Community-Aware Federated Video Summarization

Fan Wan[1,*], Junyan Wang[2,*], Haoran Duan[1], Yang Song[2], Maurice Pagnucco[2], Yang Long[1,†]
{fan.wan, haoran.duan, yang.long}@durham.ac.uk,
{junyan.wang, yang.song1}@unsw.edu.au and morri@cse.unsw.edu.au
[1]Department of Computer Science, Durham University, Durham, UK
[2]Department of Computer Science and Engineering, University of New South Wales, Sydney, Australia

2023 International Joint Conference on Neural Networks (IJCNN) | 978-1-6654-8867-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/IJCNN54540.2023.10191101

***Abstract*—Video summarization aims to extract representative frames to retain high-level information. Increasing concerns about privacy issues have been raised because conventional large-scale training requires users to upload video samples that may inevitably release sensitive information. In this paper, we thoroughly discuss the Federated Video Summarization problem, *i.e.,* how to obtain a robust video summarization model when video data is distributed on private data islands. Our key contribution includes 1) We propose a fundamental Frame-Based aggregation method to video-related tasks, which differs from the sample-based aggregation in conventional FedAvg. 2) To mitigate the heterogeneous distribution due to community diversity, we propose the Community-Aware Clustering Federated Video Summarization Framework (CFed-VS) that clusters clients via a novel data-driven clustering approach. 3) We further tackle the challenging non-IID setting with a proposed Mixture Transformer, which manifests state-of-the-art performance via extensive quantitative and qualitative experiments on TVSum and SumMe datasets.**

***Index Terms*—Federated Learning, Vision Transformer, Video Summarization**

I. Introduction

With the tremendous growth of video material, automatic tools for understanding and analyzing video content have become an increasingly urgent need. Recent statistics have shown that it will take more than 82 years for a person to watch all videos uploaded to YouTube per day [1]. A promising remedy is that automatic video summarization can enable human users to quickly identify the key content of videos and accelerate knowledge gain and information retrieval. Such a technology has been applied in many scenarios, such as fast indexing and online video recommendation. However, to provide information-rich video summarization and satisfy the wide variety of user needs, existing approaches rely on the large-scale collection of video data and important score annotations to train a robust model. Increasing awareness and concerns about privacy restrictions, e.g., the EU's General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA), have become one of the largest challenges in this domain. Moreover, huge communication costs are incurred during data transmission, which also impedes the development of video summarization technologies.

This project is supported by International Exchanges 2022 IEC\NSFC \223523 and Securing the Energy/Transport Interface EP/X037401/1.

[*] Contributed Equally.

[†] Corresponding Author.

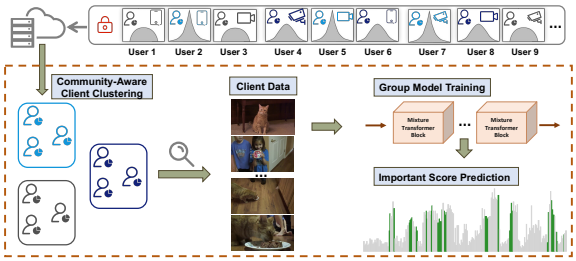


Fig. 1: Community-Aware Federated Video Summarization aims to deploy large-scale VS task training when video data are distributed on edge devices. Based on the similarity of data distribution across clients, the server will cluster clients before FL model training, and then maintain multi-group models to address statistical heterogeneity challenges.

As a burgeoning machine learning scheme, Federated Learning (FL) [2] aims to tackle the problem of data islands while preserving the privacy of data. The key idea of FL is to jointly train global models across various edge devices without collecting any private raw data from the client, thus effectively alleviating user concerns about data privacy and reducing the high costs associated with transmission. Along with its promising prospect, research on FL faces key challenges from high statistic heterogeneity [3]. Concretely, Federated learning relies on Stochastic Gradient Descent (SGD). Jointly training a global model with the Independent and Identically Distributed (IID) statistics from various clients is equivalent to the IID sampling of training data in a centralized training paradigm, which ensures the stochastic gradient is an unbiased estimate of the total gradient during FL training [4]. FL has been applied in smart city [5], recommender system [6] and healthcare [7] fields. Despite the successful applications of FL in the above areas, introducing FL to solve data privacy problems and high transmission costs in traditional VS tasks is not trivial. In the real world, data distribution can easily appear in the distribution of non-IID, and video data has an even stronger bias and diversity according to the photographer's preference. Modeling on the non-ID data with FL paradigm will lead to client drift issues [8], which will lead to performance degradation and slow convergence speed of the global model.

A unique property for VS problems is the user community heterogeneity due to the diverse user profile, preference and behavior. To tackle problems due to statistical heterogeneity in FL, recent attempts [9]–[11] aim to cluster the clients based on model parameters or gradients and maintain a multi-group

model. Nevertheless, the server may be required to wait for extra communication rounds before receiving parameters of the client model with significant changes to calculate the similarity for the clustering procedure, which will lead to the deterioration of model training efficiency and an increase in communication costs. [12] proposed a novel data-driven approach to calculate the similarity of client data distribution, in which clients are grouped based on their similarity in two types of summaries of client data distribution: label distribution, and conditional features distribution. However, it is unrealistic to apply the proposed methods [12] directly to cluster clients in VS tasks, since some video data lack specific categories, using the average feature is also impractical due to the different lengths of the videos.

To our best knowledge, this is the first work to explore the feasibility of the Federated Learning Video Summarization (FLVS) task, and we first established the baseline of FedAvg [2] in FLVS. According to our initial observations and analysis, we proposed a technical roadmap with three key directions for the FLVS problems: 1) In contrast to traditional FL using sample-based aggregation, we explore the **Frame-Based** FedAvg in FLVS tasks so that the length of video is taken into account when assigning weight contributions of client models. 2) We observe that the community factor is the key impact on the heterogeneous data distribution. A novel **Community-Aware** Clustering FL framework for Federated Video Summarization (CFed-VS) is thus proposed, which *clusters clients based on the relative distance between the data distribution of each client*, as shown in Figure 1. It is worth noting that our CFed-VS requires only one-off clustering operation compared to traditional clustering-based FL training, which improves the training efficiency of global models. 3) We then propose the **Mixture Transformer** for obtaining better model generalization in the non-IID setting for learning time-series data. In summary, the key contributions of our work are as follows:

- We propose a more effective frame-based aggregation method of FedAvg for video-related tasks, and systematically analyze the assignment of client model contribution.
- A novel clustering federated framework is proposed to leverage the relative distance between the data distributions of each client, to tackle the challenge of heterogeneous data and reduce the computation cost during the clustering process.
- Mixture Transformer is proposed to enhance model generalization in the non-IID setting, and extensive experiments demonstrate state-of-the-art performance on SumMe and TVSum datasets.

II. Related Work

A. Video Summarization

Video summarization is one of the most important directions in video recognition [13]–[17] to generate [18], [19] a short video clip while keeping the main content or stories of the original video [13], [20]. Recently, several video summarization approaches have been proposed, and they can fall into two broad categories. One of them refers to unsupervised learning, which uses manually designed criteria to prioritize and select frames or subshots from original videos [20], [21]. Another one is supervised learning, which utilizes human-edited examples to learn how to summarize novel videos [22], [23]. Also, some LSTM-based deep learning approaches have been proposed for both supervised and unsupervised video summarization. [24] specified a generative adversarial framework that consists of the summarizer and discriminator for unsupervised video summarization. Wang *et al.* [25] proposed a novel model named Dual Mixture Attention (DMASum) with meta-learning, which solved the softmax bottleneck problem in video summarization.

B. Federated with Statistic Heterogeneity

Statistic heterogeneity (also named non-IID) is one of the major challenges in federated learning. The widespread aggregation strategy in federated learning, FedAvg [2] suffers performance deterioration on non-IID due to client drift issues [8]. To address this problem, a line of research focuses on learning a single global model under the non-IID setting [26]–[28]. For example, FedProx [26] adds a proximal term to the local objective of the client to effectively limit the impact of abnormal local model updating. Another line of research overcomes this problem via personalized federated learning (PFL) [29]–[31], which seeks to personalize the global model for each client. PFL has been adopted in many approaches, including model-agnostic meta-learning [29], model regularization [30], and multi-task learning [31]. Cluster-Based Federated Learning (CFL) [9]–[11] incorporates personalization at the group level while keeping the benefits of PFL. Prior works cluster clients based on the similarity of client model parameters or gradients. However, obtaining a significant change in the client's parameter or gradient, requiring the server to wait for additional communication rounds, greatly reduces the training efficiency. To this end, we develop a time-series similarity method to generate a summary of data distribution under the privacy specification of FL, then the server clusters the client based on the summary before initiating FL training.

C. Vision Transformer

Following Transformer in NLPs [32], Vision Transformer (ViT) [33]–[35] has made great successes in various vision tasks, including object detection [36], semantic segmentation [37], [38], action recognition [39], and so on. For example, Li *et al.* [40] propose a hierarchical Transformer (Swin Transformer) whose representation is computed with Shifted windows that can bring greater efficiency by limiting self-attention computation to non-overlapping local windows while also allowing for cross-window connection. Moreover, the great success of image Transformers has led to the investigation of Transformer-based architectures for video understanding tasks [39]. For instance, UniFormer [41] integrates the merits of 3D convolution and spatio-temporal self-attention in a concise transformer format, and achieves a preferable balance between computation and accuracy.

III. Methodology

A. Rethinking FedAvg in Video Summarization

Recently, many tasks have successfully adapted federated learning (FL) paradigms, while few studies [42] have examined FL applications on video data, especially on video summarization. To explore the feasibility of federated learning on video summarization, we first investigate the characteristic of FedAvg [2], the widely used federated learning method. Then we provide an in-depth analysis of federated model training on video summarization datasets.

Revisiting FedAvg. For a learning problem, we define $f(\theta) = \mathcal{L}(x, y, \theta)$ as the loss of the prediction on an input pair (x, y) with model parameter θ. In federated learning, we assume there are K participating clients each with its own data distribution $\mathcal{P}_k$. Thus, the objective function is defined as:

$$\min_{\theta} f(\theta) = \sum_{k=1}^{K} \frac{n_k}{n} F_k(\theta), \;\; F_k(\theta) = \frac{1}{n_k} \sum_{i \in \mathcal{P}_k} f_i(\theta) \;, \qquad (1)$$

where n is the number of total samples across all clients, n_k is the number of samples that reside in k-clients. Conventional federated learning methods, like FedAvg, optimize Eq. 1 with the following steps. (i) At each communication round t, the server randomly selects K clients available for training, then sends the global model $\hat{\theta}^t$ to the selected clients and deploys it as $\theta_k^{(t)}$. (ii) Each selected client then trains its model $\theta_k^{(t)}$ locally with its own data distribution $\mathcal{P}_k$ for E_{local} epochs. (iii) The server waits until all selected devices have uploaded corresponding parameters $\theta_k^{(t+1)}$ to aggregate the new global model via $\hat{\theta}^{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} \theta_k^{(t+1)}$. The above process will be repeated until the model reaches convergence.

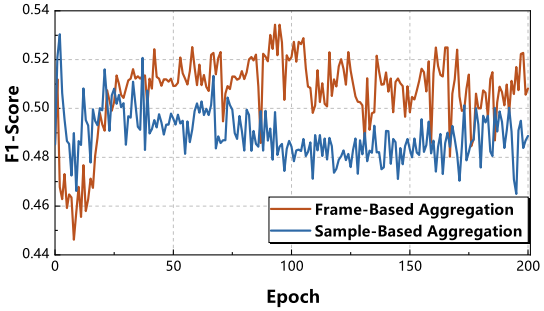


Fig. 2: Comparison of two weight aggregation strategies on the TVSum dataset. Experiments are conducted in IID data distribution with the same settings.

Frame-based FedAvg in Video Summarization. According to the FedAvg algorithm, the weight contribution of the client model is based on the number of samples, *i.e.,* $\frac{n_k}{n}$. However, we assume that directly applying this sample-based FedAvg to video summarization is impractical, as the length of the video is also important for model training in video understanding tasks [43]. To this end, we propose to apply video frames $\frac{v_k}{v}$ instead of video samples $\frac{n_k}{n}$ as:

$$\min_{\theta} f(\theta) = \sum_{k=1}^{N} \frac{v_k}{v} F_k(\theta) \;, \qquad (2)$$

where v_k is the total number of frames that resides in k-client, and v represents the total number of frames across all clients. To verify our assumption, we conducted different aggregation strategy experiments by using vsLSTM [44] on the TVSum dataset, which is shown in Figure 2. It can be observed that the frame-based aggregation can quickly outperform the baseline and converge to more reliable performance. In this analysis, the frame-based aggregation strategy is 1.2% higher than the sample-based aggregation strategy (conventional FedAvg aggregation strategy). The proposed frame-based aggregation is particularly useful when the length of user video is different.

TABLE I: F1-score (%) of different data distribution on both TVSum and SumMe datasets, using vsLSTM [44] baseline. "Max-Min F1-Score" represents the difference between the best and worst results tested by the global model on all clients. "Mean F1-Score" represents the average F1-Score across all clients. "# Round to Reach Target F1-Score" donates the communication rounds of the global model to reach target F1-Score. The "Target F1-Score" was chosen 48% and 30% for TVSum and SumMe datasets separately.

Dataset	Federated				Centerlized
	# Classes	Max-Min F1-Score	Mean F1-Score	# Round to Reach Target F1-Score	F1-Score
TVSum	1	10.98	49.29	75	54.27
	4	5.21	52.08	46	
	10 (IID)	2.23	53.14	17	
SumMe	1	12.19	33.62	64	37.72
	3 (IID)	4.87	36.48	18	

B. Non-IID Data Distribution Analysis

To analyze the data distribution on video summarization datasets by applying federated learning, we simulate non-IID and IID settings by forcing each client to have limited classes, as shown in Table I. We follow the setting in [9] and simulate ten clients to run the global model within 200 communication rounds. Each participating client runs the global model with 20 local epochs. We then obtain three folds of insights as follows.

- Compared to the centralized paradigm, the performance of using FedAvg (frame-based) would not decrease too much, which also verifies the feasibility of using federated learning in video summarization tasks.
- With the decrease of classes in each client, the performance of the global model will decrease from 53.14% to 48.29%, and the performance of "round to reach" will increase from 17 to 75. These results indicate that higher data heterogeneity would affect the model performance and convergence speed.
- The value of "Max-Min F1-Score" increased with the decrease of categories in each client, which indicates that the global model is difficult to generalize on all the clients in non-IID settings.

Therefore, using federated learning in video summarization tasks will face high data heterogeneity (non-IID) challenges, which affects the global model in terms of performance and convergence speed. To address the above challenges, we

propose the Community-Aware Clustering Federated Video Summarization strategy by clustering the clients with similar data distribution, and training corresponding group models.

C. Community-Aware Federated Video Summarization

There has been significant research investigating high data heterogeneity challenges [3], [10], [45] due to the community diversity of users. Specially, cluster-based federated learning clusters the clients based on the similarity of the model parameters [9]–[11] or data distribution [12] across different clients recently. Parameters-based clustering algorithm requires the server to consume additional communication rounds to obtain significantly varying gradients or parameters, thereby reducing model training efficiency and increasing communication costs. To this end, we propose a data-driven approach via leveraging the relative distance between the data distribution of each client, and clients with a close distance of data distribution can be clustered as a community before FL model training.

Algorithm 1 Community-Aware Federated Video Summarization

Input: $\mathcal{C} \leftarrow$ Client groups;
$X_k \leftarrow$ Dataset of k-client with U samples;
$M_d \leftarrow$ the distance matrix of the data distribution;
$DTW() \leftarrow$ dynamic time warping function;
Output: Uploaded group model parameters $\theta_{c,t}$

procedure CFED-VS
  Server broadcast $\mathbf{x}_p$ to all K clients
  for $k \in K$ **do**
    $d^k_{\text{center}} = \frac{1}{U}\sum_{u=1}^{U} \text{DTW}(\mathbf{x}_u, \mathbf{x}_p)$
    Client k upload d^k_{center} to server
  end for
  $M_d = \{|d^k_{\text{center}} - d^q_{\text{center}}|\} \mid k, q \in K \text{ and } k \neq q$
  $\mathcal{C} \leftarrow \mathcal{F}_{K-means}(M_d, m)$
  while $c \in \mathcal{C}$ **do**
    $\theta_{c,0} \leftarrow \theta_0$
    for `each round` $t = 1, 2, \ldots, T$ **do**
      $\theta_{c,t+1} \leftarrow$ FEDAVG$(\theta_{c,t}, K_c)$
    end for
  end while
end procedure

function DTW($\mathbf{x}_1$, $\mathbf{x}_2$)
  $l_1 \leftarrow$ `length of` $\mathbf{x}1$; $l_2 \leftarrow$ `length of` $\mathbf{x}2$
  for $i = 1, 2, \ldots, l_1$ **do**
    for $j = 1, 2, \ldots, l_2$ **do**
      $cost = d(\mathbf{x}_1[i], \mathbf{x}_2[j])$
      $\mathcal{D}[i, j] = cost + Min(\mathcal{D}[i-1, j], \mathcal{D}[i, j-1], \mathcal{D}[i-1, j-1])$
    end for
  end for
  return $\mathcal{D}[l_1, l_2]$
end function

Community Distribution Estimation. Due to the data privacy policy in federated learning, it is not allowed to directly calculate the distance between any two private data distributions among clients. We thus set a proxy sample from another public dataset to collect the comparison of distances simultaneously. By calculating the distance between proxy samples and the center of all training samples of each client, we can obtain the essential information on the data distribution in each client, which can be regarded as the summary of data distribution for all clients. Due to the property of video data (time-series), we adopt the widely used Dynamic Time Warping (DTW) [46] as the distance measurement.

Concretely, the centre server first broadcast a single proxy sample $\mathbf{x}_p$ to each available client k. Then the client calculates the pair-wise distance $\mathcal{D}[\mathbf{x}_u, \mathbf{x}_p]$ between each training sample $\mathbf{x}_u$ and the proxy sample $\mathbf{x}_p$ via the DTW distance function, where $\mathbf{x}_u$ is the sample in the local dataset $X_k = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_u, \ldots, \mathbf{x}_U\}$. Finally, we estimate the distance d^k_{center} from the center of data distribution of k-client to the proxy sample $\mathbf{x}_p$. We regard d^k_{center} as the summary of each client's data distribution, and the detailed implementation can be seen in Algorithm 1.

Clustering Procedure. Once each device sends its summary of data distribution d^k_{center} to the central server, the server calculates the pair-wise distance between any two clients' summary via L1 distance, i.e. $|d^k_{\text{center}} - d^q_{\text{center}}|$, where k, q denotes the different client. We store all pair-wise values of various clients to form the distance matrix M_d. Then we adopt the K-Means algorithm [47] to cluster the client into the m group based on the distance matrix. An overview of the clustering procedure is also described in Algorithm 1 and an illustration of the clustering procedure is shown in Fig 3. After completing the clustering operation, we train m cluster-wise models via the proposed Frame-Based FedAvg.

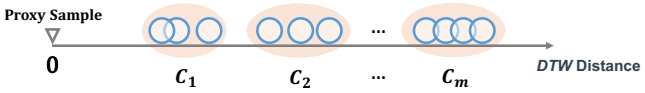


Fig. 3: The illustration of clustering client methods in CFed-VS. The blue circle denotes the distance from the proxy sample to the data centre of various clients. The orange oval represents different client groups and indexes by $\mathcal{C}_m$.

Our proposed CFed-VS strategy allows for a one-off clustering of participating clients into various groups before starting federated training. In the event that new clients join, the server can send only the proxy sample and assign these clients to the appropriate group based on their feedback summary. Compared to other grouping methods based on data distribution, our approach does not reveal the user's data category information and only requires a certain amount of computing resources on the client side.

D. Mixture Transformer

Even though the proposed CFed-VS provides a solution to non-IID performance decline, there is still a performance gap when using conventional video summarization models

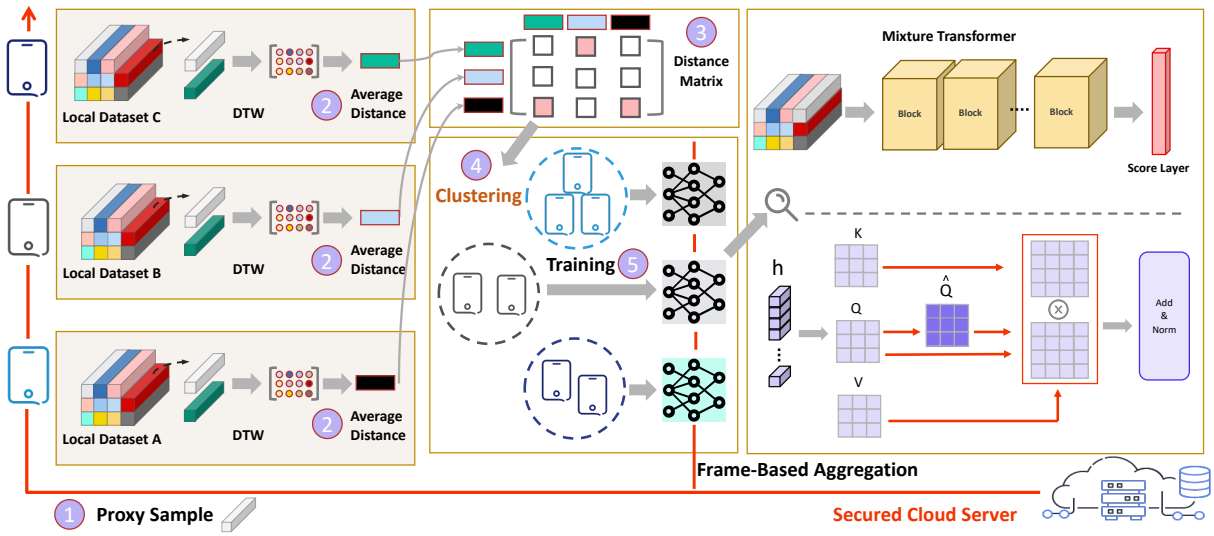


Fig. 4: Overview of the Community-Aware Federated Video Summarization system. The secured cloud server firstly clusters clients into multi-groups based on the similarity of data distribution before the FL training procedure. Then the Mixture Transformer model can be deployed to each client to carry out training.

with a non-IID data setting. We then focus on improving the model generalization in the non-IID setting. Considering the outstanding performance of the Transformer [32] for learning time-series data, the global receptive field of the self-attention mechanism can help the model focus on summarizing global information of a video sequence, which is suitable for the video summarization task. Therefore, we propose the Mixture Transformer to address the above challenges.

Transformer Block. In a basic Transformer block, the queries, keys, and values $Q = HW_q$, $K = HW_k$, and $V = HW_v$ are linear projections of the input frame feature H with $Q, K, V \in \mathbb{R}^{N\times d}$, where N and d denote the frame number and channel dimension. The process is defined as:

$$\mathcal{A}(\boldsymbol{K},\boldsymbol{Q},\boldsymbol{V}) = \mathcal{F}_{Softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{D_a}})V \ , \tag{3}$$

$$H_a^l = MLP(LN(\mathcal{A}^l) + \mathcal{A}^{l-1}) + H_a^{l-1}, \tag{4}$$

where $\mathcal{A}$ denotes the attention map, which is computed as the scaled dot-product function, and l denotes the block number. LN and MLP represent Layer Normalization and Muti-Layer Perception, respectively. In the context of video summarization, the log probability matrix $\mathcal{A}$ becomes a high-rank matrix when the visual contents are complex and the changes between frames are severe. Such high-rank matrix applied *softmax* function will face the **Softmax bottleneck**, as discussed in [48]. It reflects the circumstance that *softmax* function does not have the capacity to express the true attention distribution when d is smaller than $rank(\mathcal{A}) - 1$. Inspired by the work of [25], we apply the Associated Query $\hat{Q} = tanh(W^{\hat{Q}}Q)$ to capture the second-order changes between queries so that complex video content can be represented in a more smoothed attention representation. Then the attention map in Eq. 3 is re-computed as:

$$\hat{\mathcal{A}} = \mathcal{A}(\boldsymbol{K},\boldsymbol{Q},\boldsymbol{V})\cdot\mathcal{A}(\boldsymbol{K},\hat{\boldsymbol{Q}},\boldsymbol{V})^T \ , \tag{5}$$

where $\hat{\mathcal{A}} \in \mathbb{R}^{T\times T}$, namely mixture attention map. $W^{\hat{Q}}$ is the Associated Query parameter. As $\hat{\mathcal{A}}$ is a non-linear function of the attention distribution, the rank of $\hat{\mathcal{A}}$ can be arbitrarily higher than the standard attention map $\mathcal{A}$, which can be used to alleviate the bottleneck problem.

Overall Framework. The architecture overview is shown in Figure 4. Firstly, the secured cloud server sends a proxy sample to each user. According to the received feedback, users with similar distances will be clustered as a community. Each community trains a unique Mixture Transformer model using the proposed Frame-Based Aggregation. New users will be assigned to the correct community by comparing their feedback summary in the distance between the proxy sample and the center of their data distribution.

IV. Experiment

A. Experimental Setup

Datasets. We evaluate our model on two public datasets: TVSum [49] and SumMe [50]. TVSum was collected from Youtube, which contains 50 videos in 10 categories. The duration of most videos ranges from 1 to 10 minutes. SumMe includes 25 videos with various holidays, events and sports. The video lengths vary from 1.5 to 6.5 minutes. Both datasets contain annotations labeled for key-frames by 25 human annotators. Considering the non-IID setting is category-based [3], we manually flag each video's category as the absence of detailed categories for each video on two datasets. Furthermore, the limited data provided by TVSum and SumMe cannot meet the splitting method in non-IID setting, thus we used the ball-and-urn technique [51] to split a single video sample into multiple fragments, increase the number of samples in TVSum and SumMe to 150 and 100 respectively, and then we followed Pathological non-IID setting [3] to assign client data.

Evaluation Metrics As for the evaluation metrics for the **VS** task, we used the key-shot-based F-score [23] as the

metric, and the converted frame-level importance scores to shot-based summaries for all datasets. The kernel temporal segmentation (KTS) [20], where the method can segment the video into separate intervals in time, was used to change the user annotation from frame to key shot level in our experiment. Then we calculate the harmonic average F-score as the evaluation metric. We also used Kendall's τ [52] and Spearman's ρ [53] correlation coefficients to compare the ordinal correlation between the generated summary and the ground truth. As the metric for **FL**, given numerous devices, we evaluate the corresponding group model based on the client's local test set for the CFL-based framework under the same number of groups. For FedAvg and FedProx, we evaluate the global model on the local test data of all clients.

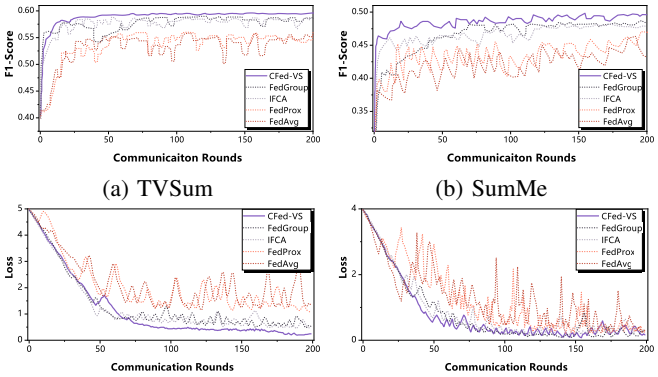
Implement Details For CFed-VS, we set the total number of clients at ten, and the fraction of clients participating in each round of FL is 0.8. The global communication round T is 200, the learning rate is 0.0001, and the local epoch E_{local}=20. For Mixture Transformer, the 1024 dimensional visual features extracted from the $pool5$ layer of the GoogLeNet [54] are used for training, to be consistent with existing methods. Since cluster structures in the real world may be ambiguous, ignoring the knowledge learned by the group model from other communities will reduce the performance of models trained in a single client cluster [55]. Thus, we also adopt the weight-sharing approach [55]. The proxy sample $\mathbf{x}_p$ is selected by selecting a random test sample in SumMe when performing experiments on TVSum, and vice versa. Besides following previous work, frames feature H are first extracted by I3D [56] whose dimensionality is 1024, and each video is segmented into shots by KTS [57], which is a widely used video temporal segmentation method in the video summarization task.

B. Experimental Results

Comparison to FL Methods. As shown in Table II, we compared our proposed CFed-VS with four FL baseline frameworks and found that CFed-VS achieves the best performance both in two datasets. The result shows that IFCA [55], FedGroup [9] and CFed-VS outperform significantly to other frameworks in most non-IID settings. We attribute this to the CFL-based approaches, which group clients with similar data distribution into the same community, so group models can effectively learn common properties from communities to mitigate the challenge of data heterogeneity. As the convergency speed shown in Figure 5, our proposed approach has a fast convergency speed compared with IFCA and FedGroup, which shows the efficiency of the CFed-VS approach in FLVS tasks.

TABLE II: Comparisons with FedAvg, FedProx, IFCA, FedGroup on TVSum and SumMe dataset.

Dataset	TVSum			SumMe	
# of Classes	1	2	4	1	2
FedAvg [2]	54.23 ± 2.2	54.38 ± 1.5	56.08 ± 0.5	43.78 ± 0.2	49.59 ± 0.5
FedProx [26]	55.65 ± 1.9	54.63 ± 0.6	57.42 ± 1.3	43.62 ± 0.8	50.77 ± 1.2
IFCA [55]	57.29 ± 1.3	57.94 ± 1.7	58.51 ± 1.6	47.23 ± 1.4	51.41 ± 0.5
FedGroup [9]	57.41 ± 1.7	58.39 ± 0.4	58.74 ± 1.9	47.72 ± 1.7	51.59 ± 2.1
CFed-VS	$\mathbf{57.91 \pm 1.4}$	$\mathbf{58.86 \pm 0.4}$	$\mathbf{59.98 \pm 2.7}$	$\mathbf{48.20 \pm 0.8}$	$\mathbf{52.13 \pm 0.5}$



(a) TVSum (b) SumMe

(c) TVSum (d) SumMe

Fig. 5: Evaluation of model performance on TVSum and SumMe under 2-class non-IID and 1-class non-IID separately.

TABLE III: F1-score (%) of DMASum with state-of-the-art approaches on both SumMe and TVSum dataset.

Method	SumMe	TVSum
DPP-LSTM [44]	38.6	54.7
SUM-GAN [58]	41.7	54.3
Cycle-SUM [59]	41.9	57.6
DMASum [25]	54.3	61.4
SumGraph [60]	51.4	**63.9**
RSGN [61]	45.0	61.0
Mixture Transformer	**55.1**	63.8

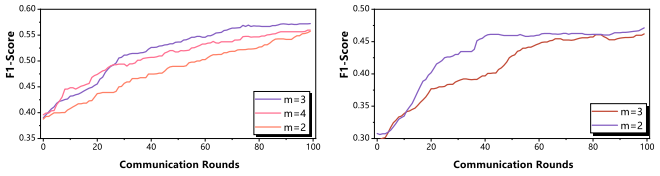
TABLE IV: Rank-order correlation coefficients computed between predicted importance scores by different models and human-annotated scores on both SumMe and TVSum datasets using Kendall's τ and Spearman's ρ correlation coefficients.

Method	SumMe		TVSum	
	τ	ρ	τ	ρ
Random	0.000	0.000	0.000	0.000
DPP-LSTM [44]	-	-	0.042	0.055
SUM-GAN [58]	0.049	0.066	0.024	0.031
SumGraph [60]	-	-	0.094	0.138
RSGN [61]	0.083	0.085	0.083	0.090
Mixture Transformer	**0.102**	**0.107**	**0.098**	**0.149**

Comparison to VS Methods. Our model comparison with state-of-the-art VS methods is summarized in III and IV. From the result of III, it can be seen that Mixture Transformer can achieve competitive performance on both datasets, which indicates that pure Transformer structure can obtain outstanding performance than other models. From IV, we can see the correlation coefficients given by DMASum are significantly higher than other state-of-the-art models, which verify that the mixture attention mechanism itself is capable of improving model generalization.

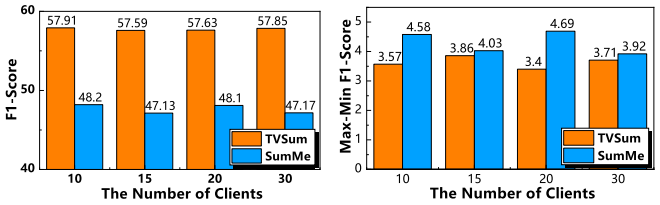
C. Ablation Study

Effect on Group Number We then adopted the proposed CFed-VS approach to cluster the ten clients into 2 to 4 and 2-3 groups in TVSum and SumMe. The performance of CFed-VS



(a) TVSum (b) SumMe

Fig. 6: Evaluation of model performance on TVSum and SumMe with different group numbers.



(a) F1-Score (b) Max-Min F1-Score

Fig. 7: Model performance analysis on TVSum and SumMe with different numbers of clients.

with different group numbers m was conducted in 100 global communication rounds under the 1-class non-IID setting. As shown in Figure 6, CFed-VS can efficiently converge on both datasets and achieves the best performance with groups $m = 3$ and $m = 2$ in TVSum and SumMe.

Effect on Client Number We finally examine the effect of device number K on CFed-VS, in which K = 10, 15, 20, 30 under three groups and two groups in TVSum and SumMe separately. Figure 7 illustrates the performance of CFed-VS for TVSum and SumMe datasets under the 1-class non-ID setting. We observe that the number of devices does not significantly affect the model performance. Meanwhile, the low variance of the Max-Min F1-Score, which measures the difference between the best and worst F1-Score of group models evaluate in all clients, indicates that group models can effectively generalize to the various clients. The results of these experiments indicate that device numbers have a stable impact on the CFed-VS framework.

D. Discussion on Efficiency and Privacy

Conventional parameter-based clustering methods, usually require training a global model in certain communication rounds to determine the difference between client models. Based on the similarity of parameters, the server groups the client and FL training is then performed independently for each client cluster to produce multiple federated models. Our proposed approach clusters the clients before the FL training, which saves the communication rounds for the clustering procedure. In terms of privacy, it is worth noting that the summary of the data distribution sent by the client to the server does not reveal any specific information regarding the video of the user. Since the client sends a single value to the server, our approach is more private than clustering based on label distribution or feature distribution.

V. Conclusion

In this work, we established the first FLVS benchmark with three key technical directions. 1) For the federated learning foundation, we investigated the Frame-Based Aggregation tailored for the VS problem with different video lengths. 2) A Community-Aware Clustering Federated Learning (CFed-VS) framework was proposed. By completing the proxy registration before FL training, community clients with similar distances of data distribution effectively mitigated the data heterogeneity. 3) The proposed Mixture Attention Transformer alleviated the bottleneck problem and significantly improved the model generalization in the non-IID setting. Our thorough evaluation on both datasets suggested favorable outcomes of our method compared to existing established FL and VS frameworks. Future development of the CFed-VS framework can investigate how it could be adapted or improved to handle even larger and more diverse datasets.

References

[1] M. Otani, Y. Nakashima, E. Rahtu, and J. Heikkila, "Rethinking the evaluation of video summaries," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7596–7604.

[2] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[3] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," *arXiv preprint arXiv:1806.00582*, 2018.

[4] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

[5] J. C. Jiang, B. Kantarci, S. Oktug, and T. Soyata, "Federated learning in smart city sensing: Challenges and opportunities," *Sensors*, vol. 20, no. 21, p. 6230, 2020.

[6] A. Hard, K. Rao, R. Mathews, S. Ramaswamy, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, and D. Ramage, "Federated learning for mobile keyboard prediction," *arXiv preprint arXiv:1811.03604*, 2018.

[7] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang, "Federated learning for healthcare informatics," *Journal of Healthcare Informatics Research*, vol. 5, no. 1, pp. 1–19, 2021.

[8] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5132–5143.

[9] M. Duan, D. Liu, X. Ji, R. Liu, L. Liang, X. Chen, and Y. Tan, "Fedgroup: Efficient federated learning via decomposed similarity-based clustering," in *2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*. IEEE, 2021, pp. 228–237.

[10] C. Briggs, Z. Fan, and P. Andras, "Federated learning with hierarchical clustering of local updates to improve training on non-iid data," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–9.

[11] C. Feng, H. H. Yang, D. Hu, Z. Zhao, T. Q. Quek, and G. Min, "Mobility-aware cluster federated learning in hierarchical wireless networks," *IEEE Transactions on Wireless Communications*, 2022.

[12] Y. Wang, J. Wolfrath, N. Sreekumar, D. Kumar, and A. Chandra, "Accelerated training via device similarity in federated learning," in *Proceedings of the 4th International Workshop on Edge Systems, Analytics and Networking*, 2021, pp. 31–36.

[13] J. Wang, Y. Long, M. Pagnucco, and Y. Song, "Dynamic graph warping transformer for video alignment," 2020.

[14] J. Wang, L. Qin, P. Zhang, Y. Long, B. Hu, M. Pagnucco, S. Wang, and Y. Song, "Towards unified multi-excitation for unsupervised video prediction," 2022.

[15] J. Wang, Z. Sun, Y. Qian, D. Gong, X. Sun, M. Lin, M. Pagnucco, and Y. Song, "Maximizing spatio-temporal entropy of deep 3d cnns for efficient video recognition," *arXiv preprint arXiv:2303.02693*, 2023.

[16] J. Wang, B. Hu, Y. Long, and Y. Guan, "Order matters: Shuffling sequence generation for video prediction," *arXiv preprint arXiv:1907.08845*, 2019.

[17] Y. Bai, J. Wang, Y. Long, B. Hu, Y. Song, M. Pagnucco, and Y. Guan, "Discriminative latent semantic graph for video captioning," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3556–3564.

[18] S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks, "Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models," *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[19] Y. Long, L. Liu, F. Shen, L. Shao, and X. Li, "Zero-shot learning using synthesised unseen visual data with diffusion regularisation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 10, pp. 2498–2512, 2017.

[20] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," in *European conference on computer vision*. Springer, 2014, pp. 540–555.

[21] O. Morere, H. Goh, A. Veillard, V. Chandrasekhar, and J. Lin, "Co-regularized deep representations for video summarization," in *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 3165–3169.

[22] M. Rochan, L. Ye, and Y. Wang, "Video summarization using fully convolutional sequence networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 347–363.

[23] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *European conference on computer vision*. Springer, 2016, pp. 766–782.

[24] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial lstm networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 202–211.

[25] J. Wang, Y. Bai, Y. Long, B. Hu, Z. Chai, Y. Guan, and X. Wei, "Query twice: Dual mixture attention meta learning for video summarization," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 4023–4031.

[26] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine Learning and Systems*, vol. 2, pp. 429–450, 2020.

[27] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," *arXiv preprint arXiv:1907.02189*, 2019.

[28] A. K. Sahu, T. Li, M. Sanjabi, M. Zaheer, A. Talwalkar, and V. Smith, "On the convergence of federated optimization in heterogeneous networks," *arXiv preprint arXiv:1812.06127*, vol. 3, p. 3, 2018.

[29] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning: A meta-learning approach," *arXiv preprint arXiv:2002.07948*, 2020.

[30] F. Hanzely and P. Richtárik, "Federated learning of a mixture of global and local models," *arXiv preprint arXiv:2002.05516*, 2020.

[31] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, "Federated multi-task learning," *Advances in neural information processing systems*, vol. 30, 2017.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[33] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[34] H. Duan, Y. Long, S. Wang, H. Zhang, C. G. Willcocks, and L. Shao, "Dynamic unary convolution in transformers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[35] H. Duan, S. Wang, and Y. Guan, "Sofa-net: Second-order and first-order attention network for crowd counting," *arXiv preprint arXiv:2008.03723*, 2020.

[36] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[37] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, 2021.

[38] Z. Wang, X. Li, S. Yu, H. Duan, X. Zhang, J. Zhang, and S. Chen, "Vsp-fuse: Multifocus image fusion model using the knowledge transferred from visual salience priors," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[39] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6836–6846.

[40] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[41] K. Li, Y. Wang, P. Gao, G. Song, Y. Liu, H. Li, and Y. Qiao, "Uniformer: Unified transformer for efficient spatiotemporal representation learning," *arXiv preprint arXiv:2201.04676*, 2022.

[42] Y. Liu, A. Huang, Y. Luo, H. Huang, Y. Liu, Y. Chen, L. Feng, T. Chen, H. Yu, and Q. Yang, "Fedvision: An online visual object detection platform powered by federated learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 08, 2020, pp. 13 172–13 179.

[43] Y. Zhu, X. Li, C. Liu, M. Zolfaghari, Y. Xiong, C. Wu, Z. Zhang, J. Tighe, R. Manmatha, and M. Li, "A comprehensive study of deep video action recognition," *arXiv preprint arXiv:2012.06567*, 2020.

[44] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *European conference on computer vision*. Springer, 2016, pp. 766–782.

[45] M. Xie, G. Long, T. Shen, T. Zhou, X. Wang, J. Jiang, and C. Zhang, "Multi-center federated learning," *arXiv preprint arXiv:2005.01026*, 2020.

[46] M. Müller, "Dynamic time warping," *Information retrieval for music and motion*, pp. 69–84, 2007.

[47] J. MacQueen, "Classification and analysis of multivariate observations," in *5th Berkeley Symp. Math. Statist. Probability*, 1967, pp. 281–297.

[48] Z. Yang, Z. Dai, R. Salakhutdinov, and W. W. Cohen, "Breaking the softmax bottleneck: A high-rank rnn language model," *arXiv preprint arXiv:1711.03953*, 2017.

[49] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "Tvsum: Summarizing web videos using titles," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5179–5187.

[50] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in *European conference on computer vision*. Springer, 2014, pp. 505–520.

[51] G. Conti, "Analytic combinatorics."

[52] M. G. Kendall, "The treatment of ties in ranking problems," *Biometrika*, vol. 33, no. 3, pp. 239–251, 1945.

[53] D. Zwillinger and S. Kokoska, *CRC standard probability and statistics tables and formulae*. Crc Press, 1999.

[54] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[55] A. Ghosh, J. Chung, D. Yin, and K. Ramchandran, "An efficient framework for clustered federated learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 19 586–19 597, 2020.

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

[57] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," in *European conference on computer vision*. Springer, 2014, pp. 540–555.

[58] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial lstm networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 202–211.

[59] L. Yuan, F. E. Tay, P. Li, L. Zhou, and J. Feng, "Cycle-sum: Cycle-consistent adversarial lstm networks for unsupervised video summarization," *arXiv preprint arXiv:1904.08265*, 2019.

[60] J. Park, J. Lee, I.-J. Kim, and K. Sohn, "Sumgraph: Video summarization via recursive graph modeling," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*. Springer, 2020, pp. 647–663.

[61] B. Zhao, H. Li, X. Lu, and X. Li, "Reconstructive sequence-graph network for video summarization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2793–2801, 2021.