Data Science and Engineering (2026) 11:66–83
https://doi.org/10.1007/s41019-025-00315-9

RESEARCH PAPERS

# CAFL: Conditional Attention Federated Learning for Image Emotion Analysis

Chang Liu[1,2,3,4] · Zengmao Wang[1,2,3,4] · Yongchao Xu[1,2,3,4] · Bo Du[1,2,3,4]

Received: 9 September 2024 / Revised: 15 July 2025 / Accepted: 18 August 2025 / Published online: 13 November 2025


## Abstract

The rapid proliferation of images on online platforms has made emotion analysis a task of paramount significance. However, these images are often privacy-sensitive, making Federated Learning (FL) a compelling paradigm over traditional centralized methods. A critical yet largely unaddressed challenge in applying FL to this domain is the severe concept drift stemming from the subjective and culturally diverse nature of emotional expression, which causes conventional FL algorithms to fail. In this paper, we propose CAFL (Conditional Attention Federated Learning) to fill this gap. CAFL empowers clients to learn collaboratively yet personally. It intelligently routes information through an adaptive gate that separates features into a personalized stream and a global stream. These streams are then processed by dedicated local and global prediction heads. Crucially, collaboration is guided by a conditional attention mechanism, where the server computes a personalized reference model for each client based on an attention-weighted aggregation of peer models, promoting knowledge sharing among kindred clients. Extensive experiments on various lightweight foundation models show that CAFL consistently outperforms existing FL methods, demonstrating its robustness and superior performance as a solution for distributed, privacy-sensitive image emotion analysis.

**Keywords** Federated Learning · Personalized Federated Learning · Foundation Models · Image Emotion

## 1 Introduction

The rapid development of online platforms and social media has led to an explosive growth in multimedia content, particularly images [1, 2]. This proliferation of visual data has made image emotion analysis an increasingly significant field of research, with wide-ranging applications in social media analytics, human-computer interaction, and digital marketing [3, 4]. Image emotion analysis aims to automatically recognize and analyze the emotions conveyed in visual content, providing valuable insights into user sentiments and preferences [5, 6].

Traditionally, image emotion analysis has been approached as a classification task, where emotions are represented as discrete labels [7, 8]. However, this simplistic view fails to capture the nuanced and subjective nature of emotions. Recent research has shifted towards emotion distribution learning, which represents emotions as a probability distribution over multiple categories, providing a more comprehensive and realistic representation of emotional content in images [9–11].

✉ Zengmao Wang
wangzengmao@whu.edu.cn

✉ Bo Du
dubo@whu.edu.cn

Chang Liu
computerscience@whu.edu.cn

Yongchao Xu
yongchao.xu@whu.edu.cn

1 School of Computer Science, Wuhan University, Wuhan 430072, China

2 Institute of Artificial Intelligence, Wuhan University, Wuhan 430072, China

3 National Engineering Research Center for Multimedia Software, Wuhan University, Wuhan 430072, China

4 Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan 430072, China

The extraction of discriminative features is key to effective image emotion analysis [12, 13]. Early works relied on hand-crafted features, including low-level holistic features, color, composition, and texture information [5, 7]. With the advent of deep learning, many models have been designed to explore emotions in images, leveraging the power of convolutional neural networks (CNNs) to extract more semantic and context-aware features [8, 14–16].

However, the centralized nature of traditional machine learning approaches presents significant challenges when dealing with large-scale distributed image data. Privacy concerns have become increasingly prominent, with users and organizations becoming more reluctant to share raw image data for central processing [17, 18]. Social images often contain sensitive user information, raising serious privacy issues when using traditional centralized machine learning methods [19, 20].

Federated Learning (FL) has emerged as a promising paradigm to address these challenges [17, 18]. FL enables multiple parties (clients) to collaboratively train models without sharing raw data, thereby preserving data privacy while leveraging the value of distributed data. This approach is particularly well-suited for image emotion analysis tasks, where data is often distributed across numerous end devices with varying computational capabilities and network conditions [21].

However, applying FL to image emotion analysis tasks still faces significant hurdles, with concept drift being the most prominent [22]. Concept drift in federated image emotion analysis refers to the significant differences in data distribution across different clients. This is particularly prevalent in emotion analysis tasks due to the subjective nature of emotions and the diversity of visual expressions across different cultures and individuals [21]. Traditional FL methods often perform poorly when handling such highly heterogeneous data, struggling to learn a global model that performs well across all clients [18, 23].

To address these challenges, we propose CAFL (Conditional Attention Federated Learning), a novel approach that combines the strengths of attentive mechanisms and federated learning to enable effective image emotion analysis in privacy-preserving, distributed environments. CAFL is designed to tackle the concept drift problem by adaptively facilitating collaboration between clients with similar data distributions.

Our work builds upon recent advancements in attention mechanisms [16, 24] and personalized federated learning. By incorporating a conditional attention mechanism, CAFL can dynamically adjust the importance of different clients' contributions based on their relevance to the target task. This approach is particularly well-suited to image emotion analysis, where emotional expressions may have varying degrees of similarity across different client subgroups.

We conducted extensive experiments on multiple lightweight foundation models, such as MobileNet V3 [25], MobileViT V2 [26], ConvNeXt V2 [27], to evaluate the performance of CAFL. Our results reveal that both the pretraining method and the structure of the foundation model significantly impact the performance of federated learning for image emotion analysis [28, 29].

The key contributions of our work are threefold:

1. We propose CAFL, a novel personalized federated learning framework specifically designed to address the concept drift challenge in image emotion analysis by unifying three core mechanisms for targeted collaboration and personalization.
2. We introduce a conditional attention mechanism at the server level, which adaptively facilitates collaboration by computing pairwise model similarities. This encourages clients with similar emotional interpretation patterns (i.e., similar data distributions) to learn more from each other.
3. We design a client-side personalization architecture featuring an adaptive feature gate and a bipartite prediction head. This structure learns to disentangle and process personalized and global features on a per-sample basis, mitigating the negative impacts of data heterogeneity.

## 2 Related Work

### 2.1 Image Emotion Analysis

Image emotion analysis has gained significant attention in recent years due to its wide applications in multimedia retrieval, advertising recommendation, and user experience enhancement [30]. Early works focused on designing hand-crafted features to represent images, including low-level holistic features [7], various element information such as color, composition, and texture [5], and high-level content such as adjective–noun pairs [6].

With the rapid development of deep learning, many deep models have been designed to explore emotions in images [1, 8, 14]. These deep features often achieve better performance than traditional hand-crafted features due to their stronger semantic representational ability. However, current works are still hampered by how to extract more effective emotional content from images. Owing to the subjectivity of human cognition, different regions and contents of an image may contribute differently to the arousal of emotions [31, 32].

Recent studies have proposed attention-based methods to focus on emotion-related regions in images [3, 16, 33]. Nevertheless, these methods do not fully refine how each emotion responds to different regions or explore potential relationships between each emotion and emotion-aware regions. Moreover, the same image may trigger different emotions depending on the viewer's cultural background and social experiences [34], leading to the challenge of concept drift in image emotion analysis. This concept-drift problem naturally motivates adopting federated learning (FL), which leverages diverse data from multiple clients while preserving privacy.

### 2.2 Federated Learning

Federated Learning (FL) is a distributed learning paradigm that enables training models on decentralized data [17, 19]. FedAvg [17] is the most widely used FL algorithm, which iteratively aggregates locally trained models by weighted averaging. However, FedAvg faces challenges when dealing with non-IID data distributions across clients.

To alleviate these issues, several enhanced FL algorithms have been proposed. FedProx [18] introduces a proximal term to limit the impact of local updates and improve stability. FedDyn [35] employs a dynamic regularizer to stabilize training by correcting for client drift. MOON [36] utilizes contrastive learning to align local models with a global model and each other. FedNTD [37] preserves global knowledge during local training via not-true distillation on non-ground-truth classes. In addition, cross-silo ranking FL (CS-F-LTR) [38] improves model stability and efficiency in heterogeneous environments, and Fed-LTD [39] introduces a federated learning framework, enabling collaborative decision-making. While Hu-Fu [40] explores secure, efficient computation under strict privacy constraints.

### 2.3 Personalized Federated Learning

Personalized Federated Learning (PFL) tailors the global model to each client's specific data distribution while still benefiting from collaboration. Representative approaches include regularization-based FedRoD [41] and representation-based FedRep [42]. Attention message passing is exploited in FedAMP [43]. Layer-wise personalization is realized by FedBN [44] and by retraining only heads in FedBABU [45].

Beyond these, FedPer [46] separates a shared base network and local personalization layers, whereas FedCross [47] employs cross-model aggregation to boost accuracy. FedPAC [48] aligns local and global features and coordinates client-specific classifiers, and FedPHP [49] leverages hashed private models for efficiency. FedGH [50] introduces a generalized global header to unify heterogeneous backbones. FedCP [51] separates global and personalized information at the feature level via conditional policies. FedARC [52] incorporates adaptive regularization to align local objectives with global loss and model-contrastive learning to enhance generalization across heterogeneous data sources.

Although these methods address either non-IID optimization or model personalization, they lack a unified mechanism for conditional attention and comprehensive evaluation. CAFL bridges this gap by introducing conditional attention with a gated feature split, accompanied by a refined evaluation paradigm for image emotion analysis.

## 3 Methodology

### 3.1 Problem Statement

Personalized Federated Learning (PFL) addresses a fundamental challenge in distributed machine learning: how to leverage the rich, diverse data held by multiple clients without violating data privacy, particularly when the data is non-independently and identically distributed (non-IID). This section formally defines the PFL problem and introduces the notation used throughout this paper (summarized in Table 1).

We consider a federated system of $m$ clients, $C_1, \ldots, C_m$. Each client $C_i$ holds a private dataset $\mathcal{D}_i$ drawn from a local data distribution $P_i$. A key characteristic of real-world FL is data heterogeneity, where for any two clients $i \neq j$, their distributions differ, i.e., $P_i \neq P_j$. In the domain of image emotion analysis, this non-IID condition is not an exception but the norm, arising from differing cultural contexts, individual subjectivity, and environmental factors that lead to significant *concept drift* ($P(y|x)$ shift) across clients.

A single global model trained on such heterogeneous data often fails to perform well for any individual client. PFL addresses this by aiming to train a unique, personalized model $M_i$ for each client, defined by a distinct set of parameters $\theta_i$. The goal is to optimize performance of every local model on its own data distribution, while still benefiting from the collaborative knowledge of the entire network. Our framework operates under the common PFL assumption that all clients share an identical model architecture, though the parameters of these models are personalized during training.

Formally, we define a local objective function $F_i(\theta_i)$ for each client, which typically measures the empirical risk on its local dataset $\mathcal{D}_i$. The global objective of the PFL system is to find the optimal set of all client parameters, $\Theta = [\theta_1, \ldots, \theta_m]$, by minimizing a global function $G(\Theta)$

**Table 1** Summary of notations used in the CAFL framework

| Symbol | Description |
|---|---|
| ***System-Level Parameters*** | |
| $m$ | Total number of clients in the federated system. |
| $C_i$ | The $i$-th client, where $i \in \{1, \ldots, m\}$. |
| $\mathcal{D}_i$ | The private local dataset of client $C_i$. |
| $P_i$ | The unique data distribution from which $\mathcal{D}_i$ is sampled. |
| $t$, $T$ | The current communication round index and the total number of rounds. |
| ***Model Architecture Components*** | |
| $\theta_i$ | The complete set of learnable parameters for the model of client $C_i$. |
| $\theta_{i,\text{base}}$ | Parameters of client $i$'s personalized base model. |
| $\theta_{i,\text{gate}}$ | Parameters of client $i$'s private adaptive gate network (Module A). |
| $\theta_{i,\text{head}}$ | Parameters of client $i$'s private personalized prediction head (Module B). |
| $\theta^t_{\text{glob}}$ | Parameters of the server-aggregated global head at round $t$. |
| $\theta^t_{i,\text{glob}}$ | Client $i$'s local copy of the global head, updated during local training. |
| $f_{\text{base}}(\cdot)$ | The function of the base model architecture. |
| $\text{Gate}_i(\cdot)$ | The function of client $i$'s gate network. |
| $h_{i,\text{head}}(\cdot)$ | The function of client $i$'s personalized head. |
| $h_{\text{glob}}(\cdot)$ | The function of the global head architecture. |
| ***Collaborative Mechanism (Server-Side)*** | |
| $K(\cdot)$ | A kernel function measuring squared distance between model parameters. |
| $\sigma$ | A sensitivity hyperparameter for the kernel function. |
| $\alpha_{ij}$ | The attention weight, representing the influence of client $j$ on client $i$. |
| $u^t_i$ | The personalized reference model for client $C_i$ at round $t$. |
| ***Optimization and Objectives*** | |
| $F_i(\theta_i)$ | The local objective function for client $C_i$. |
| $\mathcal{L}_{\text{class}}$ | The supervised classification loss (e.g., cross-entropy). |
| $\mathcal{L}_{\text{collab}}$ | The collaborative regularization loss (Module C). |
| $\lambda$ | Hyperparameter controlling the strength of collaborative regularization. |
| $G(\Theta)$ | The global PFL objective function over all client parameters $\Theta$. |
| $R(\Theta)$ | The regularization term for inter-client collaboration. |
| $\Theta$ | The collection of all client parameters, $\Theta = [\theta_1, \ldots, \theta_m]$. |
| $\eta^t$ | The learning rate (step size) at round $t$. |
| $B$ | Bounding constant for gradients and subgradients (from Assumption 1). |
| $\Theta^*$ | An optimal solution to the global objective $G(\Theta)$. |
| $G^*$ | The optimal value $G(\Theta^*)$. |
| $L$ | Lipschitz constant for gradients in non-convex analysis. |
| $U^t$ | The collection of all personalized reference models at round $t$, $U^t = [u^t_1, \ldots, u^t_m]$. |
| $Y$ | An element of the subdifferential $\partial F(\Theta^t)$. |

that aggregates local objectives and incorporates a collaboration mechanism:

$$\min_{\Theta} \left\{ G(\Theta) := \sum_{i=1}^{m} F_i(\theta_i) + \lambda R(\Theta) \right\}, \tag{1}$$

where $\lambda > 0$ is a hyperparameter that balances local performance with collaboration, and $R(\Theta)$ is a regularization term that defines the nature of inter-client collaboration. The design of $F_i$ and $R(\Theta)$ is the core of any PFL method. In the subsequent sections, we will detail how CAFL instantiates these terms through its modular architecture.

## 3.2 Conditional Attention Federated Learning

Building upon the PFL formulation (Eq. (1)), we now introduce the architectural and procedural details of our proposed method, Conditional Attention Federated Learning (CAFL). CAFL instantiates a multi-level personalization strategy, realized through a sequence of three distinct modules that

address data heterogeneity at the feature, model, and communication levels.

#### 3.2.1 Overview of CAFL's Three-Module Architecture

To effectively address the concept drift challenge, CAFL is constructed from three key modules, whose interplay is illustrated in Figure 1.

- **Module A: Adaptive Feature Gating.** At the feature level, this module employs a learnable, sample-conditioned policy to adaptively separate the feature representation from a common backbone architecture into a personalized stream and a global stream. It learns which features are client-specific versus which are universally applicable.
- **Module B: Bipartite Prediction Head.** At the model level, CAFL utilizes a bipartite (two-part) head architecture. It consists of a private personalized head to process the personalized feature stream and a global head (synchronized with the server) to process the global stream. This structure enables tailored predictions for each client while still leveraging shared knowledge.
- **Module C: Collaborative Similarity Matching.** At the communication level, this module facilitates intelligent collaboration. The server computes inter-client model similarities to provide each client with a personalized reference model. This reference model is then used as a regularization target during local training to encourage effective knowledge transfer.

This modular design provides a clear and robust framework for balancing personalization and collaboration, directly responding to the challenges of learning from non-IID data.

#### 3.2.2 Module A: Adaptive Feature Gating

The first stage of personalization in CAFL occurs at the feature level. After client $C_i$ extracts a feature representation $r \in \mathbb{R}^K$ from its personalized base model, $r = f_{\text{base}}(x; \theta_{i,\text{base}})$, we introduce a lightweight gate network, $\text{Gate}_i(\cdot)$, parameterized by the client's private parameters $\theta_{i,\text{gate}}$. This module's purpose is to learn a sample-specific policy for routing features, which is a key mechanism for handling concept drift by determining which aspects of an image are subject to personal interpretation versus universal understanding.

Specifically, the gate network is a conditional policy network that takes the feature representation $r$ as input and outputs two complementary weight vectors (or masks), $w_p$ and $w_g$, both in $\mathbb{R}^K$:

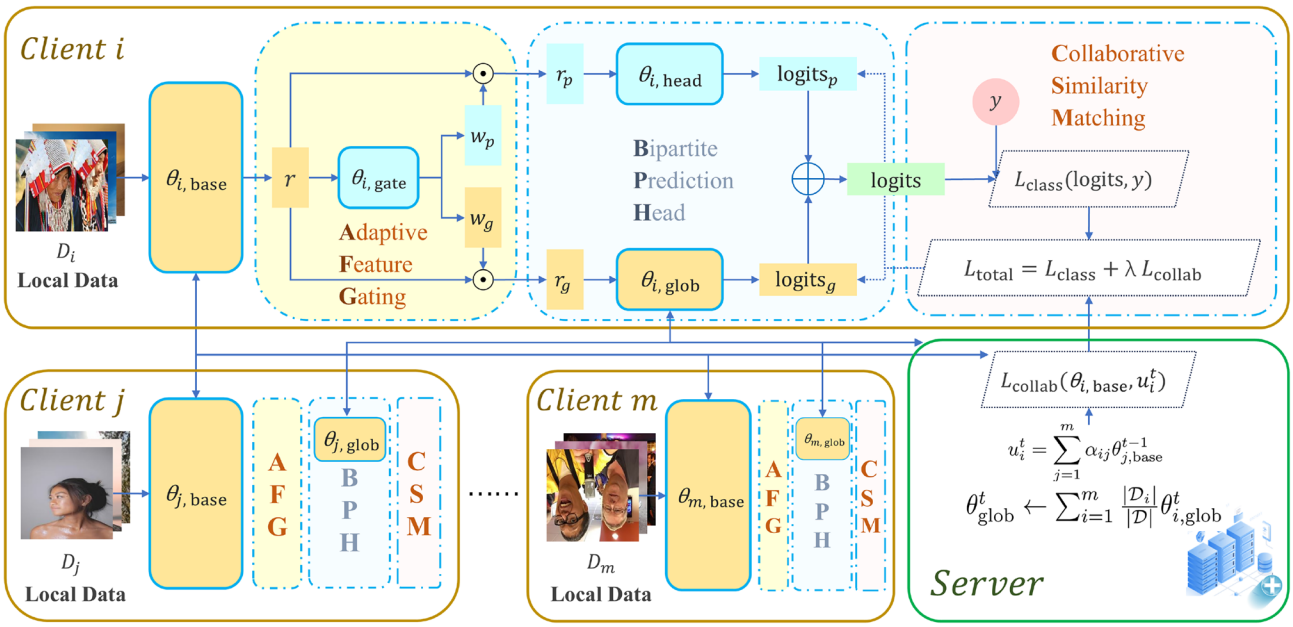



**Fig. 1** The architecture and end-to-end workflow of CAFL, which operates in iterative rounds between clients and a server. **(1) Local Client Training:** Within each round, a client (e.g., Client i) processes its local data. A feature vector $r$ is extracted from the personalized base model ($\theta_{i,\text{base}}$). **(2) Module A (Adaptive Feature Gating)** splits $r$ into personalized ($r_p$) and global ($r_g$) streams. **(3) Module B (Bipartite Prediction Head)** uses a private head ($h_{i,\text{head}}$) and its local copy of the global head ($\theta_{i,\text{glob}}$) to produce the final `logits`. The client model is then updated using a loss function that includes **(4) Module C (Collaborative Similarity Matching)**, which incorporates the personalized reference model $u_i^t$ provided by the server. **(5) Server-Side Coordination:** After local training, clients upload their updated parameters ($\theta_{i,\text{base}}$ and $\theta_{i,\text{glob}}$) to the server. The server computes a new personalized reference model $u_i^{t+1}$ for each client and aggregates a new global head $\theta_{\text{glob}}^{t+1}$ for the next round. This entire process preserves privacy by only exchanging model parameters

$$(w_p, w_g) = \mathrm{Gate}_i(r; \theta_{i,\mathrm{gate}}). \tag{2}$$

To ensure reproducibility, we implement this gate as a two-layer MLP. It first projects the input feature $r$ to a hidden dimension, applies a ReLU activation, and then projects it to a dimension of $2K$. This output is reshaped to size $(K, 2)$, and a softmax function is applied along the second dimension. This ensures that for each feature element $k \in \{1, \ldots, K\}$, the weights are non-negative and sum to one: $w_p^{(k)} + w_g^{(k)} = 1$.

