LEARNING TO FORGET UNDESIRED ANOMALIES: CATEGORY UNLEARNING FOR WEAKLY SUPERVISED VIDEO ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Weakly supervised video anomaly detection (WSVAD) localizes anomalous events using only video-level labels. However, models trained under this paradigm are often deployed in various real-world scenarios, where certain categories previously treated as anomalies may no longer be considered anomalous. This issue is particularly challenging because WSVAD lacks frame-level category supervision, while many recent state-of-the-art methods couple a class-agnostic binary branch with a fine-grained classification branch. Naively masking a category output suppresses only its fine-grained prediction, leaving the shared binary anomaly evidence largely unaffected. To address this limitation, we propose Category-Conditioned Dual-Branch Modulation (C-DBM), a lightweight adapter that enables on-demand unlearning of specific anomaly categories for frozen WSVAD detectors. Given a set of retained categories, C-DBM employs category-conditioned cross-attention to extract category-specific temporal information and modulates both the fine-grained representations and the binary anomaly logits accordingly. After a one-time adapter training stage, different unlearning requests can be satisfied simply by updating the retained-category state, without any request-time optimization or access to the original training data. Experiments across four base-model–dataset combinations show that C-DBM supports both single- and multi-category unlearning and substantially reduces residual fine-grained responses for forgotten categories. Moreover, compared with other unlearning baselines, C-DBM more effectively suppresses forgotten-category predictions, while preserving both binary and fine-grained detection performance under the relabeled full-test protocol. Overall, C-DBM demonstrates effective anomaly-category unlearning while maintaining performance across both branches of modern WSVAD frameworks.

1 INTRODUCTION

Video anomaly detection (VAD) aims to temporally localize anomalous events in long, untrimmed videos, while weakly supervised VAD (WSVAD) reduces annotation cost by learning from video-level labels under the multiple-instance learning (MIL) paradigm (Sultani et al., 2018; Wu et al., 2020; Feng et al., 2021; Tian et al., 2021; Chen et al., 2023). Conventional WSVAD primarily produces category-agnostic anomaly scores, whereas recent vision–language approaches increasingly incorporate category-level semantics via textual priors, fine-grained alignment, and category-aware modeling (Yang et al., 2024; Chen et al., 2024; Wu et al., 2024; Huang et al., 2025; Wang & Chen, 2025; Yin et al., 2026). In particular, modern dual-branch detectors can simultaneously provide category-agnostic anomaly evidence and fine-grained category predictions (Wu et al., 2024; Wang & Chen, 2025; Yin et al., 2026), making anomaly categories an explicit part of model behavior.

However, what constitutes an anomaly depends on the deployment context and may change over time (Sultani et al., 2018; Zanella et al., 2024). As illustrated in Figure 1, the same behavior may be anomalous in one environment yet legitimate in another; evolving privacy or data-governance requirements may likewise require the learned influence of a particular anomaly category to be removed (Ginart et al., 2019; Sekhari et al., 2021). Representative WSVAD detectors, however, are not designed to selectively forget a learned anomaly category while preserving the remaining ones

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.

(Wu et al., 2024; Huang et al., 2025; Wang & Chen, 2025; Yin et al., 2026). Machine unlearning provides a natural framework for such removal (Cao & Yang, 2015; Ginart et al., 2019; Bourtoule et al., 2021; Sekhari et al., 2021), but modern dual-branch WSVAD simultaneously models category-specific semantics and shared category-agnostic anomaly evidence (Wu et al., 2024; Wang & Chen, 2025; Yin et al., 2026). Consequently, suppressing a target fine-grained output alone cannot remove its contribution to the shared binary anomaly response.

Existing machine-unlearning methods are predominantly developed for settings in which the forget samples or classes are explicitly identifiable and can be directly optimized through retraining, distillation, gradient updates, or parameter modification (Bourtoule et al., 2021; Kurmanji et al., 2023; Chundawat et al., 2023; Fan et al., 2024; Foster et al., 2024). Applying class unlearning to WSVAD instead raises three distinctive challenges: *latent temporal attribution*, because target-category snippets are unknown under video-level supervision (Sultani et al., 2018; Tian et al., 2021; Chen et al., 2023); *retain-aware MIL supervision*, because retained and forgotten anomalies may coexist within the same video (Wu et al., 2020); and *dual-branch consistency*, because category-specific semantics and shared binary anomaly evidence must be modified jointly without suppressing retained anomalies (Wu et al., 2024; Yin et al., 2026). We therefore study anomaly-category unlearning in WSVAD, aiming to withdraw a forgotten anomaly category while preserving retained-anomaly detection.

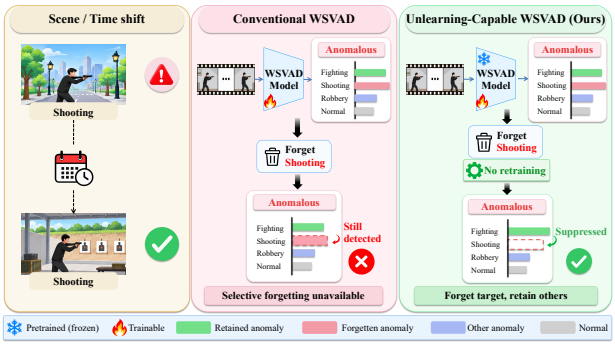

Figure 1: Motivation for anomaly-category unlearning in WSVAD. Left: Anomaly status may change across scenes or over time. Middle: Conventional WSVAD methods lack selective class withdrawal. Right: Our unlearning-capable VAD detector can suppress the target category while preserving retained anomalies.

To address these challenges, we propose Category-Conditioned Dual-Branch Modulation (C-DBM), a lightweight adapter for frozen WSVAD detectors. C-DBM uses category-conditioned cross-attention to obtain per-category temporal states, trains them under retain-aware MIL supervision, and coordinates fine-grained feature modulation with retained-category binary evidence routing. The base WSVAD model remains frozen, and after adapter training, different category-unlearning requests are executed by changing only the retain state without request-specific optimization.

Our main contributions are summarized as follows:

- To the best of our knowledge, we are the first to study anomaly-category unlearning in WSVAD. We formulate the problem under weak video-level MIL supervision, where a requested anomaly category must be selectively forgotten while preserving the detection of retained anomalies across both category-specific and category-agnostic outputs.
- We propose C-DBM, a lightweight unlearning adapter for frozen WSVAD detectors that combines category-conditioned temporal reading, retain-aware MIL optimization, and coordinated fine-grained and binary modulation. A single trained adapter supports different category-unlearning requests by changing only the retain state, without request-time optimization or retraining.
- We evaluate C-DBM with two frozen base WSVAD models on UCF-Crime and XD-Violence under both single- and multi-category unlearning requests. C-DBM substantially reduces forgotten-category responses while maintaining binary and fine-grained detection performance under the relabeled full-test protocol.

2 Related Work

Weakly supervised video anomaly detection. Early WSVAD methods primarily learn category-agnostic anomaly scores through ranking-based MIL, self-training, feature-magnitude learning, and

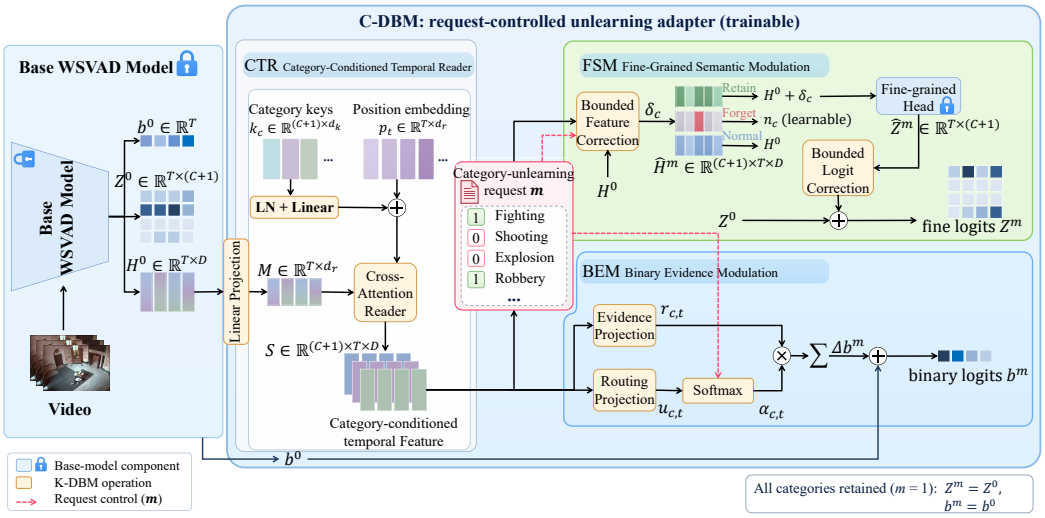

Figure 2: Overview of our proposed C-DBM. Given a frozen base WSVAD model, CTR reads category-conditioned temporal representations shared by both branches, while FSM modulates fine-grained semantic responses and BEM modulates the binary anomaly response according to the category-unlearning request.

magnitude-contrastive modeling (Sultani et al., 2018; Feng et al., 2021; Tian et al., 2021; Chen et al., 2023). Recent vision–language approaches incorporate category semantics through textual prompting and semantic priors (Yang et al., 2024; Chen et al., 2024), while VADCLIP (Wu et al., 2024) combines coarse-grained anomaly classification with fine-grained vision–language alignment. Ex-VAD, LEC-VAD, and DSANet further develop fine-grained, category-aware, or disentangled semantic modeling (Huang et al., 2025; Wang & Chen, 2025; Yin et al., 2026). These methods focus on improving detection, localization, or semantic recognition; our work instead studies post-training removal of a learned anomaly category while preserving retained-anomaly detection.

Machine unlearning and class unlearning. Machine unlearning removes designated training influence while preserving retained utility (Cao & Yang, 2015; Ginart et al., 2019; Bourtoule et al., 2021; Sekhari et al., 2021). Approximate methods reduce retraining cost through data partitioning, distillation, saliency- or importance-guided parameter updates, conflict-aware optimization, and subspace removal (Bourtoule et al., 2021; Kurmanji et al., 2023; Chundawat et al., 2023; Fan et al., 2024; Foster et al., 2024; Patel & Qiu, 2025; Qian et al., 2026). Class unlearning extends this goal to an entire target class, but existing formulations typically assume explicitly identifiable forget samples or instances. In WSVAD, target-category temporal evidence is latent under video-level supervision, retained and forgotten anomalies may coexist within a video, and category-specific semantics can share category-agnostic anomaly evidence.

Model designs for future unlearning. Some methods introduce class-dependent structures during training to facilitate later removal, including mnemonic codes in One-Shot Machine Unlearning, class-specific filters in WF-Net, and retain/forget-conditioned keys in NOVO (Yamashita et al., 2025; Poppi et al., 2024; Roy et al., 2025). These studies show that class-dependent control can facilitate future unlearning. Unlike them, C-DBM leaves a pretrained base WSVAD model frozen and learns an external adapter under weak video-level supervision to coordinate forgotten-category suppression across fine-grained semantic and shared binary anomaly outputs.

3 METHOD

As illustrated in Figure 2, C-DBM is a lightweight request-conditioned adapter for frozen dual-branch WSVAD models. Given a category-unlearning request represented by the retain vector $\mathbf{m}$, CTR reads category-conditioned temporal representations that are subsequently used by FSM and BEM to modulate the fine-grained semantic and shared binary responses, respectively. The base

model remains frozen during adapter training, and different category-unlearning requests are executed by updating only $\mathbf{m}$ without request-specific optimization.

3.1 PRELIMINARIES

Following the standard multiple-instance learning formulation of WSVAD (Sultani et al., 2018; Tian et al., 2021), let $\mathbb{V} = \{(\mathcal{V}_i, \mathbf{y}_i)\}_{i=1}^{N}$ denote N training videos with video-level multi-hot labels $\mathbf{y}_i \in \{0,1\}^{C+1}$ over Normal (indexed by 0) and C anomaly categories $\mathcal{C}_a = \{1, \ldots, C\}$. Here, $y_{i,0} = 1$ denotes a normal video and $y_{i,c} = 1$ indicates the presence of anomaly category $c \in \mathcal{C}_a$; Normal is mutually exclusive with anomaly labels, and no snippet-level category annotations are available during training.

Each video $\mathcal{V}_i$ is represented by T temporal snippets, yielding $\mathbf{X}_i \in \mathbb{R}^{T\times D}$ with feature dimension D. The frozen dual-branch base WSVAD model $\mathcal{F}_{\boldsymbol{\theta}}$ produces shared temporal features together with binary and fine-grained snippet-level logits:

$$(\mathbf{H}^0, \mathbf{b}^0, \mathbf{Z}^0) = \mathcal{F}_{\boldsymbol{\theta}}(\mathbf{X}), \tag{1}$$

where $\mathbf{H}^0 \in \mathbb{R}^{T\times D}$, $\mathbf{b}^0 \in \mathbb{R}^T$, and $\mathbf{Z}^0 \in \mathbb{R}^{T\times(C+1)}$ denote the shared temporal features, category-agnostic binary anomaly logits, and fine-grained logits over Normal and the C anomaly categories, respectively.

Following the class-wise retain/forget conditioning in (Roy et al., 2025), each category-unlearning request is encoded by a retain vector $\mathbf{m} \in \{0,1\}^C$, where $m_c = 1$ and $m_c = 0$ indicate retaining and forgetting category $c \in \mathcal{C}_a$, respectively. Normal is always retained and excluded from $\mathbf{m}$, and at least one anomaly category must remain retained. Conditioned on $\mathbf{m}$, C-DBM produces request-specific binary and fine-grained outputs $(\mathbf{b}^{\mathbf{m}}, \mathbf{Z}^{\mathbf{m}})$ from the frozen base-model representations. When $\mathbf{m} = \mathbf{1}$, no category unlearning is requested, so FSM and BEM are bypassed by design and the frozen outputs are returned exactly, i.e., $\mathbf{b}^{\mathbf{m}} = \mathbf{b}^0$ and $\mathbf{Z}^{\mathbf{m}} = \mathbf{Z}^0$.

3.2 CATEGORY-CONDITIONED TEMPORAL READER

Because WSVAD lacks snippet-level category supervision, selective unlearning first requires extracting category-conditioned temporal evidence from the shared base representation. We therefore introduce a shared *Category-Conditioned Temporal Reader* (CTR), which uses learnable category keys and cross-attention to read category-conditioned temporal states from $\mathbf{H}^0$ for both FSM and BEM. For efficiency, $\mathbf{H}^0$ is projected from dimension D to d_r to form the shared memory

$$\mathbf{M} = W_{\text{mem}}\,\mathrm{LN}(\mathbf{H}^0) \in \mathbb{R}^{T\times d_r}. \tag{2}$$

For each category $c \in \{0\}\cup\mathcal{C}_a$, a learnable key $\mathbf{k}_c \in \mathbb{R}^{d_k}$ is projected and combined with a temporal position embedding $\mathbf{p}_t$ to initialize a category–time query:

$$\mathbf{R}_{c,t}^{(0)} = W_k\,\mathrm{LN}(\mathbf{k}_c) + \mathbf{p}_t. \tag{3}$$

Here, $\mathbf{k}_c$ specifies the category-related evidence to retrieve, while $\mathbf{p}_t$ anchors the query at snippet t for subsequent local temporal reading from $\mathbf{M}$. At layer ℓ, CTR applies locally masked cross-attention using $\mathbf{R}_c^{(\ell-1)}$ as the query and the shared memory $\mathbf{M}$ as both key and value:

$$\mathbf{R}_c^{(\ell)} = \mathrm{CrossAttn}\Big(\mathbf{R}_c^{(\ell-1)}, \mathbf{M}, \mathbf{M}; \boldsymbol{\Omega}\Big), \tag{4}$$

where $\mathrm{CrossAttn}$ denotes a standard cross-attention block with residual and feed-forward updates (Vaswani et al., 2017), and $\boldsymbol{\Omega}$ restricts attention to snippets within the same local temporal window. After L layers, we obtain $\mathbf{S} = \mathrm{LN}(\mathbf{R}^{(L)}) \in \mathbb{R}^{(C+1)\times T\times d_r}$, where $\mathbf{S}_{c,t}$ denotes the category-conditioned temporal state for category c at snippet t and is shared by FSM and BEM.

3.3 FINE-GRAINED SEMANTIC MODULATION

Because the fine-grained branch explicitly encodes category semantics, forgetting a category requires modifying its video-conditioned semantic pathway rather than merely suppressing the corresponding output, while preserving retained categories and Normal. We therefore introduce *Fine-Grained Semantic Modulation* (FSM), which uses the CTR states to modulate category-wise semantic streams according to the retain vector $\mathbf{m}$. For each anomaly category $c \in \mathcal{C}_a$, $\mathbf{S}_{c,t}$ is projected to

the base-model feature dimension D and norm-bounded relative to $\mathbf{H}_t^0$:

$$\widetilde{\boldsymbol{\delta}}_{c,t} = W_\delta \mathbf{S}_{c,t}, \qquad \boldsymbol{\delta}_{c,t} = \widetilde{\boldsymbol{\delta}}_{c,t} \frac{\rho \|\mathbf{H}_t^0\|_2}{\rho \|\mathbf{H}_t^0\|_2 + \|\widetilde{\boldsymbol{\delta}}_{c,t}\|_2}. \tag{5}$$

This scaling preserves the correction direction while enforcing $\|\boldsymbol{\delta}_{c,t}\|_2 < \rho \|\mathbf{H}_t^0\|_2$, where $\rho > 0$ specifies the relative correction budget. FSM then constructs the category-wise semantic streams according to the retain state:

$$\widetilde{\mathbf{H}}_{c,t}^{\mathbf{m}} = \begin{cases} \mathbf{H}_t^0, & c = 0, \\ \mathbf{H}_t^0 + \boldsymbol{\delta}_{c,t}, & c > 0,\ m_c = 1, \\ \mathbf{n}_c, & c > 0,\ m_c = 0, \end{cases} \tag{6}$$

where $\mathbf{n}_c \in \mathbb{R}^D$ is a learned video-independent null vector. Retained anomaly categories receive bounded feature corrections, forgotten categories are replaced by null features, and the Normal stream is taken directly from $\mathbf{H}_t^0$.

The modulated category-wise features are passed through the frozen fine-grained readout of the base WSVAD model, allowing FSM to reuse its pretrained semantic prediction space without introducing an additional classifier. Applying this readout independently to each category stream yields the intermediate fine-grained logits $\widehat{\mathbf{Z}}^{\mathbf{m}} \in \mathbb{R}^{T \times (C+1)}$.

Rather than replacing the frozen predictions with $\widehat{\mathbf{Z}}^{\mathbf{m}}$, FSM converts their deviation from $\mathbf{Z}^0$ into a bounded residual correction:

$$\Delta \mathbf{Z}_{\text{fine}}^{\mathbf{m}} = B_{\text{fine}} \tanh\left(\frac{\widehat{\mathbf{Z}}^{\mathbf{m}} - \mathbf{Z}^0}{B_{\text{fine}}} \right), \qquad \mathbf{Z}^{\mathbf{m}} = \mathbf{Z}^0 + \Delta \mathbf{Z}_{\text{fine}}^{\mathbf{m}}. \tag{7}$$

Here, $B_{\text{fine}} > 0$ ensures $|\Delta Z_{\text{fine},t,c}^{\mathbf{m}}| < B_{\text{fine}}$, keeping $\mathbf{Z}^0$ as the prediction reference.

3.4 Binary Evidence Modulation

Because the binary branch is category-agnostic, fine-grained modulation alone cannot remove forgotten-category evidence from the shared anomaly response while preserving co-occurring retained anomalies. We therefore introduce *Binary Evidence Modulation* (BEM), which maps each anomaly-category CTR state $\mathbf{S}_{c,t}$, $c \in \mathcal{C}_a$, to a bounded signed evidence value $r_{c,t}$ and a routing logit $u_{c,t}$. Specifically, $e_{c,t} = \mathbf{w}_{\text{evi}}^\top \mathbf{S}_{c,t} + b_{\text{evi}}$ is bounded as $r_{c,t} = B_{\text{bin}} \tanh(e_{c,t}/B_{\text{bin}})$, while $u_{c,t} = \mathbf{w}_{\text{route}}^\top \mathbf{S}_{c,t} + \beta_c$ controls the relative participation of category c in the binary correction. The routing logits are normalized only over retained anomaly categories:

$$\alpha_{c,t} = \frac{m_c \exp(u_{c,t})}{\sum_{j=1}^{C} m_j \exp(u_{j,t})}, \qquad c \in \mathcal{C}_a. \tag{8}$$

Here, $\mathbf{w}_{\text{evi}}, \mathbf{w}_{\text{route}} \in \mathbb{R}^{d_r}$ are the evidence and routing projections, $b_{\text{evi}} \in \mathbb{R}$ is the evidence bias, $\beta_c \in \mathbb{R}$ is a category-specific routing bias, and $B_{\text{bin}} > 0$ bounds the signed evidence magnitude. Because $m_c = 0$ forces $\alpha_{c,t} = 0$, forgotten categories are excluded from routing, while the weights over retained categories sum to one at each snippet. BEM combines the retained-category participation $\alpha_{c,t}$ with the signed evidence $r_{c,t}$ to form the binary correction:

$$\Delta b_t^{\mathbf{m}} = \sum_{c=1}^{C} \alpha_{c,t} r_{c,t}. \tag{9}$$

BEM applies the routed correction to the frozen binary logit as $b_t^{\mathbf{m}} = b_t^0 + \Delta b_t^{\mathbf{m}}$. Because $r_{c,t}$ is signed, the correction can either increase or decrease the anomaly response, allowing BEM to preserve retained anomaly evidence in mixed-label videos rather than enforcing monotonic suppression.

3.5 Weakly Supervised Optimization for Unlearning

During training, we sample retain vectors $\mathbf{m}$ to simulate category-unlearning requests and derive request-conditioned supervision from the original video-level labels. Videos whose annotated

anomalies are all forgotten are excluded from the retention losses rather than relabeled as Normal, whereas mixed-label videos remain eligible through their retained annotated categories.

To preserve the base model's weakly supervised training objective, we retain its native losses $\mathcal{L}_{\text{bin}}$ and $\mathcal{L}_{\text{fine}}$, implemented as top-$k$ MIL binary cross-entropy and fine-grained category-wise cross-entropy, respectively. Under each retain state, $\mathcal{L}_{\text{bin}}$ assigns normal videos a negative target and anomalous videos with at least one retained category a positive target, while $\mathcal{L}_{\text{fine}}$ keeps the Normal target for normal videos and distributes the target mass uniformly over the retained annotated categories of anomalous videos.

Beyond the retention losses, C-DBM introduces two request-specific objectives for forgotten-category suppression and retained-category routing. Let $P_{i,c}^{\text{fine}}$ denote the video-level fine-grained probability of category c, and let $\overline{\alpha}_{i,c}$ denote $\alpha_{i,c,t}$ averaged over the top-k snippets selected by the frozen binary logits $\mathbf{b}_i^0$. We apply $\mathcal{L}_{\text{forget}}$ to each forgotten positive pair satisfying $m_c = 0$ and $y_{i,c} = 1$, and $\mathcal{L}_{\text{route}}$ to each video containing at least one retained annotated category:

$$\mathcal{L}_{\text{forget}}(i,c) = -\log(1 - P_{i,c}^{\text{fine}}), \qquad \mathcal{L}_{\text{route}}(i) = -\log\left(\sum_{c=1}^{C} y_{i,c} m_c \overline{\alpha}_{i,c}\right). \tag{10}$$

$\mathcal{L}_{\text{forget}}$ suppresses forgotten-category predictions, whereas $\mathcal{L}_{\text{route}}$ concentrates BEM routing mass on retained annotated categories. Because $\mathcal{L}_{\text{route}}$ supervises the retained category set rather than a single class, multiple co-occurring retained categories may jointly receive routing mass.

The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{bin}} + \mathcal{L}_{\text{fine}} + \mathcal{L}_{\text{forget}} + \lambda_r \mathcal{L}_{\text{route}}. \tag{11}$$

Here, $\mathcal{L}_{\text{bin}}$ and $\mathcal{L}_{\text{fine}}$ preserve retained detection, $\mathcal{L}_{\text{forget}}$ suppresses forgotten-category responses, and $\mathcal{L}_{\text{route}}$ guides retained-category routing, with λ_r controlling its contribution.

4 Experiments

We evaluate C-DBM from three perspectives: (1) whether a single trained adapter can execute different category-unlearning requests on demand; (2) whether forgotten-category responses can be suppressed while preserving binary and fine-grained detection under the request-dependent relabeled full-test protocol; and (3) how each category-conditioned component contributes to selective anomaly-category unlearning.

4.1 Experimental Setup

Datasets and base models. We evaluate on UCF-Crime (Sultani et al., 2018) and XD-Violence (Wu et al., 2020) using their official train–test splits. UCF-Crime contains 13 anomaly categories, while XD-Violence contains six anomaly categories with multiple categories potentially co-occurring within a video. We use VADCLIP (Wu et al., 2024) and DSANet (Yin et al., 2026) as dual-branch base WSVAD models. Each base model is first trained with all available anomaly categories and then frozen, while only C-DBM is optimized during unlearning-adapter training.

Evaluation. For each category-unlearning request, represented by a forgotten-category set F and its corresponding retain vector $\mathbf{m}$, we construct a request-dependent frame-level ground truth on the complete test set.

Binary and Fine are evaluated against this relabeled full-test ground truth using frame-level AUC on UCF-Crime and AP on XD-Violence, where Binary uses the category-agnostic anomaly score and Fine uses $1 - p_t(\text{Normal})$. Forget Top-1 and Fine P. are computed on anomalous frames from videos associated with the requested categories. On XD-Violence, this video-level selection can include frames from mixed-label videos whose temporal annotations contain only retained categories, so these metrics are reported as residual-response diagnostics separately from the Relabeled Full detection metrics. Retain AUC/AP are computed on videos containing none of the requested categories and are reported only as supplementary diagnostics. All metrics are computed separately for each request and macro-averaged over requests within the same evaluation protocol.

Implementation details. Both base models output $D = 512$-dimensional temporal features, and CTR uses two cross-attention layers with $d_r = 128$, $d_k = 64$, and local windows of 8 and 64

Table 1: Multi-category anomaly-category unlearning on DSANet–XD-Violence under the relabeled full-test protocol. All values are percentages and are macro-averaged over requests with the same number of forgotten categories. † Hard Closure masks the forgotten fine-grained logits at inference.

# Forgotten categories	# Requests	Base model		Hard Closure†		C-DBM		C-DBM Fine P. /category (%)↓
		Bin AP↑	Fine AP↑	Bin AP↑	Fine AP↑	Bin AP↑	Fine AP↑	
1	6	74.03	73.06	74.03	77.89	**74.58**	**78.76**	0.67
2	15	61.51	60.74	61.51	70.29	**62.81**	**72.16**	0.38
3	20	48.28	47.73	48.28	61.89	**50.04**	**64.66**	0.37

snippets on UCF-Crime and XD-Violence, respectively. We set $\rho = 0.1$, $B_{\text{bin}} = 2$, $B_{\text{fine}} = 8$, and $\lambda_r = 0.1$, and train C-DBM for 20 epochs in FP32 using AdamW with two sampled requests per mini-batch. During training, sampled requests contain one to three forgotten categories on UCF-Crime and one forgotten category on XD-Violence. All reported experiments use random seed 234.

4.2 On-Demand Category Unlearning

We first evaluate the on-demand setting in which one trained adapter executes different anomaly-category unlearning requests by changing only the retain vector $\mathbf{m}$. To focus the analysis on a dataset with category-time annotations, Table 1 reports complete one-, two-, and three-category request results on DSANet–XD-Violence. Although UCF-Crime provides anomaly categories at the video level, it does not provide equivalent category-time co-occurrence annotations; therefore, we do not use it for analyzing retained and withdrawn anomaly interactions.

Table 2: Single-category anomaly-category unlearning results under the relabeled full-test protocol. All entries are percentages and are macro-averaged over 13 UCF-Crime requests or 6 XD-Violence requests. Full Bin/Fine are computed on the complete test set with request-dependent frame relabeling. Top-1 and Fine P. are computed on the original forgotten-category evaluation range described in the text. Bold values indicate the best retained detection or lowest forgotten-category response among the unlearning methods for each base-model–dataset column group.

Base model	Method	UCF-Crime: Relabeled Full				XD-Violence: Relabeled Full			
		Bin AUC ↑	Fine AUC ↑	Top-1 ↓	Fine P. ↓	Bin AP ↑	Fine AP ↑	Top-1 ↓	Fine P. ↓
VADCLIP	The base model	87.56	86.14	28.96	22.72	66.93	72.75	63.97	58.15
	FT (2023)	86.02	85.05	**0.00**	0.47	76.34	73.15	8.17	7.92
	NegGrad+ (2023)	84.08	83.47	**0.00**	0.17	73.32	77.92	**0.00**	**0.00**
	SSD (2024)	84.51	69.63	2.47	6.30	71.53	63.40	21.80	24.94
	SCRUB (2023)	87.41	86.00	17.29	13.81	75.54	80.25	40.50	36.62
	SalUn (2024)	86.47	85.15	**0.00**	3.35	74.10	75.66	3.24	11.81
	Label Flipping (2021)	85.00	83.63	**0.00**	0.18	76.76	75.34	**0.00**	0.04
	Bad Teacher (2023)	86.22	85.42	4.12	6.02	**77.33**	80.24	0.01	3.07
	LUR (2025)	84.96	83.28	**0.00**	**0.06**	76.81	77.16	**0.00**	**0.00**
	MUpHT (2026)	86.43	85.22	0.03	1.78	54.20	61.02	7.27	9.49
	C-DBM (Ours)	**87.53**	**86.17**	**0.00**	0.38	67.30	**80.68**	**0.00**	1.48
DSANet	The base model	88.49	87.92	33.03	25.40	74.03	73.06	60.82	48.16
	FT (2023)	87.08	86.19	**0.00**	0.18	75.86	73.20	**0.00**	0.55
	NegGrad+ (2023)	83.86	85.63	0.21	0.68	69.70	69.91	**0.00**	0.01
	SSD (2024)	86.68	85.66	0.02	2.87	74.70	71.51	12.55	14.20
	SCRUB (2023)	88.30	87.50	12.40	12.53	**80.97**	**80.30**	3.63	9.35
	SalUn (2024)	86.87	86.20	0.09	2.38	77.82	75.59	1.51	11.08
	Label Flipping (2021)	86.68	85.26	**0.00**	**0.09**	77.75	75.99	**0.00**	0.10
	Bad Teacher (2023)	87.44	87.16	7.28	7.33	79.56	**80.24**	0.09	3.87
	LUR (2025)	85.33	85.13	2.87	2.34	78.12	70.32	**0.00**	**0.00**
	MUpHT (2026)	88.25	87.75	0.02	1.32	65.24	65.38	5.70	9.70
	C-DBM (Ours)	**88.62**	**87.97**	**0.00**	0.13	74.58	78.76	**0.00**	0.67

Varying request sizes. On DSANet–XD-Violence, C-DBM achieves higher Binary/Fine AP than Hard Closure by 0.55/0.87 percentage points for one-category requests, by 1.30/1.86 percentage points for two-category requests, and by 1.76/2.77 percentage points for three-category requests under the relabeled full-test protocol. The residual probability assigned to forgotten categories remains

low, with Forgotten Fine P./category values of 0.67%, 0.38%, and 0.37%, respectively. These results are macro-averaged over requests within this base-model–dataset setting and are not intended as a universal comparison across all datasets and baselines.

Deployment efficiency. C-DBM adds only 0.579M and 0.575M parameters on UCF-Crime and XD-Violence, respectively, corresponding to 0.299–0.361% of the frozen base model, while a single adapter is shared across all category requests for each base-model–dataset pair. At deployment, each category-unlearning request is executed by changing only the retain vector $\mathbf{m}$, without gradient updates, access to training data, or request-specific optimization.

4.3 Comparison with Unlearning Baselines

We compare C-DBM with adapted request-specific unlearning methods under the single-category relabeled full-test protocol.

Forgetting effectiveness. On anomalous frames from videos containing the forgotten categories, the frozen base models assign the corresponding categories a Top-1 rate of 28.96–63.97% and a Fine P. of 22.72–58.15%. C-DBM reduces Top-1 to 0 across all four settings and lowers Fine P. to 0.13–1.48% without hard output masking.

Figure 3 shows that forgetting extends beyond Top-1: Hit@1/2/3 decreases from 60.82/76.23/85.35% for the frozen DSANet to 0.00/1.95/3.75% with C-DBM.

Relabeled full-test detection. On UCF-Crime, C-DBM maintains competitive Full Bin/Fine AUC among the trainable unlearning methods for both VADCLIP and DSANet. On VADCLIP–XD-Violence, C-DBM achieves the highest Full Fine AP among the trainable methods while maintaining comparable Full Binary AP. On DSANet–XD-Violence, C-DBM preserves the relabeled full-test detection performance while substantially reducing residual forgotten-category responses. These results demonstrate that C-DBM can suppress forgotten-category responses without significantly compromising retained detection under the relabeled evaluation protocol.

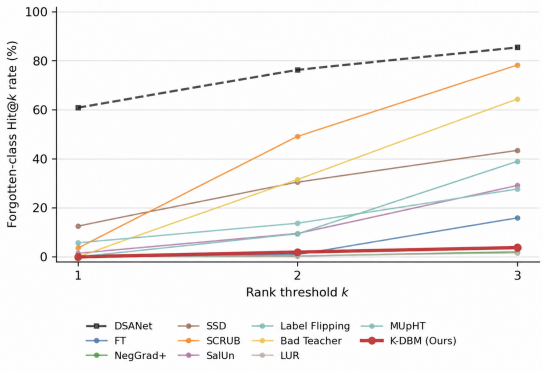

Figure 3: Forgotten-category Hit@k on DSANet–XD-Violence, macro-averaged over six single-category requests. Hit@k is the fraction of anomalous frames from videos associated with the forgotten category where the forgotten category remains among the top-k predictions; lower is better.

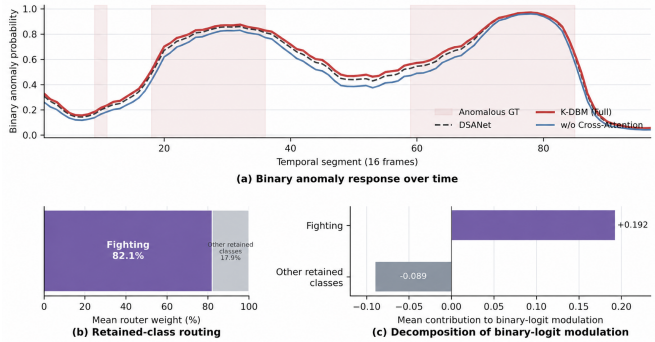

Figure 4: Temporal analysis of a mixed-label XD-Violence video with Shooting forgotten and Fighting retained. (a) Binary anomaly responses of DSANet, C-DBM, and the variant without cross-attention. Shaded regions indicate ground-truth anomaly intervals. (b) Mean routing weights over retained categories. (c) Mean signed contributions to binary-logit modulation.

4.4 Ablation and Analysis

We analyze C-DBM on DSANet–XD-Violence, where the trade-off between relabeled full-test detection and forgotten-category suppression is most visible in Table 2. All ablation results are macro-averaged over six single-category requests without Hard Closure, and Exact Retrain is used only as a reference for distributional diagnostics.

We evaluate three variants: removing CTR cross-attention by setting $\mathbf{R}_c^{(L)} = \mathbf{R}_c^{(0)}$, disabling FSM by setting $\mathbf{Z}^{\mathbf{m}} = \mathbf{Z}^0$, and disabling BEM by setting $\mathbf{b}^{\mathbf{m}} = \mathbf{b}^0$.

Table 3: Module-wise ablation of C-DBM on DSANet–XD-Violence, macro-averaged over six single-category forgetting requests.

Removed component	Diagnostic	Full	w/o component
CTR cross-attention	Binary selectivity S_B (pp)↑	**2.57**	0.54
	Forgotten Binary-JS↓	**0.0841**	0.0892
FSM	Hit@3 (%)↓	**3.75**	85.35
	Forgotten Fine-JS↓	**0.1600**	0.2778
BEM	Forgotten Binary-JS↓	**0.0841**	0.0966
	Binary MIA gap (pp)↓	**8.13**	9.25

Diagnostic metrics. Diagnostic metrics. Hit@3 measures whether the forgotten category remains among the top-three predictions. Forgotten Binary-JS and Forgotten Fine-JS measure the per-frame Jensen–Shannon divergence from Exact Retrain on frames associated with the forgotten categories, while Binary MIA gap measures the difference in binary loss-threshold attack AUC. Binary selectivity is defined as $S_B = \Delta_{\text{excl}} - \Delta_{\text{mix}}$, where $\Delta = P_{\text{active}} - P_{\text{withdraw}}$ is computed on category-exclusive and mixed-label videos, respectively; larger S_B indicates more selective binary modulation.

Effect of CTR cross-attention. Removing cross-attention reduces S_B from 2.57 to 0.54, while Forgotten Binary-JS changes from 0.0841 to 0.0892. This indicates that category-conditioned temporal reading mainly improves where binary evidence is modified, concentrating the response change on category-exclusive videos while limiting interference with retained anomalies.

Contributions of FSM and BEM. Removing FSM increases Hit@3 from 3.75% to 85.35% and Forgotten Fine-JS from 0.1600 to 0.2778, confirming that FSM provides the primary forgotten-category semantic suppression. Removing BEM increases Forgotten Binary-JS from 0.0841 to 0.0966 and Binary MIA gap from 8.13 to 9.25 points, supporting its complementary role in modifying the shared binary response toward Exact Retrain.

Mixed-label routing behavior. Figure 4 analyzes a mixed-label video where Shooting is forgotten and Fighting is retained. Over anomalous segments, C-DBM produces a mean binary response of 77.30%, compared with 75.66% for the frozen base model and 72.10% without cross-attention. The routing weights assign 82.1% of the mean mass to Fighting, whose $+0.192$ contribution offsets the -0.089 contribution from the other retained categories. This illustrates how category-conditioned routing can preserve retained anomaly evidence when forgotten and retained anomalies coexist.

5 Conclusion

This work introduces C-DBM, a lightweight adapter for on-demand anomaly-category unlearning in dual-branch WSVAD. Category-conditioned cross-attention jointly supports fine-grained semantic modulation and binary evidence modulation while keeping the base WSVAD model frozen. Across four base-model–dataset combinations, C-DBM supports both single- and multi-category unlearning requests and substantially reduces forgotten-category responses while maintaining binary and fine-grained detection performance under the relabeled full-test protocol. Ablations show improved binary selectivity; a mixed-label case further illustrates how routing adjusts the binary response using retained-category evidence. As C-DBM targets conditional category-level behavior, certified parameter erasure and stronger unlearning guarantees remain future directions.

LLM USE DISCLOSURE

We used large language models to (i) improve the grammar and wording of this manuscript and (ii) assist with experimentation by drafting portions of the code and suggesting debugging steps. All scientific ideas, claims, experiment designs, and final implementations are by the authors. We reviewed and verified all LLM-assisted text and code before use, and the authors remain fully responsible for the content of this paper.

ETHICS STATEMENT

This work complies with the ICLR Code of Ethics (`https://iclr.cc/public/CodeOfEthics`). Our study does not involve human subjects, sensitive personal data, or any form of biometric information. All datasets used are publicly available and widely adopted in the research community. We have taken care to avoid generating or reinforcing harmful content, stereotypes, or biases in both model design and evaluation. No proprietary or confidential data was used. There are no known conflicts of interest or external sponsorship that could influence the outcomes of this research. We acknowledge the importance of ethical considerations in AI research and have made efforts to ensure transparency, reproducibility, and fairness throughout the development of our framework.

REFERENCES

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy*, pp. 141–159, 2021. doi: 10.1109/SP40001.2021.00019.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pp. 463–480, 2015. doi: 10.1109/SP.2015.35.

Junxi Chen, Liang Li, Li Su, Zheng-Jun Zha, and Qingming Huang. Prompt-enhanced multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18319–18329, 2024.

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. doi: 10.1609/aaai.v37i1.25112.

Vikram S. Chundawat, Ayush K. Tarun, Murari Mandal, and Mohan Kankanhalli. Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7210–7217, 2023. doi: 10.1609/aaai.v37i6.25879.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. SalUn: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=gn0mIhQGNM`.

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.

Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast machine unlearning without retraining through selective synaptic dampening. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 12043–12051, 2024. doi: 10.1609/aaai.v38i11.29092.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Zou. Making AI forget you: Data deletion in machine learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11516–11524, 2021. doi: 10.1609/aaai.v35i13.17371.

Chao Huang, Yushu Shi, Jie Wen, Wei Wang, Yong Xu, and Xiaochun Cao. Ex-VAD: Explainable fine-grained video anomaly detection based on visual-language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25750–25761. PMLR, 2025.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Gaurav Patel and Qiang Qiu. Learning to unlearn while retaining: Combating gradient conflicts in machine unlearning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4211–4221, 2025. doi: 10.1109/ICCV51701.2025.00401.

Samuele Poppi, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Multi-class unlearning for image classification via weight filtering. *IEEE Intelligent Systems*, 39(6):40–47, 2024. doi: 10.1109/MIS.2024.3412742.

Chengyao Qian, Jing Wu, Trung Le, Dinh Phung, and Mehrtash Harandi. ph-strips for selective forgetting: A blunt but fast diagnostic baseline for machine unlearning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3306–3315, 2026.

Soumya Roy, Soumya Banerjee, Vinay Verma, Soumik Dasgupta, Deepak Gupta, and Piyush Rai. NOVO: Unlearning-compliant vision transformers. *arXiv preprint arXiv:2507.03281*, 2025. doi: 10.48550/arXiv.2507.03281.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

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.

Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W. Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4975–4986, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yu Wang and Shiwei Chen. Learning event completeness for weakly supervised video anomaly detection. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 62505–62517. PMLR, 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 *European Conference on Computer Vision*, pp. 322–339. Springer, 2020. doi: 10.1007/978-3-030-58577-8_20.

Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. 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. doi: 10.1609/aaai.v38i6.28423.

Tomoya Yamashita, Masanori Yamada, and Takashi Shibata. One-shot machine unlearning with mnemonic code. In *Proceedings of the 16th Asian Conference on Machine Learning*, volume 260 of *Proceedings of Machine Learning Research*, pp. 255–270. PMLR, 2025.

Zhiwei Yang, Jing Liu, and Peng Wu. Text prompt with normality guidance for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18899–18908, 2024.

Wenti Yin, Huaxin Zhang, Xiang Wang, Yuqing Lu, Yicheng Zhang, Bingquan Gong, Jialong Zuo, Li Yu, Changxin Gao, and Nong Sang. Learning to tell apart: Weakly supervised video anomaly detection via disentangled semantic alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12027–12035, 2026. doi: 10.1609/aaai.v40i14.38191.

Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. Harnessing large language models for training-free video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18527–18536, 2024.

A Exact Retraining Reference

For each single-category request, Exact Retrain removes every training video annotated with the requested category and trains the corresponding WSVAD host from random initialization using its native training procedure. All crops of a removed video are excluded; on XD-Violence, this also removes videos carrying both the requested category and retained categories. The output vocabulary and official test set remain unchanged. We run one seed (234) for 10 epochs per request and use the native best-checkpoint rule: binary frame-level AUC on UCF-Crime for both hosts, fine-grained AP on VADCLIP–XD-Violence, and binary AP on DSANet–XD-Violence. The native host procedure uses the official test metric for checkpoint selection.

Tables 4 and 5 report each retrained model's *full-test* detection performance, followed by the equal-weight mean over 13 or six deletion requests. These values characterize the delete-and-retrain reference; they are not retained-subset results and must not be compared directly with the retained-video AUC/AP in Table 2. In the ablation of Table 3, the corresponding per-request retrained checkpoints instead serve as paired references for forgotten-scope prediction JS and membership-inference AUC gaps. Those diagnostics are computed from matched predictions and are not derived from the full-test values below.

Table 4: Exact Retrain after deleting each UCF-Crime category: full-test frame-level AUC (%). Binary and Fine are the retrained host's binary and fine-grained branches.

Deleted category	VADCLIP		DSANet	
	Bin	Fine	Bin	Fine
Abuse	87.97	86.79	88.35	88.35
Arrest	87.70	85.91	88.35	88.35
Arson	87.38	86.16	87.71	87.71
Assault	87.56	85.83	88.07	88.07
Burglary	87.35	85.67	89.07	89.07
Explosion	86.70	85.23	87.73	87.73
Fighting	87.89	86.18	87.92	87.92
RoadAccidents	86.74	85.27	87.92	87.92
Robbery	88.16	86.10	87.61	87.61
Shooting	87.98	86.54	87.98	87.98
Shoplifting	88.10	86.18	87.83	87.83
Stealing	87.85	85.87	87.61	87.61
Vandalism	87.45	85.57	88.97	88.97
Macro avg.	87.60	85.95	88.09	88.09

Table 5: Exact Retrain after deleting each XD-Violence category: full-test frame-level AP (%). Binary and Fine are the retrained host's binary and fine-grained branches.

Deleted category	VADCLIP		DSANet	
	Bin	Fine	Bin	Fine
Abuse	80.01	84.89	85.38	85.38
CarAccident	83.35	84.52	85.50	85.50
Explosion	80.15	83.15	85.33	85.33
Fighting	77.26	77.92	80.21	80.21
Riot	56.75	55.98	57.37	57.37
Shooting	81.12	86.20	86.58	86.58
Macro avg.	76.44	78.78	80.06	80.06

B Per-Category Unlearning Results

Tables 6 and 7 give the per-request retained-video Bin/Fine performance of C-DBM; their macro-average rows reproduce Table 2. The forgotten-category responses of all methods are compared below.

Table 6: Per-category retained-video detection by C-DBM on UCF-Crime (frame-level AUC, %).

Forgotten category	VADCLIP		DSANet	
	Bin	Fine	Bin	Fine
Abuse	87.84	86.55	88.91	88.19
Arrest	88.68	87.17	89.32	88.84
Arson	87.78	86.39	88.92	88.20
Assault	87.82	86.28	88.89	88.05
Burglary	88.70	87.11	89.67	88.82
Explosion	89.37	87.97	90.49	89.66
Fighting	87.87	86.28	88.95	88.04
RoadAccidents	88.33	86.77	89.51	88.87
Robbery	87.85	86.97	88.69	87.96
Shooting	89.22	87.95	90.09	89.39
Shoplifting	88.96	87.66	90.29	89.70
Stealing	87.43	86.03	88.50	87.91
Vandalism	87.87	86.54	88.90	88.28
Macro avg.	88.29	86.90	89.32	88.61

Table 7: Per-category retained-video detection by C-DBM on XD-Violence (frame-level AP, %).

Forgotten category	VADCLIP		DSANet	
	Bin	Fine	Bin	Fine
Abuse	79.81	84.90	85.74	84.86
CarAccident	85.87	88.45	89.31	88.26
Explosion	82.45	87.34	87.88	87.49
Fighting	79.92	86.86	87.54	87.27
Riot	61.89	61.39	63.37	58.70
Shooting	82.50	87.29	88.02	87.02
Macro avg.	78.74	82.70	83.64	82.27

Figures 5–8 place the nine baselines and C-DBM on the same per-request heatmaps. The rows are forgotten categories and the columns are methods; each cell gives the forgotten-category response in percent. The Top-1 panels show whether the forgotten category remains the highest-ranked prediction, whereas Fine P. shows its mean fine-grained probability on anomalous frames of its labeled

videos. The same response bands are used in every panel, with reported 0.00% values shown separately. Flip and BT abbreviate Label Flipping and Bad Teacher, respectively. These figures measure forgetting only; retained detection is reported in Tables 6 and 7 and compared across methods in Table 2. No inference-time closure is applied.

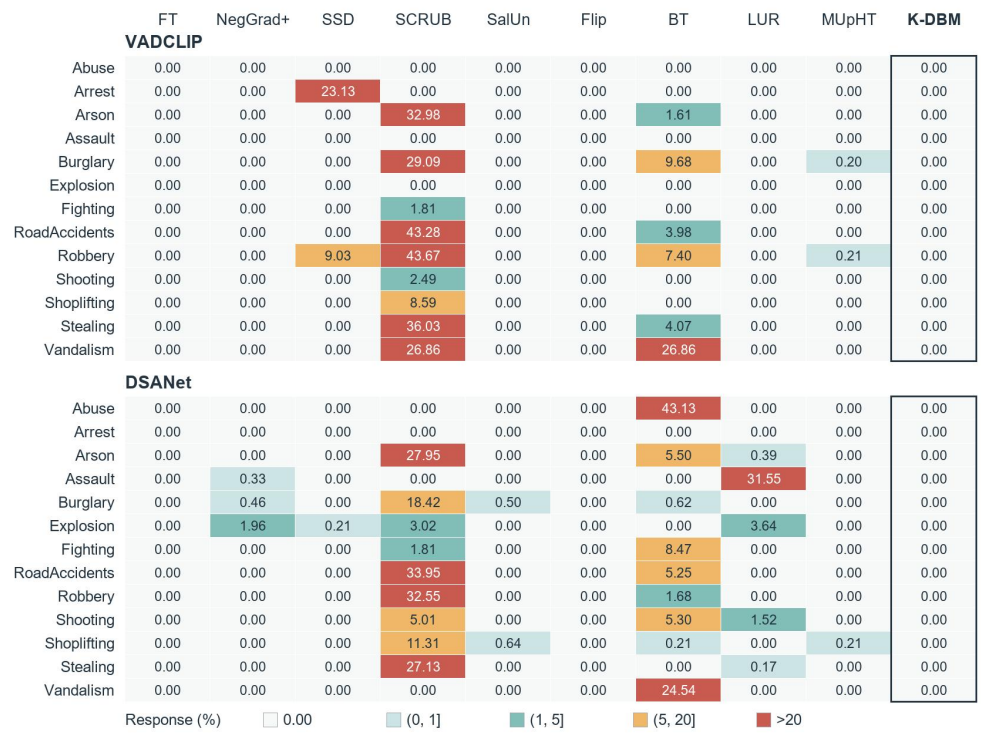

Figure 5: Per-category forgotten-class Top-1 rate on UCF-Crime (%; lower is better). C-DBM is outlined.

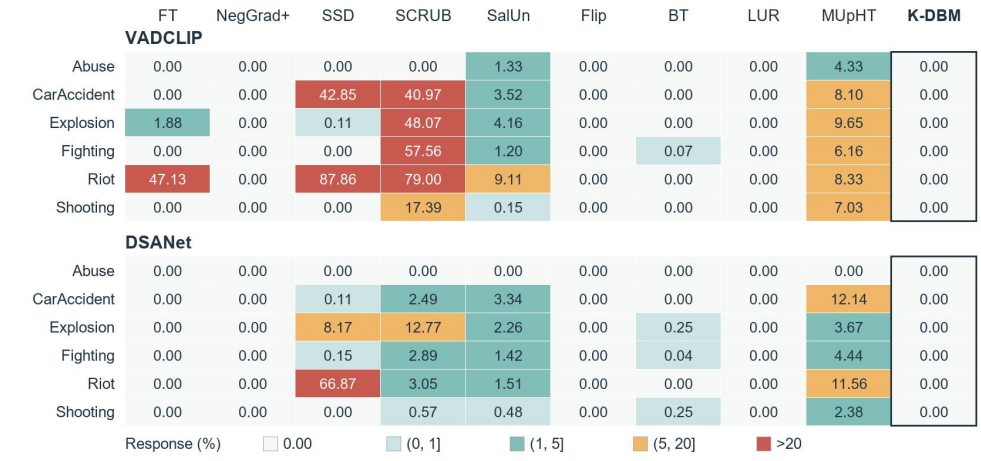

Figure 6: Per-category forgotten-class Top-1 rate on XD-Violence (%; lower is better). C-DBM is outlined.

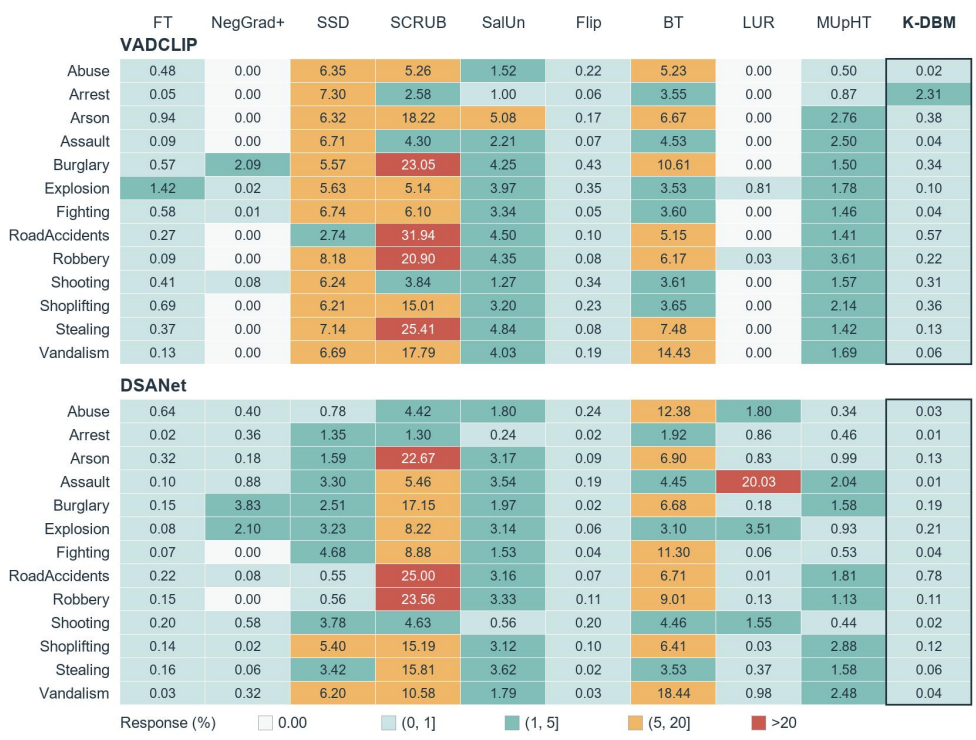

Figure 7: Per-category forgotten-class fine-grained probability on UCF-Crime (%; lower is better). C-DBM is outlined.

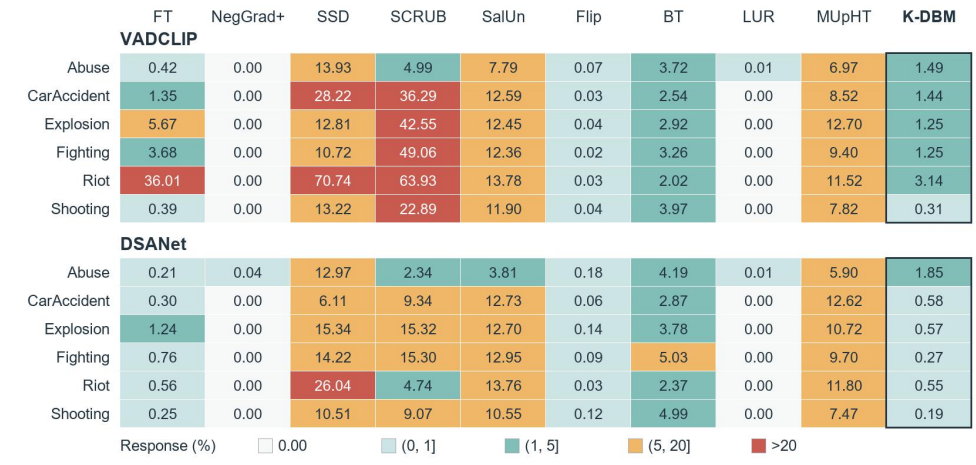

Figure 8: Per-category forgotten-class fine-grained probability on XD-Violence (%; lower is better). C-DBM is outlined.

C MIL ADAPTATION OF UNLEARNING BASELINES

Common WSVAD protocol. All nine baselines start independently from the corresponding fully trained base model for each requested category. The pretrained CLIP encoder remains frozen, while the host's downstream trainable parameters are eligible for updates or editing. A video's forget set contains every training video whose label set includes the requested category, including mixed-label XD-Violence videos; all crops of such a video share the same assignment. The retained set contains the remaining training videos. For a video with T_i valid snippets, we use $k_i = \lfloor T_i/16 \rfloor + 1$ and form the binary bag probability by averaging the top-k_i snippet-level sigmoid scores. The fine-grained bag logit for each category is the mean of that category's top-k_i snippet logits. Normal videos use the Normal target; XD-Violence videos with multiple positive labels use a normalized multi-hot target. Padded snippets are excluded from top-k_i selection. Where applicable, DSANet's native auxiliary objectives are used for retention, not reversed on forget videos. All reported baseline predictions retain every output channel; no hard closure is used.

Optimization and relabeling methods. FT optimizes the retained-video binary and fine-grained MIL objectives, without accessing forget videos during adaptation. NegGrad+ combines retained-data gradient descent with ascent on the original-label binary and fine-grained MIL losses of forget videos. Label Flipping assigns Normal to each forget video before applying the same MIL training objective; on XD-Violence this deliberately removes its co-occurring retained labels as well. LUR uses a differentiable inner update on the retained MIL objective, followed by ascent on the forgotten category's fine-grained bag loss evaluated at the look-ahead parameters. Its forget objective uses only the requested category, rather than reversing the binary loss or the other positive labels in a mixed-label video.

Distillation methods. SCRUB maximizes disagreement with the frozen base model on forget videos using bag-level binary and fine-grained KL terms, while minimizing supervised retained MIL losses and distillation on retained videos. In our adaptation, the forget maximization runs for the first five epochs and retained minimization for all 20 epochs. Bad Teacher distills the base model on retained videos at the bag level and matches a synthetic Normal-like teacher on forget videos: the binary bag target is zero and the fine-grained target places most mass on Normal. This is the Normal-like adaptation used in our comparison, not a claim that a separately initialized bad-teacher network was trained.

Importance and activation editing. SSD estimates empirical squared-gradient parameter importance from the full training set and forget videos using the MIL classification loss, then dampens coordinates with disproportionately high forget importance in one edit. SalUn selects the largest 50% of forget-loss gradient coordinates and updates only those coordinates using retained MIL supervision and random-label targets on forget videos. The random target is a single class even for a multi-label XD-Violence video. MUpHT collects activations at the host's downstream linear layer from the forgotten category's top-k_i snippets, forms a video-balanced activation subspace, and removes its top ten directions from that layer's weights in one closed-form edit. It does not take optimizer steps.

The iterative baselines use 20 adaptation epochs; SSD and MUpHT are one-shot edits. Table 2 uses each baseline's archived fixed endpoint and C-DBM's archived best checkpoint, with no inference-time output masking.