 by 1.21 points; that result uses Event clients, whereas ours uses Random clients. Both results use the same response representation and decoder under a conditional per-deployment budget of $\epsilon \leq 8$. Table 1 is a cross-paper comparison spanning different backbones, public supervision, client partitions, and privacy targets, rather than an equal-budget evaluation. To measure private transfer with the public detector held fixed, Table 2 includes the identical public-only control as "Without private knowledge." This matched comparison isolates the additional ranking benefit of the released coefficients (Section 4.3).

4.3 Ablation Studies

Table 2 fixes the final decoder parameters within each dataset. Setting the private correction to zero recovers the identical public baseline, reducing UCF AUC from 89.19 to 88.39 and XD AP from 84.03 to 83.49. The 0.81/0.54-point differences isolate the contribution of the realized private vector under the deployed decoder: public training data, recovery predictions, test videos, and ground truth are unchanged.

Response-scale calibration and smoothing. Removing scale calibration sets $(\gamma, \kappa) = (0, 1)$ and gives 88.29/83.29; removing temporal smoothing sets $w = 1$ while retaining centering and gives 83.93/80.82. Both operations affect the evidence before pooling, although the final correction is video-level. These fixed-parameter interventions show the contribution of scale control and temporal aggregation to the readout.

Public recovery and the gate. Replacing the complete public recovery branch with the raw public reference probability reduces performance to 84.67/60.39, measuring the contribution of the public

Table 2: Component ablations at $\epsilon \leq 8$, $\delta = 10^{-5}$, with fixed decoder settings. * Independent two-dimensional releases with public-selected clipping; not common-noise comparisons. Other rows reuse the main DP vector.

Configuration	UCF AUC	XD AP
Full method	**89.19**	**84.03**
Without private knowledge	88.39	83.49
Raw-response release*	88.05	82.52
Matched CLIP-PCA basis*	87.49	82.21
Without response-scale calibration	88.29	83.29
Without temporal smoothing	83.93	80.82
Without complete public recovery	84.67	60.39

Table 3: Budget sensitivity with fixed decoder settings ($\delta = 10^{-5}$). Each private row is one realized release with $q = 1$ and $R = 1$; $\epsilon = 8$ reuses the main vector.

Target ϵ	σ	UCF AUC	XD AP
3	1.717961	89.16	72.29
5	1.085479	87.79	68.49
8	0.693179	89.19	84.03
16	0.382667	89.11	83.27
Public-only	–	88.39	83.49

branch as a whole. Setting $(\alpha_g, \beta_g) = (0, 0)$ makes the gate constant. Since the decoded response is centered, its average is zero and this intervention exactly reproduces the no-private row, as verified in the saved arrays. Thus, the nonconstant gate converts centered temporal evidence into a video-level correction.

Release representation. Raw-response and matched CLIP-PCA releases yield 88.05/82.52 and 87.49/82.21, respectively, under the main decoder settings. Each uses separate public-selected clipping and an independent Gaussian draw, so these are comparisons of complete realized release designs. Separately retuned alternatives reach 89.21 UCF AUC with a local residual and 84.11 XD AP with a raw-response design. We retain one pooled rule across datasets; full alternatives and a sign-reversal stress test are in Appendix D.1.

4.4 Privacy Budget and Protocol Validation

Table 3 changes only the calibrated noise multiplier while fixing teachers, actual client contributions, basis, clipping radius, public scores, and decoder parameters. The main $\epsilon = 8$ vector is reused; each other budget uses one independent draw, without redrawing or retuning. At the main budget, private knowledge improves both datasets. At $\epsilon = 3/5$, XD falls to 72.29/68.49 AP, showing substantial noise sensitivity of the frozen readout. These single-draw results characterize realized budget sensitivity rather than expected utility; the public-only row gives the no-transfer reference. The corresponding plot is in Appendix D.2. Disclosing all four budget vectors composes to $\epsilon = 20.903040$ at $\delta = 10^{-5}$; decoder-only reuse adds no cost. Full participation permits direct Gaussian accounting. Exact release and composition records, mechanism tests, and replay checks are in Appendix A.

4.5 Cross-Video Calibration Analysis

Across all 290 UCF and 800 XD test videos, every within-video strict ordering and tie is preserved, as predicted by Eq. 12. Figure 2 audits all abnormal–normal frame pairs from different videos; compact metric summaries report the corresponding full-test UCF AUC and XD AP changes. Cross-video corrections outnumber regressions on both datasets. UCF's 0.696 billion net corrections exactly account for the 0.807-point AUC gain. XD likewise shows a net ordering improvement, alongside an independently computed AP increase from 83.492 to 84.027 percent. With the public detector fixed and within-video ordering preserved, these gains arise from cross-video alignment. Counting details and supplementary examples are provided in Appendices D.3 and E.

4.6 Qualitative Results

Figure 3 presents abnormal and normal examples from both datasets. In UCF (a), a negative offset lowers elevated background scores while retaining high responses over much of the event. In XD (c), a positive offset raises scores while preserving the three event-associated peaks. Normal examples (b,d) receive negligible corrections. Signed decoded evidence is pooled into one offset per

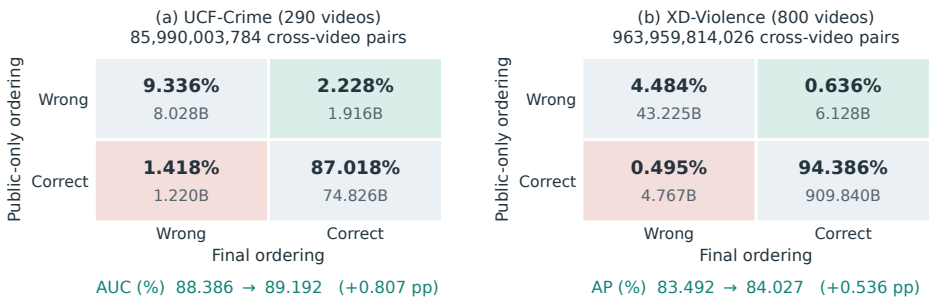

Figure 2: Exhaustive cross-video ordering audit. Cells show percentages and counts (B: billion); green/red denote corrected/worsened pairs and gray unchanged orderings. Empty tie rows and columns are omitted. Below: independently computed full-test UCF AUC and XD AP, public-only → final, with gains in percentage points (pp). Pair counts do not decompose AP.

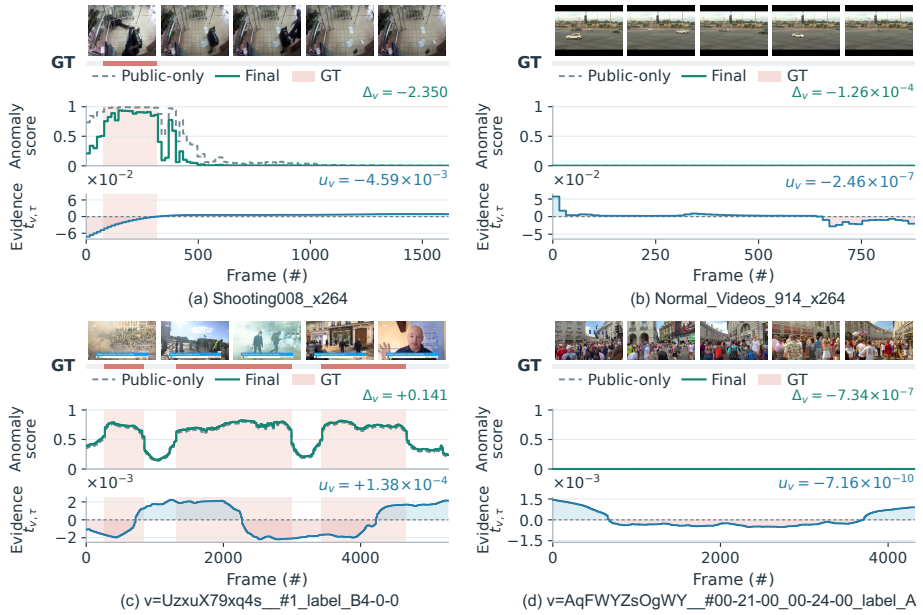

Figure 3: Qualitative examples on UCF (a,b) and XD (c,d), with five uniformly spaced keyframes and frame-level ground truth (GT). Gray dashed/green solid curves denote public-only/final scores on the same $[0, 1]$ scale; red shading marks abnormal intervals. Lower plots show signed decoded evidence $t_{v,\tau}$; annotations give pooled evidence u_v and the video-level log-odds shift Δ_v.

video, adjusting score levels while preserving within-video ordering. These examples illustrate the calibration mechanism examined by the cross-video ranking audit in Figure 2.

5 Conclusion

PARD-VAD transfers matched temporal-response innovations through one client-private vector to calibrate scores across videos. On complete test sets, the shared mechanism achieves 89.19 UCF AUC and 84.03 XD AP, improving identical public controls by 0.81 and 0.54 points at a per-deployment budget of $\epsilon \leq 8$, conditional on fixed public assets. Exhaustive audits show net cross-video ordering improvements with unchanged within-video ranks. Budget experiments identify noise sensitivity, especially on XD, as a direction for improving the readout.

AI USE STATEMENT

In this work, we used generative AI tools to assist with polishing the manuscript, conducting literature reviews, and examining whether the ideas proposed in this paper overlap or conflict with previously published work. We also used generative AI tools to assist with the development of portions of the code. We did not use generative AI tools to propose or refine hypotheses or to formulate mathematical claims. The generation of synthetic datasets is not applicable to this work.

We have reviewed all AI-assisted work. We manually conducted literature reviews to check whether research ideas generated or suggested by large language models involved potential plagiarism or substantial overlap with existing work, and we verified the correctness and authenticity of the references provided by these tools. We take full responsibility for the final content of this work, including any text, claims, code, or other artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

Our work strictly adheres to the ICLR Code of Ethics. This research is conducted exclusively on publicly available, standard academic benchmark datasets, namely UCF-Crime and XD-Violence, which have been widely used for evaluating video anomaly detection methods. We acknowledge the potential dual-use nature of video anomaly detection technology. While such techniques can contribute to socially beneficial applications, including enhancing public safety through intelligent surveillance and improving industrial quality control, anomaly detection systems may also enable intrusive monitoring and may reflect or amplify biased definitions of what constitutes "abnormal" behavior.

Moreover, the client-level differential privacy (DP) mechanism considered in our work does not eliminate these broader societal risks, nor does it anonymize individuals appearing in videos at inference time. Our privacy guarantee relies on the assumption of a trusted aggregation-and-noise service and fixed public auxiliary data. It does not protect against malicious data poisoning, disclosure of raw aggregated updates, or leakage of the randomness used for noise generation. In addition, the secure-aggregation simulator employed in our experiments is intended solely for research evaluation and should not be interpreted as a production-ready cryptographic protocol. Our primary motivation is to advance video anomaly detection in a manner that supports socially beneficial applications while making the scope, assumptions, and limitations of the proposed privacy protections explicit.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 describe the release-and-calibration method, data partitions, and evaluation protocol. Appendix A provides the sensitivity proof, noise calibration, RDP accounting, and composition bounds. Appendix B specifies the teacher architecture, optimization settings, public basis construction, and decoder hyperparameters; Appendix C summarizes the public detector components and matched controls. Experiments use UCF-Crime and XD-Violence with frozen CLIP features and documented versions of PyTorch, scikit-learn, and OpenDP. Independent frame-level evaluation verifies the reported detection metrics, while 130 focused tests check the privacy mechanism and decoder. Appendix D.3 specifies the exhaustive cross-video counting procedure and its metric consistency checks. Inference can be replayed from the released noisy coefficients and fixed public assets without disclosing private teacher parameters or noise randomness.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318, 2016.

Anas Al-lahham, Muhammad Zaigham Zaheer, Nurbek Tastan, and Karthik Nandakumar. Collaborative learning of anomalies with privacy (clap) for unsupervised video anomaly detection: A new baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, pp. 1175–1191, 2017.

Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. MGFN: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 387–395, 2023.

Ishan Rajendrakumar Dave, Chen Chen, and Mubarak Shah. 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.

Keval Doshi and Yasin Yilmaz. Privacy-preserving video understanding via transformer-based federated learning. In *IEEE Conference on Dependable and Secure Computing*, pp. 1–8, 2023. doi: 10.1109/DSC61021.2023.10354099.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pp. 265–284. Springer, 2006.

Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. MIST: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14009–14018, 2021.

Joseph Fioresi, Ishan Rajendrakumar Dave, and Mubarak Shah. 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.

Joseph Fioresi, Ishan Rajendrakumar Dave, and Mubarak Shah. Privacy beyond pixels: Latent anonymization for privacy-preserving video understanding. In *International Conference on Learning Representations*, 2026.

Jonas Geiping, Hartmut Bauermeister, Hannah Droge, and Michael Moeller. Inverting gradients: How easy is it to break privacy in federated learning? In *Advances in Neural Information Processing Systems*, volume 33, pp. 16937–16947, 2020.

Ghani Haider, Majid Kundroo, Boyun Eom, Dong-Hwan Park, Chen Chen, and Taehong Kim. Fedvar: Prototype-aligned federated framework for video anomaly recognition. *Future Generation Computer Systems*, pp. 108745, 2026.

Daliang Li and Junpu Wang. FedMD: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581*, 2019. URL `https://arxiv.org/abs/1910.03581`.

Shuo Li, Fang Liu, Licheng Jiao, Jiahao Wang, Yanbiao Ma, Xinyi Wang, Yang Liu, Qianyue Bao, Xu Liu, Lingling Li, and Puhua Chen. Privacy-preserving video anomaly detection via federated learning. *Knowledge-Based Systems*, 349:116454, 2026. doi: 10.1016/j.knosys.2026.116454.

Ilya Mironov. Renyi differential privacy. In *Proceedings of the IEEE Computer Security Foundations Symposium*, pp. 263–275, 2017.

Nicolas Papernot, Martin Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1610.05755.

Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Ulfar Erlingsson. Scalable private learning with PATE. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1802.08908.

Fan Qi, Ruijie Pan, Huaiwen Zhang, and Changsheng Xu. FedVAD: Enhancing federated video anomaly detection with GPT-driven semantic distillation. In *Proceedings of the European Conference on Computer Vision*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pp. 8748–8763, 2021.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *Proceedings of the IEEE Symposium on Security and Privacy*, pp. 3–18, 2017.

Yong Su, Haohao Zhu, Yuyu Tan, Simin An, and Meng Xing. Prime: Privacy-preserving video anomaly detection via motion exemplar guidance. *Knowledge-Based Systems*, 278:110872, 2023.

Yong Su, Jiahang Li, Simin An, Meng Xing, and Zhiyong Feng. Federated weakly-supervised video anomaly detection with mixture of local-to-global experts. *Information Fusion*, 123:103256, 2025. doi: 10.1016/j.inffus.2025.103256.

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

Benfeng Wang, Chao Huang, Jie Wen, Wei Wang, Yabo Liu, and Yong Xu. Federated weakly supervised video anomaly detection with multimodal prompt. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21017–21025, 2025.

Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Proceedings of the European Conference on Computer Vision*, pp. 322–339, 2020.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

A PRIVACY PROOF AND SCOPE

Sensitivity. Fix the public pool, reference models, surrogate functions, matched controls, transform matrices, clipping radius, client indices, and training algorithm. Under $D \sim D'$, only one client's trained teacher and projected vector can change. For that index c, both clipped vectors lie in the radius-C ball, hence

$$\left\| \frac{1}{N} \sum_i \bar{c}_i(D_i) - \frac{1}{N} \sum_i \bar{c}_i(D'_i) \right\|_2 = \frac{1}{N} \|\bar{c}_c(D_c) - \bar{c}_c(D'_c)\|_2 \leq \frac{2C}{N}. \tag{13}$$

Clipping the complete projected difference bounds each client's contribution, including the matched subtraction. The two temporal coordinates share one vector-level L2 sensitivity. Local optimization may be arbitrarily data dependent within a client. Its randomization is independent of the secret central noise.

Gaussian RDP. For a fixed training-randomness realization, adjacent mean vectors μ, μ' differ by at most $\Delta = 2C/N$. Equal-covariance Gaussian outputs satisfy

$$D_\alpha\big(\mathcal{N}(\mu, s^2 I) \,\|\, \mathcal{N}(\mu', s^2 I)\big) = \frac{\alpha \|\mu - \mu'\|_2^2}{2s^2} \leq \frac{\alpha}{2\sigma^2}, \tag{14}$$

where D_α is Rényi divergence of order $\alpha > 1$ and $s = \sigma\Delta$. The bound is uniform over training randomness and survives its marginalization. Adaptive RDP composition adds the per-release bounds; conversion gives Eq. 7. Public decoder construction, fixed-input inference, and interventions on the released vector are post-processing. This proves Proposition 1 under the declared central trust model.

Numerical accountant. The order set is $\Gamma = \{1.25, 1.5, 2, 3, 4, 5, 8, 16, 32, 64, 128, 256\}$. Bisection calibrates σ before accessing clipped client contributions, with a relative 10^{-6} budget margin. The implementation validates the OpenDP Gaussian privacy map at the supplied sensitivity and checks the recorded RDP value. The main experiment uses $q = 1$ and $R = 1$, with the unsampled Gaussian RDP bound. The per-dataset certificates are summarized in Table 4.

Table 4: Conditional single-deployment privacy records for the main method.

Quantity	UCF	XD
Released temporal coordinates	2	2
Participating clients m/N	13/13	6/6
Releases R / participation q	1 / 1	1 / 1
Clip radius C	0.0407310144	0.0539423071
Noise multiplier σ	0.6931794682	0.6931794682
Coordinate noise SD s	0.0043436774	0.0124638999
Accounted ϵ	7.999992	7.999992
δ	10^{-5}	10^{-5}

Trust model and scope. The theorem conditions on a fixed public pool outside client adjacency. In the benchmark simulation, designated training-source videos and their labels are treated as public, with all their crops excluded from private clients. Constructing this pool from initially protected data would require a separate privacy mechanism. The central noise is hidden from the inference server. Experiments simulate secure aggregation; deployment requires a protocol that implements masking and dropout handling under the stated honest-execution and non-collusion assumptions.

Multiple experiments. The main deployment has a bound of $\epsilon \leq 8$, and jointly observed releases accumulate privacy loss. Observing the four budget vectors in Table 3 has a composed bound of $\epsilon = 20.903040$ at $\delta = 10^{-5}$, for either dataset. Adding the independent raw-response and matched-PCA vectors gives the bound in Table 5; decoder-only reuse adds no release cost. The table accounts for the specified outputs, while earlier releases remain in the cumulative ledger. The theorem conditions on fixed mechanism and public-design choices: an end-to-end bound for their adaptive selection would also need to account for the historical private outputs informing those choices. Historical selection is therefore outside the reported fixed-deployment bound.

Table 5: Composition of the specified released vectors, per dataset, at $\delta = 10^{-5}$. This does not include the entire research history.

Observed outputs	Vectors	ϵ bound
Main deployment	1	7.999992
Main + matched-PCA	2	11.999988
Four budget points	4	20.903040
Budget points + raw response + matched-PCA	6	25.772994

Numerical validation. We validate the OpenDP noise map against the analytic sensitivity $2C/N$ and check the corresponding privacy bounds. The 130 focused tests cover clipping, accounting, secure-aggregation cancellation, zero-response behavior, score ties, and decoder behavior. Independent frame-expanded scikit-learn evaluation reproduces all main, component, and budget metrics, verifying consistency between the mechanism, predictions, and reported results.

B Implementation and Public Reference

Training configuration. Both datasets train width-64 teachers for 32 epochs with eight steps per epoch, using eight local and eight public-replay bags per step. Each bag uses four native-cadence windows of up to 64 feature positions. AdamW starts at learning rate 3×10^{-4}. Frozen CLIP features have dimension 512. Eight public surrogate pairs fit the response basis and four additional pairs supply public reconstruction tasks for clipping selection. The final release retains two temporal coordinates and no video-head coordinates. All clients participate once: $m/N = 13/13$ on UCF and $6/6$ on XD.

Shared architecture. The frozen CLIP feature dimension is 512. Public class logits contain 14 classes on UCF and seven on XD. Temporal inputs concatenate per-position layer-normalized CLIP features with $\tanh(\text{class logits}/8)$. The temporal adapter has width 64, three dilated convolution blocks with dilations 1, 2, and 4, dropout 0.15, and a bounded residual. A separate video head is LayerNorm–Linear(64)–GELU–Dropout–Linear(1), with zero-initialized final weights and output $2\tanh(\cdot)$. Its input concatenates mean layer-normalized CLIP features scaled by $1/\sqrt{512}$, transformed mean and top-12.5% class logits, and a constant coordinate. The vocabulary sizes give input dimensions of 541/527 for UCF/XD within the same architecture.

Local objectives. Define $P(X)$ as the mean of the highest 12.5% snippet probabilities in a sampled bag. For abnormal bags, the classification loss is $-\log P(X)$; for normal bags it is the mean snippet-level negative log-likelihood. A hinge term averages $[0.2 - P(X^+) + P(X^-)]_+$ over positive–negative bag pairs with weight 0.2. The video-head loss, weighted 0.5, applies binary cross-entropy to the public full-video top-k logit plus $b_i(X)$. Positive-bag sparsity and adjacent-window probability smoothness both use weight 0.002; mean squared temporal and video residuals each use weight 0.01. Every step sums a local-support loss and a public-replay loss. AdamW uses weight decay 0.001 and cosine decay to 3×10^{-5}; gradient norms are clipped at five. This optimizer clipping is distinct from release clipping C.

Public function construction. At most 64 evenly spaced positions per public source video are used for encoding, with temporal recentering on the sampled grid. Eight public surrogate pairs fit the basis and four additional pairs supply public validation tasks. UCF surrogate roles follow public anomaly categories; XD uses fixed random public-support roles. Pairs share initialization and replay but have disjoint local support; the common pretrained public reference may already have seen both public subsets. The eigenvalue floor is $\max(10^{-12}, 0.001\lambda_1)$. Public calibration evaluates ranks $\{1, 2, 4, 8\}$ and clipping quantiles $\{0.5, 0.75, 0.9, 1\}$, with radius floored at 0.001. It measures held-out response reconstruction using 256 synthetic public-client mixtures and public Gaussian perturbations. UCF selects rank two; final XD also uses rank two for representation consistency, although its earlier public-MSE optimum was rank one. At fixed rank two, both final radii select quantile 0.5. Rank reconciliation was made after the historical experiments.

Public reference and recovery. The reference ℓ_0 supplies teacher inputs and baseline logits, whereas B is the stronger, frozen public inference branch. UCF retains its public semantic adapter,

semantic/contextual recovery, and public video-calibration branch. XD uses a public semantic adapter with six event classes for the hierarchical teacher; its recovery retains the public weak-supervision ensemble and semantic/temporal readout from the zero-private configuration. The release encoder and final odds equation are shared, while these reference assets and internal public readout settings differ. We report public-only performance alongside each private correction to measure its additional contribution.

Decoder settings. Table 6 gives the exact final settings. The reference RMS is computed over all nonzero video responses in the full unlabeled query collection. If every response is zero, the private correction is zero. Removing scale calibration sets $(\gamma, \kappa) = (0, 1)$; removing smoothing sets $w = 1$ and retains centering. Removing the gate sets $(\alpha_g, \beta_g) = (0, 0)$, giving zero pooled centered evidence by construction. The exponential in Eq. 11 is numerically bounded to $[-60, 60]$ before exponentiation; main offsets are already bounded by $a\lambda = 8$. The public branch is unchanged for all private-contribution comparisons. No covariance shrinkage or local framewise residual is used in the final decoder.

Table 6: Final decoder settings, fixed across component and budget interventions.

Parameter	UCF	XD
Window w (feature steps)	65	97
Amplitude exponent γ	1	0.5
Response power κ	4	0.5
Gate α_g	1	1
Gate β_g	0	2
Pooling gain ρ	64	128
Fusion strength λ	2	2
Response scale a	4	4

Matched evaluation. Private and public-control teachers use matched initialization and training schedules. All core decoder ablations use the main hyperparameters without retuning. PCA and raw-response comparisons test complete alternative release designs using distinct independently noised vectors and separately public-calibrated clipping; the PCA representation encodes the matched response difference. The budget curve keeps the actual main contributions, radius, and decoder fixed. Its $\epsilon = 3, 5, 16$ vectors are each drawn once; the main $\epsilon = 8$ vector is reused.

Software. Release generation and evaluation use Python 3.10.20, PyTorch 2.13.0, NumPy 2.2.6, scikit-learn 1.7.2, and OpenDP 0.14.0. Secure Gaussian sampling is internal to OpenDP; public training seeds do not determine the private noise.

C Public Detector and Matched Controls

The public detector B supplies the baseline probabilities in Eq. 10. It combines publicly trained predictors with deterministic transformations of query features and is fixed before applying the private correction. It is distinct from the teacher reference ℓ_0, which provides inputs and baseline logits for local teacher training. Neither branch consumes the released private coefficients.

Shared components. Both public detectors combine weakly supervised MIL predictions, calibrated CLIP text–video similarities, feature-derived temporal priors, and within-video affinity pooling. MIL and semantic predictors use the designated public support and validation videos. Semantic scores compare event and normal prompts; temporal priors summarize feature magnitudes and temporal differences. Affinity pooling uses similarities between normalized CLIP features within each video. These operations produce the public probability sequence used by the private pooling gate.

Dataset-specific components. UCF combines magnitude and first-difference priors with public MIL and semantic auxiliaries, affinity refinement, and public-only contextual and video-level calibration. XD combines first- and second-difference priors with public MIL and temporal predictions, affinity pooling, and an auxiliary MIL ensemble. The UCF contextual and video-level corrections are fitted using public-only controls and are part of B, not the newly released private signal. Thus, public compositions differ across datasets, while the matched private release, response decoder, and

final odds-correction rule are shared. Component weights and temporal windows are fixed within each dataset for the main, decoder-ablation, and privacy-budget evaluations.

Matched comparisons. The public-only detectors achieve 88.39 AUC on UCF and 83.49 AP on XD. The no-private intervention holds the complete public detector fixed and sets the private offset in Eq. 11 to zero. The public-recovery ablation instead replaces the entire B with the sigmoid of the raw reference logit ℓ_0, retaining the private release and decoder settings. The former isolates the additional contribution of private knowledge; the latter measures the contribution of the complete public branch.

D Additional Experimental Analyses

D.1 Alternative Releases and Readouts

Table 7 separates fixed-parameter release comparisons from independently retuned alternatives. Raw-response and matched-PCA designs each use their own public-selected clipping radius and independently drawn two-dimensional Gaussian release, allowing comparisons of complete realized mechanisms.

Table 7: Retained alternative designs and a sign-reversal stress test. Retuned alternatives are not single-component ablations.

Configuration	UCF AUC	XD AP
Matched, pooled-only (main)	89.19	84.03
Matched, local + pooled	89.21	84.02
Raw response, main parameters	88.05	82.52
Raw response, prior parameters, pooled-only	88.54	84.00
Raw response, retuned local + pooled	88.64	84.11
Negated main vector (stress test)	86.75	81.38

Local residual versus pooled correction. Adding a local framewise residual to the matched release produces 89.21/84.02, changing UCF by $+0.0206$ points and XD by -0.0041. We omit this extra term on both datasets to keep one pooled correction rule. A raw-response alternative with separately tuned local and pooled terms reaches 88.64/84.11: it is lower on UCF, but exceeds the main XD AP by 0.0783 points. Removing its local term while retaining the previously selected parameters gives 88.54/84.00. These alternatives show the trade-off between dataset-specific readouts and the shared pooled design.

Sign-reversal stress test. Negating the already released main vector, with no new private query, gives 86.75 UCF AUC and 81.38 XD AP. These are below both the main result and the corresponding public-only controls. This deterministic intervention shows that the direction of the main private signal matters to the decoder.

D.2 Budget Sensitivity Visualization

Figure 4 visualizes every budget point in Table 3. Teachers, actual client contributions, the public basis, clipping radius, and all decoder settings are fixed. The main vector is reused at $\epsilon = 8$; each other point uses one independent draw recorded in the ledger. The nonmonotone single-draw sequence illustrates the frozen decoder's sensitivity to the released noise.

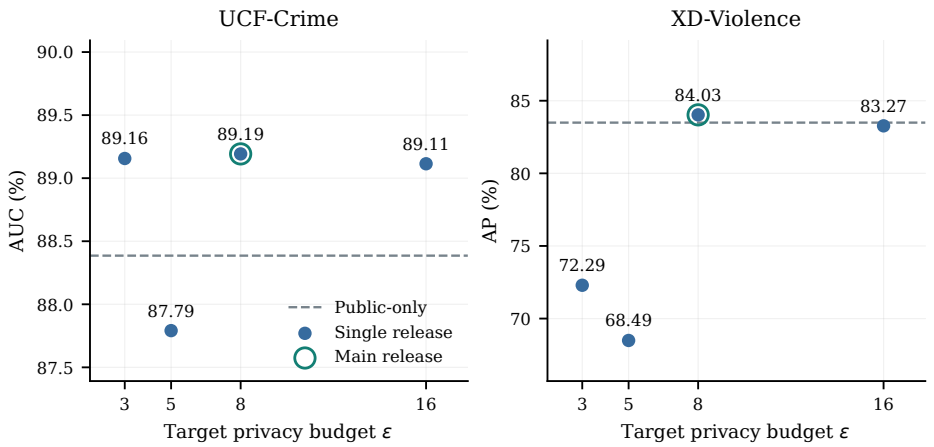

Figure 4: Budget sensitivity under the fixed final decoder. Each point is one release at $\delta = 10^{-5}$; the circled point reuses the main release. Dashed lines are the identical public-only controls.

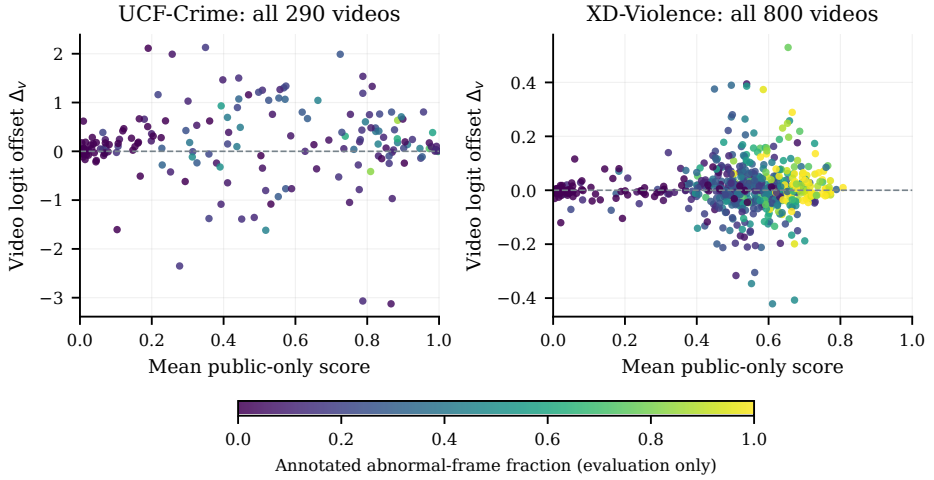

Figure 5: Video-level calibration across the full test set. Each point is one video; the dashed line denotes no correction. Color gives the ground-truth abnormal-frame fraction for analysis only.

D.3 Exact Cross-Video Counting and Metric Consistency

Within-video order audit. We reconstruct each video's constant logit offset from its public and final probabilities and verify Eq. 11 to an absolute probability tolerance of 5×10^{-13}. Across all 290 UCF and 800 XD videos, every within-video strict ordering and tie is preserved, verifying Eq. 12 on both datasets. Changes in complete-test ranking therefore occur between videos.

Exhaustive cross-video ordering audit. We count every pair consisting of a GT-abnormal frame and a GT-normal frame from different videos. A pair is correct when the abnormal frame has the higher score. Figure 2 records all public-to-final transitions, including unfavorable ones. Counting uses the exact positive and negative frame multiplicities within each 16-frame score block, with same-video pairs removed; it does not sample pairs or round scores to create ties. The implementation is checked against brute-force enumeration on 160 small weighted cases with mixed labels and ties.

On UCF, 1.916 billion pairs change from wrong to correct and 1.220 billion change from correct to wrong, yielding 0.696 billion net corrections. XD similarly has 6.128 billion corrections and 4.767

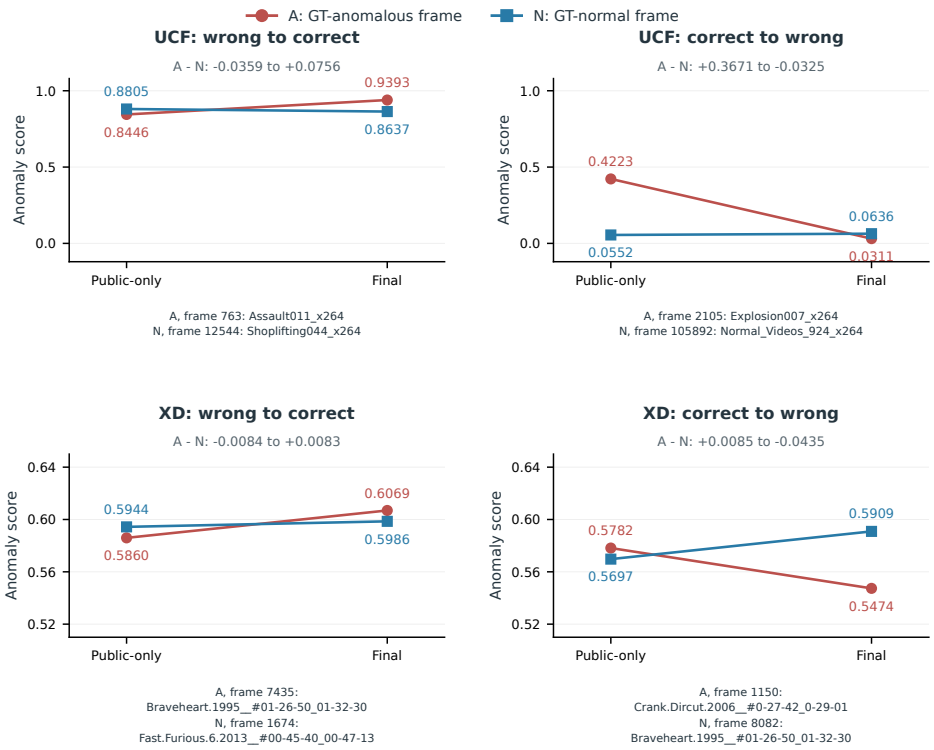

Figure 6: Cross-video rank reversals in UCF (top) and XD (bottom). Each panel compares a GT-abnormal frame (A, red) and a GT-normal frame (N, blue) from different videos. Left panels show corrected orderings; right panels show newly introduced errors. A–N denotes their anomaly-score difference. Numeric labels give the stored scores; the XD score axes are zoomed. Frame indices are zero-based.

billion regressions, for a net 1.361 billion. No cross-video ties occur in these saved predictions. On UCF, let K_{wc} and K_{cw} denote these corrected and worsened counts, and let F_+ and F_- be the total numbers of abnormal and normal test frames. Since within-video order is unchanged, with AUC expressed on $[0, 1]$,

$$\mathrm{AUC}_{\mathrm{final}} - \mathrm{AUC}_{\mathrm{public}} = \frac{K_{\mathrm{wc}} - K_{\mathrm{cw}}}{F_+ F_-}. \tag{15}$$

The denominator includes all positive–negative frame pairs, including same-video pairs whose contribution to the change is zero. This identity exactly recovers the 0.807-point UCF AUC increase, from 88.386 to 89.192 percent. XD is evaluated only with AP: independent full-test computation gives 83.492 to 84.027 percent, a 0.536-point gain. Its pair counts explain the direction of ordering changes but are not an AP decomposition. Shared frames make these pairs dependent; their large counts are not independent sample sizes for significance claims.

E Supplementary Ranking Visualizations

E.1 Video-Level Offsets

Video-dependent corrections. Figure 5 shows every video's offset Δ_v against its mean public score. UCF contains 182 positive and 108 negative offsets; XD contains 415 positive and 385 negative offsets. These video-dependent adjustments can change cross-video rankings, whereas a common monotone remapping would preserve them. Annotation colors show ground-truth abnormal-frame fractions for retrospective analysis; labels do not enter the offset computation. In mixed-label videos, either offset direction affects both normal and abnormal frames.

E.2 Illustrative Cross-Video Rank Reversals

Rank reversals. Figure 6 shows one corrected and one worsened ordering per dataset. In UCF, calibration corrects an abnormal `Assault011` frame initially ranked below a normal `Shoplifting044` frame; XD shows a smaller correction. Examples are selected near the median reversal margin in a fixed sample of 300,000 positive–negative frame pairs per dataset, excluding same-video pairs, illustrating both directions counted in Figure 2.