These weights are then applied to the original feature vector via element-wise multiplication ($\odot$) to separate it into a personalized component $r_p$ and a global component $r_g$:

$$r_p = r \odot w_p \tag{3}$$

$$r_g = r \odot w_g. \tag{4}$$

Through this mechanism, the gate network learns on a sample-by-sample basis which feature dimensions are client-specific (and should be routed to the personalized head) versus those that are generalizable (and should be routed to the global head).

#### 3.2.3 Module B: Bipartite Prediction Head

Following the feature gating, the two resulting feature streams are processed by a bipartite head architecture. The rationale behind this design is to explicitly disentangle client-specific knowledge from universal knowledge. The final prediction (logits) for client $C_i$ is computed by summing the outputs of its private personalized head, $h_{i,\mathrm{head}}(\cdot)$, and its local copy of the shared global head, $h_{\mathrm{glob}}(\cdot)$:

$$\mathrm{logits} = h_{i,\mathrm{head}}(r_p; \theta_{i,\mathrm{head}}) + h_{\mathrm{glob}}(r_g; \theta^t_{i,\mathrm{glob}}). \tag{5}$$

The parameters of the personalized head, $\theta_{i,\mathrm{head}}$, are unique to client $C_i$ and are never shared, allowing them to capture client-specific classification boundaries—a direct countermeasure to concept drift. In contrast, the parameters of the global head, $\theta^t_{\mathrm{glob}}$, are aggregated on the server at each round $t$ and distributed to clients, thereby capturing knowledge common to the entire federation. Both heads are typically single linear layers mapping from $\mathbb{R}^K$ to the output space. This dual-head structure provides a clear mechanism for balancing personalization with global collaboration at the model parameter level.

#### 3.2.4 Local Training Objective

With the components of Modules A and B defined, we can now formulate the full local objective function $F_i$ for client $C_i$. The objective consists of two parts: a standard supervised classification loss, $\mathcal{L}_{\mathrm{class}}$, and a collaborative regularization loss, $\mathcal{L}_{\mathrm{collab}}$, which is guided by Module C.

$$\min_{\theta_i} F_i(\theta_i) = \min_{\theta_i} \left( \mathcal{L}_{\mathrm{class}}(\theta_i) + \lambda \mathcal{L}_{\mathrm{collab}}(\theta_{i,\mathrm{base}}; u_i^t) \right), \tag{6}$$

where $\theta_i = \{\theta_{i,\mathrm{base}}, \theta_{i,\mathrm{gate}}, \theta_{i,\mathrm{head}}, \theta_{i,\mathrm{glob}}\}$ represents all parameters updated by the client during local training. The classification loss is typically the cross-entropy between the predicted logits and the true labels $y$:

$$\mathcal{L}_{\mathrm{class}} = \mathbb{E}_{(x,y)\in\mathcal{D}_i} \left[ \mathrm{CE}(\mathrm{logits}, y) \right]. \tag{7}$$

The collaborative loss term, whose target $u_i^t$ is provided by the server, will be detailed in the next section.

#### 3.2.5 Module C: Collaborative Similarity Matching

The final component of CAFL is the server-facilitated collaborative mechanism. We use the squared L2 distance between model parameters as a proxy for functional similarity, as it is both effective and privacy-preserving. This module encourages clients with similar data distributions (as reflected by their model parameters) to learn more from each other, directly mitigating the negative effects of concept drift from dissimilar clients.

The process is executed on the server in each communication round $t$. First, the server computes the pairwise similarity between all clients based on their base model parameters from the previous round, $\{\theta^{t-1}_{j,\mathrm{base}}\}_{j=1}^m$. The similarity score $s_{ij}$ between client $i$ and client $j$ is calculated using a kernel function $K(\cdot)$, such as the Gaussian kernel:

$$s_{ij} = K\left( \|\theta^{t-1}_{i,\mathrm{base}} - \theta^{t-1}_{j,\mathrm{base}}\|^2 \right) = \exp\left( -\frac{\|\theta^{t-1}_{i,\mathrm{base}} - \theta^{t-1}_{j,\mathrm{base}}\|^2}{\sigma^2} \right), \tag{8}$$

where $\sigma$ is a sensitivity hyperparameter. These raw scores are then normalized for each client $i$ using the softmax function to produce the final attention weights $\alpha_{ij}$:

$$\alpha_{ij} = \frac{\exp(s_{ij})}{\sum_{k=1}^m \exp(s_{ik})}. \tag{9}$$

This ensures the weights are non-negative and sum to one ($\sum_{j=1}^m \alpha_{ij} = 1$), forming a valid probability distribution of influence from peer clients.

Finally, the server constructs a personalized reference model $u_i^t$ for each client $C_i$. This reference model is a weighted average of all clients' base models, using the computed attention weights:

$$u_i^t = \sum_{j=1}^{m} \alpha_{ij} \theta_{j,\text{base}}^{t-1}. \tag{10}$$

This reference model $u_i^t$ is sent to client $C_i$ and serves as the collaborative target for the regularization loss $\mathcal{L}_{\text{collab}}$ in its local training objective (Eq. (6)). The crucial hyperparameters, $\lambda$ and $\sigma$, are determined empirically on a validation set, as detailed in our experiments.

#### 3.2.6 Privacy Preservation

CAFL ensures client privacy under the standard honest-but-curious server model. Privacy is maintained through several key design choices: (1) No raw data $\mathcal{D}_i$ ever leaves the client's local device. (2) The server only has access to model parameters designated for sharing (specifically, the set of personalized base models $\{\theta_{i,\text{base}}^t\}_{i=1}^m$ and the set of updated global heads $\{\theta_{i,\text{glob}}^t\}_{i=1}^m$), not the private components ($\theta_{i,\text{gate}}$ and $\theta_{i,\text{head}}$). (3) The reference model $u_i^t$ sent to a client is a weighted aggregate of all other models, preventing the direct reconstruction of any single client's model from this message.

#### 3.2.7 Algorithm Summary

Our framework is presented assuming full client participation in each round for clarity, though it can be extended to handle partial participation. Algorithm 1 summarizes the complete CAFL process, detailing the interplay between the server and clients and the roles of the three modules.

**Input:** Number of clients $m$, total rounds $T$, learning rate $\eta^t$, hyperparameters $\lambda, \sigma$
**Output:** Final personalized model parameters $\{\theta_i^T\}_{i=1}^m$
/* Server Initialization */
Initialize global head $\theta_{\text{glob}}^0$ and base models $\{\theta_{i,\text{base}}^0\}_{i=1}^m$.
**for** $t = 1, \ldots, T$ **do**
  /* Server computes personalized references and broadcasts */
  **for** $i = 1, \ldots, m$ **do**
    $\alpha_{ij} \leftarrow \text{softmax}_j(\exp(-\|\theta_{i,\text{base}}^{t-1} - \theta_{j,\text{base}}^{t-1}\|^2/\sigma^2))$
    $u_i^t \leftarrow \sum_{j=1}^m \alpha_{ij} \theta_{j,\text{base}}^{t-1}$
  /* Clients update in parallel */
  **for** *client* $C_i$ *in parallel* **do**
    $\theta_{i,\text{base}}^t, \theta_{i,\text{glob}}^t \leftarrow \textbf{ClientUpdate}(i, u_i^t, \theta_{\text{glob}}^{t-1}, \eta^t)$
  /* Server aggregates the global head */
  $\theta_{\text{glob}}^t \leftarrow \sum_{i=1}^m \frac{|\mathcal{D}_i|}{|\mathcal{D}|} \theta_{i,\text{glob}}^t$ // Federated Averaging

**Function** ***ClientUpdate***$(i, u_i^t, \theta_{glob}^{t-1}, \eta^t)$
  $\theta_{i,\text{glob}} \leftarrow \theta_{\text{glob}}^{t-1}$ // Load global head
  **for** *each local epoch* **do**
    **for** *each batch* $(x, y) \in \mathcal{D}_i$ **do**
      $r \leftarrow f_{\text{base}}(x; \theta_{i,\text{base}})$
      $(r_p, r_g) \leftarrow \text{Gate}_i(r; \theta_{i,\text{gate}})$ // Module A
      $\text{logits} \leftarrow h_{i,\text{head}}(r_p; \theta_{i,\text{head}}) + h_{\text{glob}}(r_g; \theta_{i,\text{glob}})$ // Module B
      $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{class}}(\text{logits}, y) + \lambda \cdot \frac{1}{2}\|\theta_{i,\text{base}} - u_i^t\|^2$ // Module C
      Update $\{\theta_{i,\text{base}}, \ldots, \theta_{i,\text{glob}}\}$ using $\nabla \mathcal{L}_{\text{total}}$ with learning rate $\eta^t$
  **return** updated $\theta_{i,\text{base}}$ and $\theta_{i,\text{glob}}$ to server

**Algorithm 1** Conditional Attention Federated Learning (CAFL)

### 3.3 Convergence Analysis of CAFL

In this section provides a theoretical analysis of the convergence properties of CAFL. We first establish some assumptions and then proceed to prove the convergence for both convex and non-convex scenarios.

#### 3.3.1 Preliminaries and Assumptions

Before delving into the convergence analysis, we make the following assumption:

**Assumption 1** There exists a constant $B > 0$ such that $\max\{\|Y\| : Y \in \partial F(\Theta^t)\} \leq B$ and $\|\nabla R(\Theta^t)\| \leq B/\lambda$ hold for every $t \geq 0$, where $\partial F$ is the subdifferential of $F$ and $\|\cdot\|$ is the Frobenius norm.

This assumption is common in the analysis of many incremental and stochastic optimization algorithms. It naturally holds if both $F(\Theta)$ and $R(\Theta)$ are locally Lipschitz continuous and $\|\Theta^t\|$ is bounded by a constant for all $t \geq 0$. The assumption is realistic for CAFL's modules, as the gate network (Module A) and prediction heads (Module B) are typically bounded in deep learning settings, and the similarity function (Module C) is smooth (e.g., Gaussian kernel).

#### 3.3.2 Convergence for Convex Case

We first establish the convergence guarantee for CAFL when both $F(\Theta)$ and $R(\Theta)$ are convex functions.

**Theorem 1** *Under Assumption 1 and assuming functions $F(\Theta)$ and $R(\Theta)$ in Eq. (1) are convex, if $\eta^1 = \cdots = \eta^T = \lambda/\sqrt{T}$ for some $T \geq 0$, then the sequence $\Theta^0, \ldots, \Theta^T$ generated by Algorithm 1 satisfies*

$$\min_{0\leq t\leq T} G(\Theta^t) \leq G^* + \frac{\|\Theta^0 - \Theta^*\|^2 + 5B^2}{\sqrt{T}}, \tag{11}$$

*where $\Theta^*$ is an optimal solution of Eq. (1) and $G^* = G(\Theta^*)$. Moreover, if $\eta^t$ satisfies $\sum_{t=1}^{\infty} \eta^t = \infty$ and $\sum_{t=1}^{\infty} (\eta^t)^2 < \infty$, then*

$$\lim_{t\to\infty} G(\Theta^t) = G^*. \tag{12}$$

***Proof*** We begin by establishing a key inequality. For any $\Theta$, including $\Theta^*$, we have:

$$\begin{aligned}\|\Theta - U^t\|^2 &= \|\Theta - \Theta^{t-1} + \Theta^{t-1} - U^t\|^2 \\ &= \|\Theta - \Theta^{t-1}\|^2 + \|\Theta^{t-1} - U^t\|^2 \\ &\quad + 2\langle\Theta - \Theta^{t-1}, \Theta^{t-1} - U^t\rangle \\ &= \|\Theta - \Theta^{t-1}\|^2 + (\eta^t)^2\|\nabla R(\Theta^{t-1})\|^2 \\ &\quad + 2\eta^t\langle\Theta - \Theta^{t-1}, \nabla R(\Theta^{t-1})\rangle\end{aligned} \tag{13}$$

From the update rule for $\Theta^t$ and its optimality condition, we can derive:

$$0 \geq \eta^t\langle Y, \Theta - \Theta^t\rangle + \lambda\langle\Theta^t - U^t, \Theta - \Theta^t\rangle \tag{14}$$

where $Y \in \partial F(\Theta^t)$. Using the convexity of $F$ and manipulating the terms, we obtain:

$$\eta^t(F(\Theta) - F(\Theta^t)) \leq \frac{\lambda}{2}(\|\Theta - U^t\|^2 - \|\Theta - \Theta^t\|^2) \tag{15}$$

Combining this with the key inequality and using the convexity of $R$, we can establish:

$$\begin{aligned}\eta^t(G(\Theta^t) - G(\Theta)) \leq \frac{\lambda}{4}\|\Theta - \Theta^{t-1}\|^2 - \frac{\lambda}{2}\|\Theta - \Theta^t\|^2 \\ + \frac{\lambda(\eta^t)^2}{2}\|\nabla R(\Theta^{t-1})\|^2 + \frac{(\eta^t)^2 B^2}{\lambda}\end{aligned} \tag{16}$$

Summing over $t = 1, \ldots, T$ and using the fact that $\eta^t = \lambda/\sqrt{T}$ for all $t$, we arrive at:

$$\sum_{t=1}^{T} \eta^t(G(\Theta^t) - G(\Theta^*)) \leq \frac{\lambda}{4}\|\Theta^* - \Theta^0\|^2 + \frac{5B^2}{2\sqrt{T}} \tag{17}$$

From this, we can conclude:

$$\min_{0\leq t\leq T} G(\Theta^t) - G^* \leq \frac{\|\Theta^0 - \Theta^*\|^2 + 5B^2}{\sqrt{T}} \tag{18}$$

The second part of the theorem, concerning the convergence when $\sum_{t=1}^{\infty} \eta^t = \infty$ and $\sum_{t=1}^{\infty} (\eta^t)^2 < \infty$, follows from standard results in convex optimization theory. These conditions on the step sizes ensure that the algorithm takes sufficiently large steps to reach the optimum, while the steps become small enough over time to allow convergence. □

This theorem implies that for any $\epsilon > 0$, CAFL needs at most $O(\epsilon^{-2})$ iterations to find an $\epsilon$-optimal solution $\tilde{\Theta}$ of Eq. (1) such that $G(\tilde{\Theta}) - G^* \leq \epsilon$. It also establishes the global convergence of CAFL to an optimal solution of Eq. (1) when $G$ is convex.

#### 3.3.3 Convergence for Non-Convex Case

Next, we provide the convergence guarantee of CAFL when $G(\Theta)$ is a smooth and non-convex function.

**Theorem 2** *Under Assumption 1 and assuming functions $F(\Theta)$ and $R(\Theta)$ in Eq. (1) are continuously differentiable and the gradients $\nabla F(\Theta)$ and $\nabla R(\Theta)$ are Lipschitz continuous with modulus L, if $\eta^1 = \cdots = \eta^T = \lambda/\sqrt{T}$, then the sequence $\Theta^0, \ldots, \Theta^T$ generated by Algorithm 1 satisfies*

$$\min_{0\le t\le T} \|\nabla G(\Theta^t)\|^2 \le \frac{18(G(\Theta^0) - G^* + 20LB^2)}{\sqrt{T}} + O\left(\frac{1}{T}\right), \tag{19}$$

*where $\Theta^*$ and $G^*$ are the same as in Theorem 1. Moreover, if $\eta^t$ satisfies $\sum_{t=1}^{\infty} \eta^t = \infty$ and $\sum_{t=1}^{\infty} (\eta^t)^2 < \infty$, then*

$$\lim_{t\to\infty} \|\nabla G(\Theta^t)\| = 0. \tag{20}$$

The proof of this theorem follows a similar structure to that of Theorem 1, but requires additional techniques to handle the non-convexity of the objective function. The key idea is to use the Lipschitz continuity of the gradients to bound the change in the objective function between iterations.

Theorem 2 implies that for any $\epsilon > 0$, CAFL needs at most $O(\epsilon^{-4})$ iterations to find an $\epsilon$-approximate stationary point $\tilde{\Theta}$ of Eq. (1) such that $\|\nabla G(\tilde{\Theta})\| \le \epsilon$. It also establishes the global convergence of CAFL to a stationary point of Eq. (1) when $G$ is smooth and non-convex.

These theoretical results demonstrate the strong convergence properties of CAFL in both convex and non-convex scenarios, providing a solid foundation for its application in personalized federated learning.

#### 3.3.4 Comparison of Convergence Properties

To contextualize CAFL's convergence, we compare it with related federated learning (FL) methods and non-federated centralized SGD, focusing on non-IID data scenarios common in image sentiment tasks.

For convex objectives, CAFL achieves an $O(1/\sqrt{T})$ rate to an $\epsilon$-optimal solution, matching FedAvg [17] and centralized SGD [53]. However, FedAvg often diverges or slows on non-IID data due to client drift [18], requiring more rounds (e.g., 2.8× slowdown on MNIST non-IID vs. IID [17]). CAFL's attentive mechanism ensures stability by promoting similar-client collaboration.

In non-convex settings (e.g., deep sentiment models), CAFL converges at $O(1/\sqrt{T})$ for $\|\nabla G\|^2 \le \epsilon$, comparable to Per-FedAvg [54], which uses a meta-learning approach for robustness on non-IID data. FedProx [18] improves FedAvg with a proximal term, achieving $O(\Delta/(\rho\epsilon))$ iterations (where $\rho$ accounts for heterogeneity), boosting accuracy by up to 22% on non-IID tasks. Centralized SGD shares $O(1/\sqrt{T})$ rates but assumes data centralization, ignoring privacy; CAFL approaches this efficiency in FL while handling non-IID better empirically (see Section 4 for faster convergence curves vs. baselines).

Overall, CAFL matches state-of-the-art FL rates while excelling in non-IID stability, making it suitable for heterogeneous sentiment data.

## 4 Experiments

To evaluate the effectiveness of our proposed Conditional Attention Federated Learning (CAFL) method, we conducted a series of comprehensive experiments. This section details our experimental setup, including datasets, implementation details, and evaluation metrics.

### 4.1 Datasets

We utilized three well-established datasets in the domain of image emotion analysis: ARTphoto [5], ABSTRACT [5], and PARA [55]. These datasets were chosen to construct a comprehensive concept drift scenario, where different clients may associate varying emotions with the same image, reflecting the real-world challenge of divergent $P(y|x)$ distributions across different centers.

#### 4.1.1 Dataset Descriptions

- **ARTphoto**: This dataset consists of artistic photographs professionally captured to evoke specific emotions, providing a rich source of visually expressive content.
- **ABSTRACT**: Comprised of abstract paintings and patterns, this dataset challenges emotion recognition algorithms with non-representational imagery.
- **PARA**: This dataset focuses on personalized image aesthetics assessmet, incorporating rich attributes that align well with our goal of personalized emotion classification.

For all datasets, we adopted an 80-20 split, utilizing 80% of the data for training and the remaining 20% for testing. This configuration ensures a robust evaluation of our Model's performance and generalization capabilities.

### 4.2 Experimental Setup

Our experiments were designed for a comprehensive and rigorous evaluation of CAFL against state-of-the-art PFL methods. This section details the classification tasks, the

rationale for our architectural and pre-training strategies, the federated learning configuration, and all implementation details required for reproducibility.

#### 4.2.1 Classification Tasks

We assess all methods under two distinct classification scenarios to test their capabilities at different levels of granularity:

1. **Binary Classification**: Images were classified into two broad categories: positive and negative emotions. This setup provides a high-level evaluation of sentiment analysis capabilities.
2. **Multi-class Classification**: We employed a more challenging 8-class emotion task, distinguishing between Amusement, Anger, Awe, Contentment, Disgust, Excitement, Fear, and Sadness. This fine-grained classification tests the model's ability to discern subtle emotional differences under significant concept drift.

#### 4.2.2 Model Architectures and Pre-training Strategy

The choice of model architectures and training strategies is crucial for a robust evaluation of federated algorithms.

**Backbone Model Selection**. To evaluate the generalizability of CAFL, we selected two state-of-the-art lightweight models that represent distinct and competing architectural philosophies for mobile vision. Our selection criteria prioritized models suitable for resource-constrained edge devices while ensuring significant architectural diversity.

- **MobileNet V3-Small** [56]: A highly optimized pure convolutional neural network (CNN), representing the classic paradigm for mobile efficiency. The specific variant we use has approximately 2.5M parameters and achieves 67.5% Top-1 accuracy on ImageNet-1K.
- **MobileViT V2-0.50** [26]: A modern hybrid architecture combining CNNs with Vision Transformers. This model introduces separable self-attention to achieve high efficiency. The 0.5x scaled version we employ is even more compact, with only 1.3M parameters, while reaching a competitive 69.2% Top-1 accuracy on ImageNet-1K.

By selecting these two models, we can rigorously test CAFL's performance on both a classic, highly efficient CNN and a modern, even more parameter-efficient hybrid model. This demonstrates that our method's advantages are robust and not contingent on a specific architectural choice.

**Pre-training Strategy**. To thoroughly investigate the algorithms' learning capabilities, we conducted all experiments in two scenarios:

1. **Training from Scratch**: Models are trained without pre-trained weights. This setup rigorously tests an algorithm's ability to learn effective representations purely from the decentralized and non-IID data within the federation.
2. **Fine-tuning from Pre-trained Models**: Models are initialized with weights pre-trained on ImageNet. This scenario assesses an algorithm's effectiveness in adapting general, large-scale knowledge to the specific, personalized task of emotion analysis while navigating the challenges of concept drift.

#### 4.2.3 Federated Learning Configuration

- **Client Simulation**: To construct a realistic PFL environment with significant concept drift, we curated a cohort of 24 unique clients. These clients were formed by combining user data from the ARTphoto [5], ABSTRACT [5], and PARA [55] datasets, ensuring a diverse and heterogeneous setup.
- **Client Participation**: To simulate real-world conditions where not all clients are available, we assumed 50% client participation in each communication round. A random subset of 12 clients was selected to perform local training in each round.

#### 4.2.4 Implementation Details and Hyperparameters

All experiments were conducted on a simulated distributed network using NVIDIA RTX 4090 GPUs. For all baseline methods, we adhered to their official papers' implementations and tuned their respective hyperparameters on a validation set to ensure a fair and robust comparison.

- **Optimizer**: We used the SGD optimizer with a batch size of 16 for all local training procedures.
- **Training Schedule**: All experiments were run for 200 communication rounds. In each round, participating clients performed 1 epoch of local training. The initial learning rate was set to 1e-4, with a decay factor of 0.99 applied after each communication round.
- **CAFL-Specific Hyperparameters**: Our proposed method introduces two key hyperparameters that govern its personalization and collaboration behavior: the regularization strength $\lambda$ and the kernel sensitivity $\sigma$. The values for these were set based on their distinct functional roles within our framework and were kept constant across all experiments to ensure a fair comparison.
  - **Regularization Strength ($\lambda$)**: This parameter balances the primary classification loss ($\mathcal{L}_{\text{class}}$) against

**Table 2** Performance comparison on the 8-class classification task **without pre-trained models**. CAFL demonstrates a significant lead in both AUC and Accuracy across two distinct lightweight architectures. Mean and standard deviation are reported over three runs

| Models | MobileNet V3 [56] | | MobileViT V2 [26] | |
|---|---|---|---|---|
| Methods | AUC | ACC | AUC | ACC |
| FedAvg [17] | 69.10 ± 0.42 | 38.46 ± 0.17 | 61.43 ± 0.50 | 28.72 ± 7.07 |
| FedBABU [45] | 67.75 ± 0.52 | 32.04 ± 1.54 | 59.50 ± 1.75 | 27.93 ± 7.58 |
| FedBN [44] | 69.10 ± 0.42 | 38.46 ± 0.17 | 61.47 ± 0.53 | 29.04 ± 6.92 |
| FedCross [47] | 73.39 ± 0.49 | 38.64 ± 0.00 | 65.66 ± 0.55 | 33.93 ± 0.79 |
| FedGH [50] | 73.32 ± 1.04 | 31.39 ± 2.04 | 61.95 ± 0.36 | 35.36 ± 2.37 |
| FedNTD [37] | 72.64 ± 0.46 | 38.64 ± 0.00 | 65.31 ± 0.52 | 33.38 ± 1.20 |
| FedPAC [48] | 72.36 ± 0.16 | 38.64 ± 0.00 | 61.32 ± 0.16 | 24.38 ± 0.41 |
| FedPHP [49] | 68.66 ± 0.35 | 38.64 ± 0.00 | 60.23 ± 1.01 | 22.95 ± 1.77 |
| FedProx [18] | 66.85 ± 0.63 | 36.93 ± 2.32 | 58.99 ± 2.03 | 21.88 ± 13.31 |
| MOON [36] | 69.12 ± 0.42 | 38.64 ± 0.00 | 61.58 ± 0.37 | 24.61 ± 0.91 |
| Per-FedAvg [54] | 67.18 ± 0.36 | 37.90 ± 0.95 | 57.98 ± 2.72 | 18.79 ± 10.76 |
| **CAFL (ours)** | **76.60 ± 0.32** | **46.03 ± 0.24** | **66.30 ± 1.16** | **35.44 ± 1.36** |

**Table 3** Performance comparison on the 8-class classification task **with ImageNet pre-trained models**. Fine-tuning further improves performance across the board, with CAFL maintaining its significant lead and demonstrating its superior ability to adapt general knowledge to personalized tasks

| Models | MobileNet V3 [56] | | MobileViT V2 [26] | |
|---|---|---|---|---|
| Methods | AUC | ACC | AUC | ACC |
| FedAvg [17] | 72.03 ± 0.48 | 36.52 ± 0.35 | 66.82 ± 0.48 | 31.95 ± 1.56 |
| FedBABU [45] | 66.07 ± 2.12 | 29.87 ± 1.89 | 57.08 ± 1.94 | 18.61 ± 1.02 |
| FedBN [44] | 72.03 ± 0.48 | 36.52 ± 0.35 | 66.83 ± 0.48 | 32.04 ± 1.52 |
| FedCross [47] | 74.18 ± 0.89 | 40.95 ± 0.17 | 67.99 ± 0.50 | 34.86 ± 1.36 |
| FedGH [50] | 71.87 ± 0.51 | 35.41 ± 0.46 | 67.43 ± 1.73 | 28.12 ± 1.33 |
| FedNTD [37] | 73.63 ± 0.83 | 40.54 ± 0.17 | 67.85 ± 0.50 | 34.86 ± 1.42 |
| FedPAC [48] | 71.48 ± 1.06 | 40.40 ± 0.13 | 64.24 ± 0.87 | 29.09 ± 1.97 |
| FedPHP [49] | 70.28 ± 0.82 | 34.63 ± 0.63 | 62.12 ± 1.36 | 25.25 ± 1.33 |
| FedProx [18] | 59.94 ± 1.78 | 21.51 ± 1.37 | 53.96 ± 2.25 | 15.60 ± 1.25 |
| MOON [36] | 72.03 ± 0.48 | 36.52 ± 0.35 | 66.83 ± 0.49 | 31.99 ± 1.48 |
| Per-FedAvg [54] | 62.82 ± 1.68 | 25.72 ± 1.92 | 55.29 ± 2.19 | 16.30 ± 0.91 |
| **CAFL (ours)** | **81.30 ± 0.75** | **51.48 ± 0.62** | **74.78 ± 0.75** | **44.28 ± 0.80** |

the collaborative regularization loss ($\mathcal{L}_{\text{collab}}$). In a PFL setting characterized by significant concept drift, it is crucial for each model to prioritize fitting its own local data distribution. Therefore, we set a modest value of $\lambda = 0.1$ to ensure the collaborative term acts as a targeted regularizer, guiding learning without overpowering the essential local personalization.

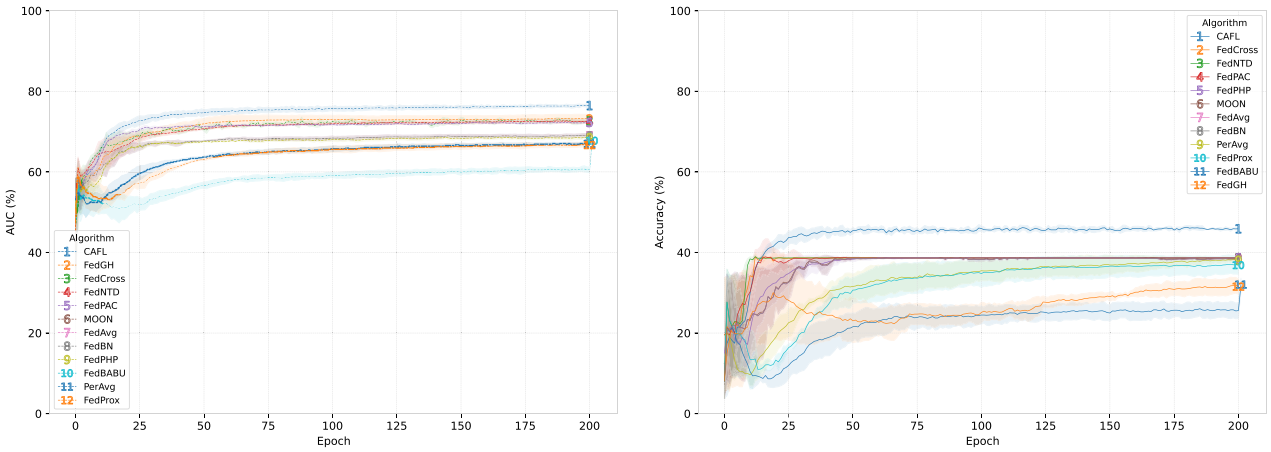
– **Kernel Sensitivity ($\sigma$)**: This parameter in the Gaussian kernel (8) dictates the"sharpness"or selectivity of the attention mechanism. A smaller $\sigma$ makes the attention highly sensitive to the distance between model parameters. We set $\sigma = 0.005$ to enforce a strict collaboration scheme. This encourages clients to assign meaningful attention weights only to a small subset of the most similar peers, effectively creating focused collaborative groups and preventing negative knowledge transfer from dissimilar clients.

## 4.3 Results

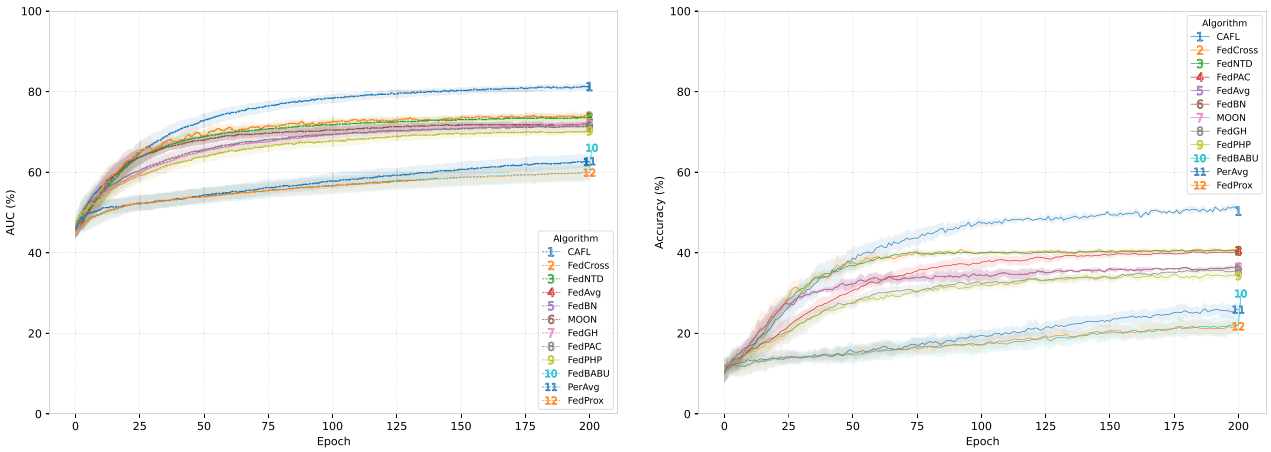
### 4.3.1 Multi-class Classification

To rigorously evaluate CAFL's ability to handle severe concept drift ($P(y|x)$ shift), our primary experiments were conducted on the challenging 8-class emotion classification task. The results unequivocally demonstrate that our proposed method consistently and significantly outperforms a wide range of state-of-the-art federated learning baselines, both when training from scratch and when fine-tuning pre-trained models.

The numerical results in Table 2 and Table 3 clearly show CAFL's superior accuracy and AUC. When training from scratch, the struggles of foundational algorithms like **FedAvg** are expected; its mandatory model averaging process fundamentally conflicts with the goal of personalization by collapsing diverse client concepts into a single, compromised model. Even more advanced methods like




(a) AUC and Accuracy without pre-training

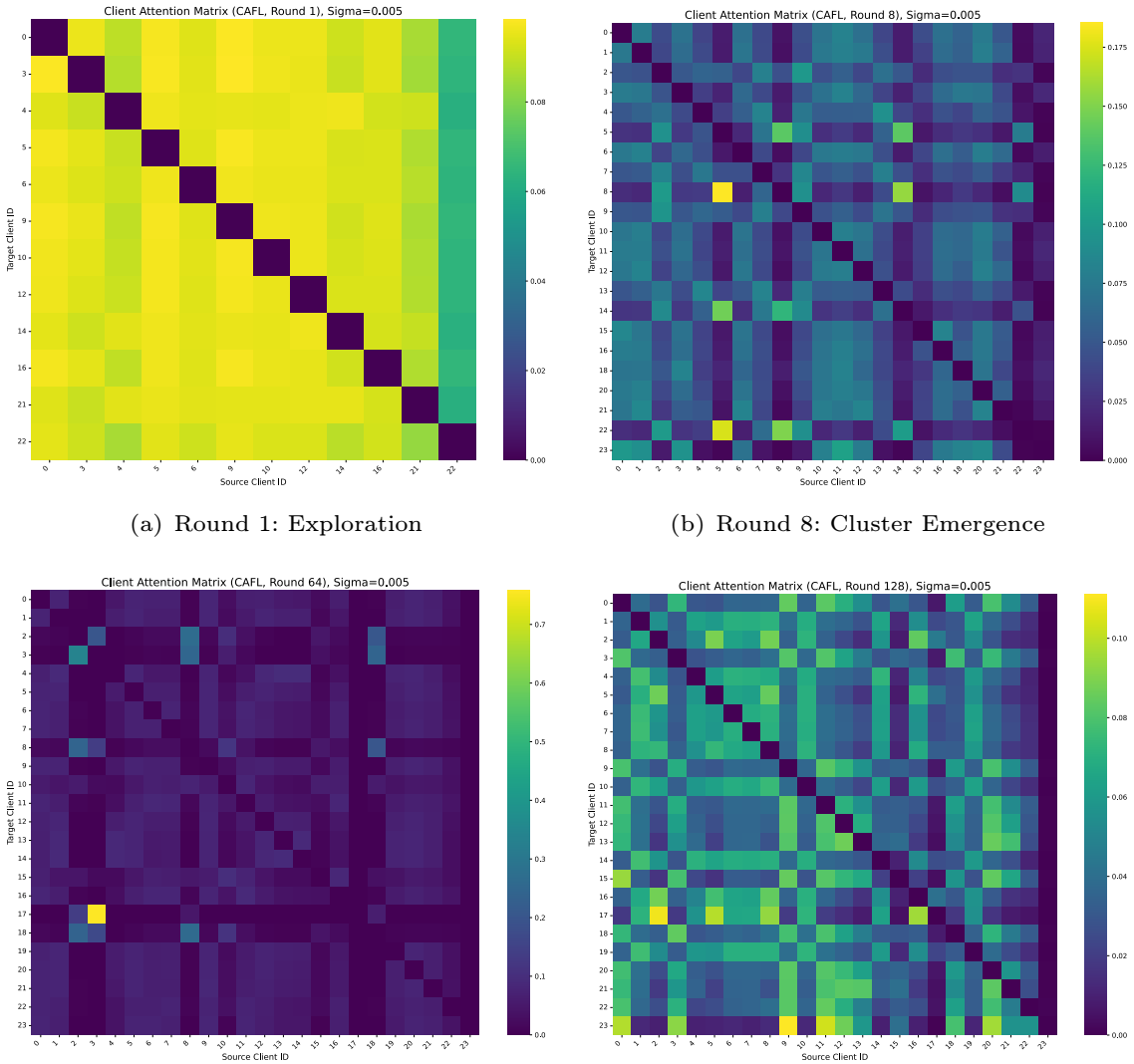



(b) AUC and Accuracy with pre-training

**Fig. 2** Convergence curves for the 8-class classification task on the MobileNet V3 backbone. The plots show the average AUC (left in each subfigure) and Accuracy (right in each subfigure) over 300 communication rounds. In both scenarios, CAFL (blue) achieves superior performance and converges significantly faster than all baselines

**FedCross**, while improving the aggregation strategy, still lack a mechanism to resolve the underlying conceptual conflicts at the client level. With pre-training, CAFL's lead is even more pronounced. This highlights a key insight: while methods like **FedNTD** effectively preserve global knowledge, they do not sufficiently address the need for personalized decision boundaries inherent to $P(y|x)$ drift. Similarly, while **FedBN** is known to excel at $P(x)$ feature shift, it is less effective here because the core problem is not feature style but the feature-to-label mapping itself. CAFL's multi-module design adeptly handles both adapting general knowledge and personalizing this crucial mapping.

Beyond final performance metrics, the convergence rate is a critical indicator of a method's efficiency and stability in federated settings. The learning curves, consolidated in Fig. 2, illustrate this point vividly. In both scenarios (with and without pre-training), CAFL not only reaches a higher final performance but also exhibits a significantly steeper initial learning slope. This demonstrates a much faster convergence rate. This efficiency is not merely a byproduct of faster optimization; it is a direct consequence of CAFL's Collaborative Similarity Matching (Module C). By actively pruning unhelpful collaborations and focusing learning on relevant knowledge from kindred peers, CAFL avoids the wasted cycles spent reconciling conflicting gradients from dissimilar clients—a problem that plagues and slows down methods like FedAvg.




(a) Round 1: Exploration

(b) Round 8: Cluster Emergence

(c) Round 64: Peak Concentration

(d) Round 128: Stabilization

**Fig. 3** The evolution of inter-client attention weights for CAFL with the **MobileNet V3** backbone, shown at key rounds. Yellow indicates high attention, while dark purple indicates low attention. Note the significant changes in the color bar scale, reflecting the dynamic concentration of attention. The process clearly shows a transition from broad exploration to highly focused, stable collaboration clusters

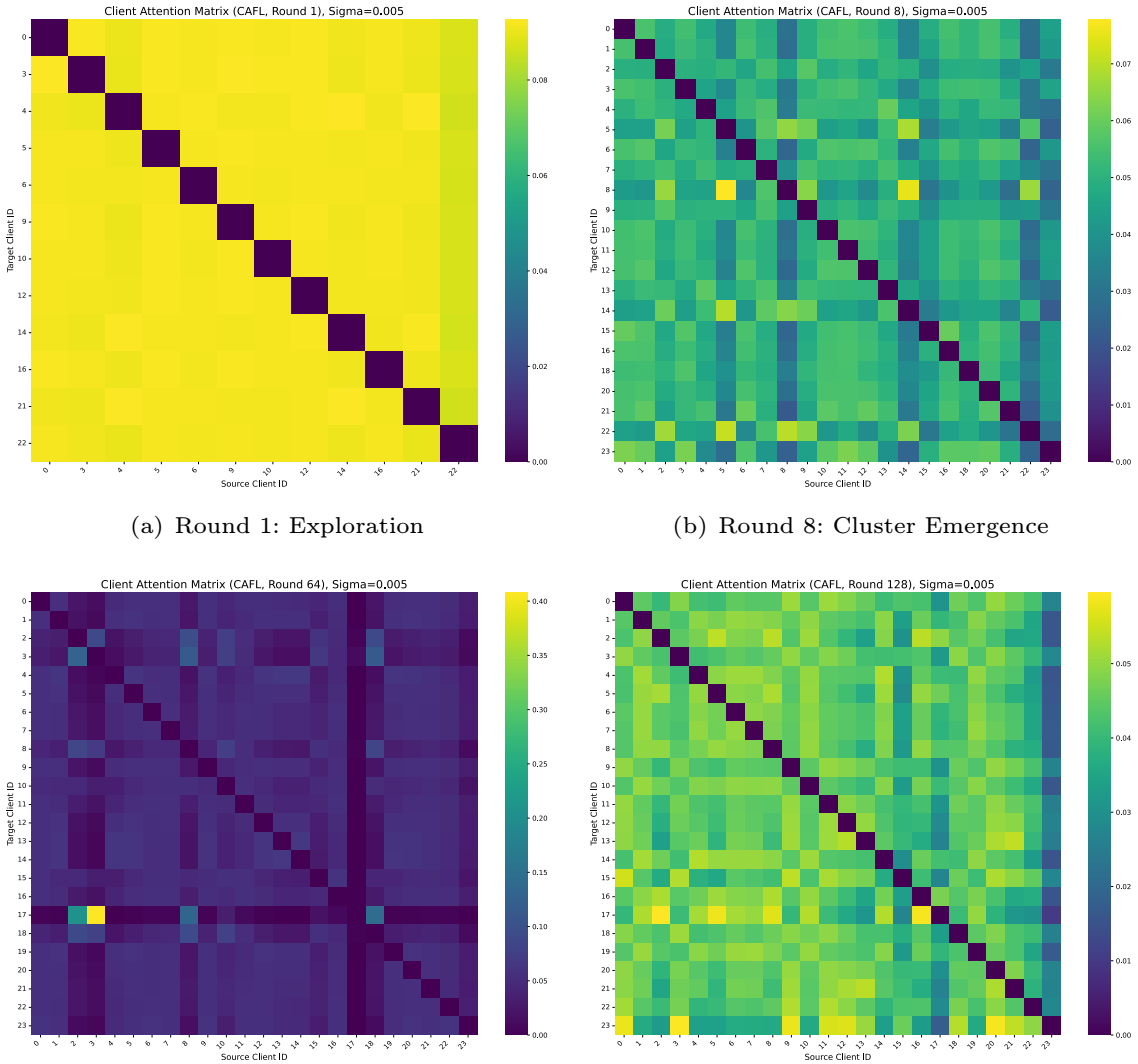
#### 4.3.2 Visualization of the Collaborative Mechanism

To provide a deeper, more intuitive understanding of how CAFL mitigates concept drift, we visualized the core of its collaborative mechanism (Module C). This involved plotting the evolution of the inter-client attention weights ($\alpha_{ij}$ from Eq. 9) as heatmaps over successive communication rounds. By analyzing the complete evolution across different architectures, including the crucial changes in the attention scale (color bar), we can observe a distinct and consistent multi-phase process that empirically validates our approach.

As depicted in Fig. 3 and Fig. 4, the attention patterns for both architectures exhibit a remarkably consistent three-phase evolution:

**Phase 1: Exploration and Low-Intensity Attention (e.g., Round 1)** In the initial rounds, the heatmaps are diffuse. The color bar's maximum value is low (approx. 0.08), indicating that attention is weak and broadly distributed. At this stage, models are near their common initialization, and parameter-space distance is not yet a meaningful proxy for conceptual similarity.

**Phase 2: Cluster Formation and Attention Concentration (e.g., Rounds 8-64)** A striking pattern emerges and intensifies dramatically by round 64. Distinct block-diagonal structures appear, signifying the formation of




(a) Round 1: Exploration

(b) Round 8: Cluster Emergence

(c) Round 64: Peak Concentration

(d) Round 128: Stabilization

**Fig. 4** The evolution of inter-client attention weights for CAFL with the **MobileViT V2** backbone. The same multi-phase evolutionary pattern observed in Fig. 3 is consistently reproduced, confirming that this dynamic behavior is a fundamental characteristic of the CAFL framework itself, rather than an artifact of a specific model architecture

collaborative clusters. Crucially, the scale of the color bar undergoes a massive shift, with peak attention values skyrocketing (e.g., to approx. 0.7 for MobileNet V3). This demonstrates that clients are not only becoming more selective about *who* to collaborate with, but the intensity of that collaboration becomes highly concentrated. CAFL enables certain clients to almost exclusively learn from a very small set of"kindred"peers who share the same $P(y|x)$ concept, while actively isolating them from all others.

**Phase 3: Stabilization and Relaxation (e.g., Round 128)** Interestingly, in later rounds, the attention pattern, while maintaining the clear cluster structure, slightly"relaxes."The peak attention value on the color bar decreases from the extreme high at round 64. This may suggest that once primary conceptual alignment is achieved through intense collaboration, the system can afford to slightly broaden its attention again for fine-tuning and to prevent overfitting to a single peer within the cluster.

This entire evolutionary pattern's consistency across both a pure CNN (MobileNet V3) and a hybrid Transformer (MobileViT V2) architecture is powerful evidence. It confirms that this intelligent, dynamic clustering is a fundamental and robust characteristic of the CAFL framework. This process directly explains CAFL's superior performance and rapid convergence; the model intelligently allocates its collaborative budget, first finding and then intensively learning from the most relevant peers before settling into a stable, personalized state.

**Table 4** Performance comparison on the binary classification task **without pre-trained models**. CAFL continues to achieve the highest performance, though the margins are narrower compared to the multi-class task, likely reflecting the reduced concept drift in a binary setting

| Models | MobileNet V3 [56] | | MobileViT V2 [26] | |
|---|---|---|---|---|
| Methods | AUC | ACC | AUC | ACC |
| FedAvg [17] | 80.41 ± 0.35 | 74.42 ± 0.35 | 74.28 ± 1.85 | 69.81 ± 1.74 |
| FedBABU [45] | 78.57 ± 2.01 | 73.50 ± 2.84 | 71.52 ± 3.34 | 70.22 ± 5.24 |
| FedBN [44] | 80.41 ± 0.35 | 74.42 ± 0.35 | 74.56 ± 2.13 | 69.99 ± 2.01 |
| FedCross [47] | 79.78 ± 0.41 | 74.79 ± 0.00 | 79.64 ± 0.77 | 75.25 ± 0.68 |
| FedGH [50] | 82.78 ± 0.05 | 79.22 ± 0.41 | 72.80 ± 2.54 | 72.21 ± 4.05 |
| FedNTD [37] | 79.51 ± 0.52 | 74.79 ± 0.00 | 78.93 ± 1.48 | 75.12 ± 0.62 |
| FedPAC [48] | 80.69 ± 0.44 | 75.12 ± 0.36 | 73.66 ± 1.56 | 72.81 ± 3.74 |
| FedPHP [49] | 79.43 ± 0.49 | 74.79 ± 0.00 | 70.20 ± 0.96 | 64.91 ± 1.93 |
| FedProx [18] | 77.60 ± 0.37 | 74.79 ± 0.00 | 69.55 ± 4.57 | 70.08 ± 5.47 |
| MOON [36] | 80.59 ± 0.23 | 76.16 ± 0.28 | 74.00 ± 0.49 | 70.54 ± 2.53 |
| Per-FedAvg [54] | 77.93 ± 0.43 | 74.79 ± 0.00 | 71.05 ± 7.21 | 70.59 ± 7.15 |
| **CAFL (ours)** | **85.09 ± 0.33** | **80.33 ± 0.45** | **80.49 ± 0.78** | **76.16 ± 1.44** |

**Table 5** Performance comparison on the binary classification task **with ImageNet pre-trained models**. All methods benefit from fine-tuning, with CAFL maintaining its position as the top-performing algorithm, confirming its robustness across varying task complexities

| Models | MobileNet V3 [56] | | MobileViT V2 [26] | |
|---|---|---|---|---|
| Methods | AUC | ACC | AUC | ACC |
| FedAvg [17] | 81.62 ± 0.20 | 74.84 ± 0.66 | 79.17 ± 0.20 | 75.21 ± 0.34 |
| FedBABU [45] | 77.14 ± 0.48 | 75.39 ± 0.56 | 64.29 ± 1.83 | 64.59 ± 1.97 |
| FedBN [44] | 81.62 ± 0.20 | 74.84 ± 0.66 | 79.17 ± 0.20 | 75.16 ± 0.40 |
| FedCross [47] | 82.77 ± 0.20 | 76.96 ± 0.62 | 77.82 ± 0.54 | 75.39 ± 0.07 |
| FedGH [50] | 81.95 ± 0.38 | 76.36 ± 0.99 | 73.41 ± 1.37 | 71.24 ± 1.93 |
| FedNTD [37] | 81.88 ± 0.14 | 75.85 ± 0.35 | 77.70 ± 0.56 | 75.16 ± 0.07 |
| FedPAC [48] | 81.31 ± 0.20 | 75.95 ± 0.28 | 75.62 ± 0.62 | 74.19 ± 0.83 |
| FedPHP [49] | 80.42 ± 0.03 | 74.47 ± 0.96 | 73.77 ± 0.92 | 72.02 ± 0.79 |
| FedProx [18] | 70.46 ± 0.85 | 68.05 ± 1.77 | 57.56 ± 2.78 | 57.89 ± 3.10 |
| MOON [36] | 81.61 ± 0.20 | 74.61 ± 0.43 | 79.14 ± 0.26 | 75.21 ± 0.30 |
| Per-FedAvg [54] | 74.08 ± 0.85 | 69.81 ± 1.67 | 60.68 ± 2.58 | 61.68 ± 3.25 |
| **CAFL (ours)** | **85.38 ± 0.23** | **81.58 ± 0.30** | **81.84 ± 0.32** | **79.59 ± 0.83** |

**Table 6** Ablation study of CAFL's core components on the binary classification task (AUC %). Each column represents the performance after removing the specified module(s). The full CAFL model significantly outperforms all ablated versions, confirming the necessity and synergy of its components

| **CAFL** | *w.o. A* | *w.o. B* | *w.o. C* | *w.o. A+B* | *w.o. A+C* | *w.o. B+C* | *w.o. A+B+C* |
|---|---|---|---|---|---|---|---|
| **80.49 ± 0.78** | 79.83 ± 0.19 | 76.18 ± 0.68 | 79.52 ± 1.23 | 76.60 ± 0.96 | 77.60 ± 0.84 | 75.00 ± 1.30 | 74.92 ± 0.94 |

#### 4.3.3 Binary Classification Performance

To assess the versatility and robustness of our framework, we also evaluated all methods on a binary emotion classification task (positive vs. negative). This scenario, with its potentially less severe degree of concept drift, serves as a crucial test for whether the benefits of a sophisticated personalization method like CAFL persist in less challenging conditions. The results, presented in Table 4 and Table 5, confirm that CAFL consistently maintains its superior performance.

As shown in the tables, CAFL consistently achieves the highest AUC and Accuracy scores across all settings. For instance, with a pre-trained MobileNet V3 model (Table 5), CAFL achieves an AUC of 85.38%, surpassing the next-best method, FedCross, by a notable margin of 2.61%.

An important insight emerges from comparing these results to the multi-class scenario. Here, the performance gap between CAFL and other advanced methods like FedCross and FedGH is narrower. This is likely because the conceptual divide for binary sentiment (positive vs. negative) is more universally understood across clients, leading to a less severe 'P(y|x)' drift compared to the nuanced 8-class task. However, the fact that CAFL still yields the best performance is significant. It demonstrates that even when conceptual differences are subtle, CAFL's ability to capture fine-grained personalization needs and facilitate targeted collaboration provides a distinct, winning edge. This confirms the robustness of our approach: it excels in the

face of severe concept drift and still delivers superior results when the drift is moderate. The convergence curves, which we omit here for brevity but which show a similar trend to the multi-class case, further confirm CAFL's consistent efficiency.

#### 4.3.4 Ablation Study

To empirically dissect the contribution of each core component within our framework, we conducted a comprehensive ablation study. We systematically deactivated one or more of CAFL's three modules and evaluated the performance impact: Module A (Adaptive Feature Gating), Module B (Bipartite Prediction Head), and Module C (Collaborative Similarity Matching). The results, presented in Table 6, were obtained using the MobileViT V2-050 backbone on the binary classification task without pre-training.

The results in Table 6 lead to several clear and compelling conclusions about our design choices:

- **Module C (Collaborative Similarity Matching) is effective.** Removing only Module C (*w.o. C*) results in a noticeable performance drop. This configuration is akin to a personalized federated learning method where clients train independently with a global model component but lack an intelligent peer-to-peer collaboration mechanism. The drop confirms that our attention-based similarity matching actively promotes positive knowledge transfer and is superior to naive collaboration.
- **Module A (Adaptive Feature Gating) is valuable.** Deactivating Module A (*w.o. A*) also lowers performance. This version of the model can still use separate personalized and global heads, but it must feed the entire feature representation to both. The performance loss indicates that learning a sample-specific policy to dynamically route features into personalized and global streams, as described in Sect. 3.2.2, is a more effective strategy for disentangling knowledge.
- **Module B (Bipartite Prediction Head) is critical.** The most substantial performance degradation occurs when Module B is removed (*w.o. B*, AUC drops to 76.18%). In this case, the model loses the structural capacity to maintain separate decision boundaries for personalized and global knowledge. Forcing both feature streams into a single head creates an information bottleneck and negates the benefits of feature disentanglement. This underscores that the bipartite head architecture (Sect. 3.2.3) is the cornerstone of our personalization strategy.
- **The components have a strong synergistic effect.** The performance degradation is most severe when multiple modules are removed. The worst result is observed when all three modules are deactivated (*w.o. A+B+C*, AUC of 74.92%). This configuration is conceptually equivalent to a simple PFL baseline that only personalizes the feature extractor. The stark difference between the full CAFL model and this baseline vividly illustrates that the components of CAFL are not merely additive but work in synergy to create a robust and effective solution.

In summary, this ablation study provides compelling quantitative evidence that CAFL's superior performance is not due to any single trick, but to the thoughtful integration of its three core mechanisms, each playing a vital and complementary role in addressing the challenges of federated learning under concept drift.

## 5 Conclusion

This paper introduced Conditional Attention Federated Learning (CAFL), a novel framework that effectively addresses the critical challenge of concept drift in federated image emotion analysis. CAFL's unique integration of an adaptive feature gate, a bipartite prediction head, and an attention-based collaboration mechanism was empirically proven to consistently outperform state-of-the-art methods across various tasks and model architectures. Our analyses confirmed that CAFL intelligently forms dynamic collaborative clusters to handle the subjective nature of emotional data, validating our design and its superior performance.

The success of CAFL opens new avenues for privacy-preserving personalization in fields where concept drift is prevalent, such as decentralized healthcare and user modeling. Future work will focus on integrating formal privacy guarantees, like differential privacy, and adapting the framework for continual learning environments where client concepts may evolve over time.

**Acknowledgements** Thank you to professors Bo Du, Yong Luo, Yongchao Xu, and Zengmao Wang for their patient guidance.

**Author Contributions** Chang Liu completed the main work, while professors Bo Du, Yongchao Xu, and Zengmao Wang helped by reviewing the paper and providing suggestions.

**Funding** This work was supported in part by the National Natural Science Foundation of China (Grant No. 62225113, 62222112, and 62176186).

**Data Availibility** Not applicable.

## Declarations

**Competing interests** The authors declare that they have no conflict of interest.

## References

1. Zhao S, Yao H, Jiang X, Sun X (2018) Predicting discrete probability distributions of image emotions. Inf Sci 454:344–358
2. Xu Z, Wang S (2023) Emotional attention detection and correlation exploration for image emotion distribution learning. IEEE Trans Affect Comput 14(1):357–369
3. Yang J, She D, Lai Y-K, Rosin PL, Yang M-H (2018) Weakly supervised coupled networks for visual sentiment analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7584–7592
4. Zhao S, Zhao X, Ding G, Keutzer K (2019) End-to-end image emotion distribution learning. IEEE Trans Multimedia 22(11):2983–2993
5. Machajdik J, Hanbury A (2010) Affective image classification using features inspired by psychology and art theory. In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 83–92
6. Borth D, Ji R, Chen T, Breuel T, Chang S-F (2013) Large-scale visual sentiment ontology and detectors using adjective noun pairs. In: Proceedings of the 21st ACM International Conference on Multimedia, pp. 223–232
7. Zhao S, Yao H, Yang Y, Zhang Y (2014) Affective image retrieval via multi-graph learning. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 1025–1028
8. You Q, Luo J, Jin H, Yang J (2015) Robust image sentiment analysis using progressively trained and domain transferred deep networks. Proceedings of the AAAI Conference on Artificial Intelligence 29:381–388
9. Yang J, She D, Sun M (2017) Joint image emotion classification and distribution learning via deep convolutional neural network. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence, pp. 3266–3272
10. Xiong H, Liu H, Zhong B, Fu Y (2019) Structured and sparse annotations for image emotion distribution learning. Proceedings of the AAAI Conference on Artificial Intelligence 33:563–570
11. Peng K-C, Chen T, Sadovnik A, Gallagher AC (2015) A mixed bag of emotions: Model, predict, and transfer emotion distributions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 860–868
12. Lu X, Jain AK (2012) Shape and texture combined face recognition for detection of fake faces. IEEE Trans Inf Forensics Secur 3(3):771–779
13. You Q, Luo J, Jin H, Yang J (2016) Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30
14. Chen T, Borth D, Darrell T, Chang S-F (2014) Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. In: 2014 IEEE International Conference on Computer Vision Workshop, pp. 296–303. IEEE
15. Zhu X, Li L, Zhang W, Rao T, Xu M, Huang Q, Xu D (2017) Dependency exploitation: A unified cnn-rnn approach for visual emotion recognition. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence, pp. 3595–3601
16. Song K, Yao T, Ling Q, Mei T (2018) Boosting image sentiment analysis with visual attention. Neurocomputing 312:218–228
17. McMahan B, Moore E, Ramage D, Hampson S, Arcas BA (2017) Communication-efficient learning of deep networks from decentralized data. Artificial intelligence and statistics, 1273–1282
18. Li T, Sahu AK, Zaheer M, Sanjabi M, Talwalkar A, Smith V (2020) Federated optimization in heterogeneous networks. Proc Mach Learn Syst 2:429–450
19. Yang Q, Liu Y, Chen T, Tong Y (2019) Federated machine learning: concept and applications. ACM Trans Intell Syst Technol (TIST) 10(2):1–19
20. Kairouz P, McMahan HB, Avent B, Bellet A, Bennis M, Bhagoji AN, Bonawitz K, Charles Z, Cormode G, Cummings R, et al. (2021) Advances and open problems in federated learning. Foundations and Trends® in Machine Learning **14**(1–2), 1–210
21. Sattler F, Wiedemann S, Müller K-R, Samek W (2020) