Appendix

The following items are included in our supplementary material:

- Appendix A gives the overall training procedure of our FedVAD in Algorithm.
- Appendix B gives a detailed description of the local client partitioning strategy.
- Appendix C gives a detailed description of the weighting effects of distillation loss.
- Appendix D gives a comparison of the performance of each method on each client.
- Appendix E gives the experimental results for different numbers of clusters.
- Appendix F gives the experimental results for the hyperparameters α and β that regulate the relative importance of model weights and object distributions when clustering.

A Algorithm 1

The comprehensive training procedure of our method is summarized in Algorithm 1.

B Experimental Settings

In our supplementary material, we detail the client partitioning, outlining the distribution of training and test sets for each client. Regarding the server-side public dataset construction, we allocate 10% to 20% of each client's training set to compose this public dataset.

B.1 Unsupervised Settings

ShanghaiTech [22] is one of the largest datasets in the field of video anomaly detection, and it includes videos collected from 13 cameras around the ShanghaiTech University campus. UBnormal [2], a recent synthetic supervised open-set benchmark, includes both normal and abnormal actions within its training set. We refer to the unsupervised learning setting in [13]. Fig.1 illustrates our strategy for achieving heterogeneous data partitioning among clients, applied to the ShanghaiTech and UBnormal datasets. We divide each dataset into K subsets (clients), where each subset contains data from one or two unique perspectives or scenarios. For ShanghaiTech and UBnormal, we set K to 7 and 6, respectively.

Algorithm 1 Federated Learning Video Anomaly Detection (FedVAD)

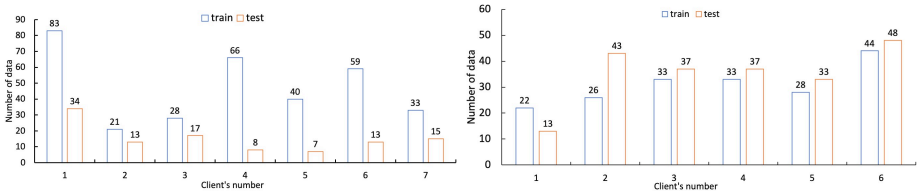
Require: Public dataset D_p, Local datasets $\{D_k^{\text{train}}, D_k^{\text{test}}\}$, Number of clients K, Local training epoch E, Number of global communication rounds R, Number of server groups M, Server with pre-trained visual model Φ and textual model Ψ.

procedure SERVER EXECUTES
  for $r = 1$ **to** R **do**
    $\{W_i\}_{i=1}^{K}, \{O_i\}_{i=1}^{K} \leftarrow$ ClientUpdate()
    for $k = 1$ **to** K **do**
      Calculate distance w_{ik}, o_{ik}
      Update cluster assignment r_{ik}
    end for
    Clustering into M groups $\{\Omega_m\}_{m=1}^{M}$
    for $m = 1$ **to** M **do**
      Select video text pairs $\langle V_{i^*}, T_{j^*} \rangle$ from D_p adapted to group m
      $Z_m \leftarrow \text{MHSA}(\text{LN}([\Phi(V_i); \Psi(T_j)]))$
      $L_{\text{distill}} \leftarrow \frac{1}{N} \sum_{n=1}^{N} \|S_m - Z_m\|^2$
    end for
    Distribute $\{W_i\}_{i=1}^{K}$ to each corresponding client
  end for
end procedure

procedure CLIENTUPDATE
  for $k = 1$ **to** K **do**
    for $e = 1$ **to** E **do**
      $W_k \leftarrow \text{Train}(D_k, W_k)$
    end for
    $O_k \leftarrow \text{ExtractObjects}(D_k)$
    Send W_k, O_k to server
  end for
end procedure

B.2 Weakly Supervised Settings

UCF-Crime [36] serves as a pivotal benchmark in weakly supervised video anomaly detection, offering an extensive video repository across 13 anomaly classes. These videos, captured in diverse settings like streets, family rooms, and shopping centers, provide a rich array of scenarios for analysis. XD-Violence [31] stands as the largest dataset in this domain, compiled from varied sources such as film clips, game footage, and in-car camera recordings, spanning six anomaly categories. Following [29], we adopt a 10-crop augmentation strategy for the UCF-Crime and ShanghaiTech datasets, encompassing the center, four corners, and their mirrored counterparts. For the XD-Violence dataset, a 5-crop strategy is implemented, consisting of the center and four corners. As depicted in Fig.3, we partition the UCF-Crime and XD-Violence datasets into subsets corresponding to their number of anomaly categories, which are 13 and 6, respectively. During this partitioning, we ensure an even distribution of normal category data across



(a) U-ShanghaiTech **(b)** U-UBnormal

Fig. 1: Number of samples on each client for different datasets. U denote unsupervised settings.

Table 1: Clients quantitative comparison of different FL methods in unsupervised and weakly supervised settings. C* indicates the number of the client.

Method	UBnormal (Unsupervised)							XD-Violence (Weakly Supervised)						
	C1	C2	C3	C4	C5	C6	Avg	C1	C2	C3	C4	C5	C6	Avg
FedAvg	87.23	69.36	59.81	60.88	66.14	60.26	67.27	83.23	58.44	95.54	55.52	29.66	64.46	64.43
FedProx	88.25	69.85	60.93	61.40	66.29	60.39	68.08	83.12	58.06	96.71	55.53	29.42	64.73	64.47
FedBN	89.11	69.72	60.87	61.41	66.36	60.51	68.02	81.92	57.94	94.11	53.98	30.01	65.32	63.93
FedRep	88.33	69.11	60.02	60.32	66.57	59.72	67.95	83.35	58.88	96.05	52.83	30.40	**66.33**	64.53
FedALA	88.52	69.16	59.38	60.28	65.79	60.44	68.04	83.39	59.69	95.55	56.74	29.78	62.50	64.66
FedPAC	89.46	69.25	60.35	60.29	65.80	59.39	68.24	83.45	60.25	95.66	57.44	30.15	62.69	64.98
FedAvg SGD	89.28	70.12	60.72	61.79	65.76	60.23	67.98	83.01	59.55	95.26	60.20	30.77	59.11	64.65
FedAdam SGD	89.72	69.78	60.72	60.18	66.66	61.96	68.17	83.20	59.62	94.07	60.98	31.40	59.05	64.72
FPS	89.20	69.45	60.37	60.07	66.20	64.75	68.34	82.67	59.74	95.74	59.23	31.28	60.56	64.87
FedFTG	88.36	69.60	60.61	60.56	66.26	65.13	68.42	83.91	**60.84**	95.68	59.34	31.82	60.33	65.32
Ours with DP	**89.82**	**70.87**	60.72	**63.62**	**68.22**	**65.73**	**69.83**	**84.42**	60.41	**96.94**	**61.82**	**31.88**	**65.21**	**66.78**
Ours	89.35	69.97	**60.98**	61.46	66.61	60.83	69.36	83.88	60.31	96.78	61.08	30.24	65.18	66.21

each subset to maintain a consistent ratio with the anomalous categories. The ShanghaiTech dataset is partitioned following a similar approach, aligning with the unsupervised learning framework.

C Weighting effects of distillation loss

The parameter λ plays an important role in our model, reflecting the weighting of distillation losses in each group on the global model importance. In order to delve deeper into the specific effects of λ variations on the model performance, the results of the experiments are presented in detail in Fig.2a. Through the experiments, we observe that there is a slight improvement in the model's AUC metric as λ increases from 0.1 to 0.5. This suggests that at lower values of λ, the distillation loss has a limited effect in enhancing the model performance. However, it is worth noting that a significant improvement in the model's AUC metrics occurs when the λ value is increased to 0.5, a result that suggests that distillation loss plays a key role in optimizing the model's performance after semantic enhancement. This significant performance improvement emphasizes the importance of distillation loss in the training process, especially after semantic information has been integrated into the model. By adjusting λ, we can fine-tune the sensitivity of the model to distillation loss.

Table 2: Clients quantitative comparison of different FL methods in weakly supervised settings. C* indicates the number of the client.

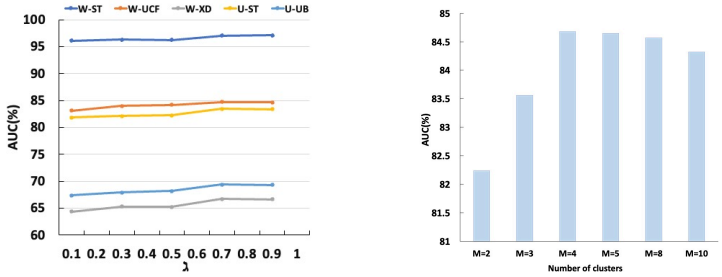
Method	UCF-Crime (Weakly Supervised)													
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	Avg
FedAvg	94.23	83.75	76.62	89.69	87.89	68.43	90.67	85.24	76.51	77.93	59.41	85.38	95.91	82.81
FedProx	93.75	83.86	76.51	88.83	86.31	66.14	85.31	82.40	76.80	77.12	55.15	80.33	88.13	80.39
FedBN	94.07	82.48	76.95	91.22	88.87	70.35	90.54	84.73	77.08	78.68	57.06	86.16	97.23	82.86
FedRep	94.68	83.70	77.13	92.00	89.35	70.51	91.14	84.21	76.29	78.52	59.37	84.90	96.19	82.92
FedALA	94.26	83.17	77.35	92.28	89.30	70.58	91.09	84.73	77.13	**78.79**	57.47	86.04	96.87	82.97
FedPAC	94.04	83.88	77.11	93.67	89.51	70.72	91.56	84.93	76.70	78.39	59.44	86.30	97.75	83.34
FedAvg SGD	93.84	82.45	76.30	93.35	88.81	71.32	91.15	85.36	76.75	77.88	60.18	85.70	96.17	83.02
FedAdam SGD	94.72	82.69	76.05	93.40	88.42	73.56	91.62	84.71	77.17	77.21	60.53	86.57	96.90	83.35
FPS	93.07	82.08	77.51	94.81	89.38	72.43	91.57	84.09	77.87	77.01	61.25	85.98	97.53	83.43
FedFTG	94.08	82.06	76.11	95.70	89.03	72.44	91.61	85.75	76.74	78.50	60.83	85.61	97.69	83.55
Ours with DP	**95.47**	**84.63**	**77.80**	96.44	**91.60**	**75.66**	**93.09**	**86.68**	**78.32**	**80.87**	60.32	**87.74**	97.07	**85.13**
Ours	94.77	84.27	77.67	**96.51**	90.47	74.28	92.09	85.47	77.19	78.51	**63.37**	87.08	**98.47**	84.68

D Experimental results on each client

Experimental results for each client on the UBnormal dataset in the unsupervised setting, and on the UCF-Crime and XD-Violence datasets in the weakly supervised setting, are compared across different methods. As illustrated in Tab.1 and Tab.2, it is evident that our method outperforms other federated learning approaches in most clients. Traditional FL methods (FedAvg and FedProx) often employ a global optimization strategy aiming for uniformity across clients. Unfortunately, they may overlook the heterogeneity in client data distributions, potentially hampering model performance for specific clients. In contrast, personalized FL methods prioritize tailoring to individual client data distributions, thereby embedding client-specific insights into local training. In contrast to other pre-trained FL methods, our approach introduces Adaptive Semantic-Enhanced Distillation to produce a unique global model for each cluster. This strategy not only amplifies the client's learning proficiency on local datasets but also markedly improves the model's generalization capabilities with unseen data. Empirical evidence demonstrates that our approach substantially outperforms both traditional federated learning frameworks and existing personalized methods across various datasets.

E Effect of different number of clusters M

Our evaluation on the UCF-Crime dataset focuses on how different clustering numbers (M-value) influence model performance. Fig.2b demonstrates that increasing the number of clusters, particularly in scenarios with numerous clients, results in a noticeable performance pattern. Initially, as clusters become more granular, the model's performance is enhanced. However, performance declines beyond a certain threshold, possibly due to overfitting or exceeding the data's inherent structure, leading to redundancy and inefficiency. Excessive clustering might also limit data availability per cluster, potentially impacting the model's training stability and generalization. These findings highlight the critical role of optimal cluster number in boosting model efficacy.



(a) Effect of distillation loss weighting factor λ.

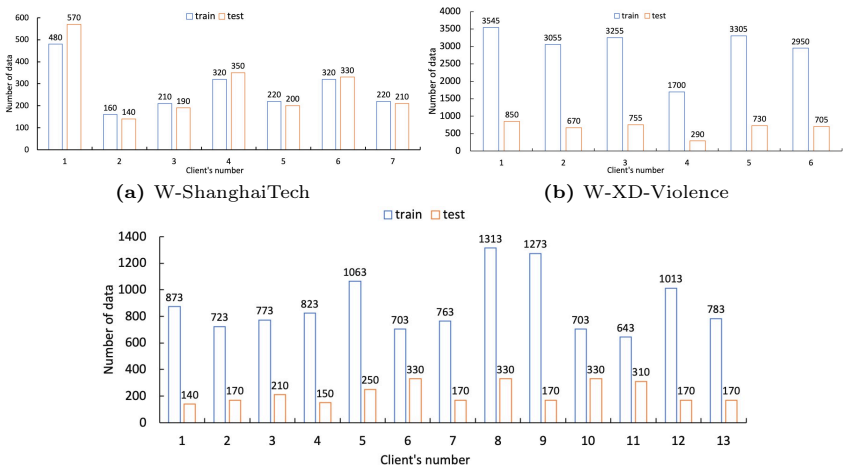
(b) The effect of different number of clusters on the UCF-Crime dataset.

Fig. 2: Effects of different parameters.

F Effects of hyperparameters α and β

In our study, we introduce two important hyperparameters: the weighting factors α and β, which regulate the effects of model weights and object distribution on the determination of cluster membership, respectively. By systematically varying these parameters, we assess their specific impact on the quality of clustering, and the results of these experiments are presented in Fig.4a to Fig.4e.

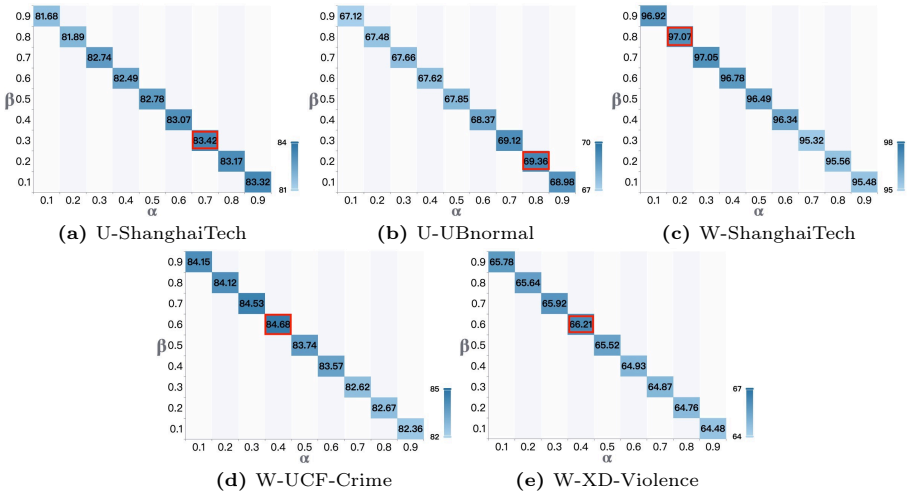
We observe that the weighting factor α plays an important role in the unsupervised learning environment for model performance improvement. Specifically, we partition the clients according to the camera's viewpoint and emphasize the importance of video foreground features in anomaly detection. Enhancing the model's sensitivity to foreground features by increasing the weight factor α can effectively improve the clustering quality and the accuracy of anomaly detection. This strategy is more suitable for scenes where foreground actions or events are key to discriminating between normal and abnormal situations. In contrast, in the weakly supervised learning setting, the client is segmented based on video categories. Since each category may contain multiple foreground and background elements, the model needs to deal with more complex scenarios. The experimental results show that when the weighting factor β is appropriately higher than α, the model is able to better synthesize these complex features, thus improving the performance. This suggests that balancing the weights of different features is crucial for optimizing model performance on datasets containing diverse scenarios and categories.



(a) W-ShanghaiTech **(b)** W-XD-Violence

(c) W-UCF-Crime

Fig. 3: Number of samples on each client for different datasets. W denote weakly supervised settings.



(a) U-ShanghaiTech **(b)** U-UBnormal **(c)** W-ShanghaiTech

(d) W-UCF-Crime **(e)** W-XD-Violence

Fig. 4: Effect of variation of weight factors α and β of clustering on experimental results. U and W denote unsupervised and weakly supervised settings, respectively.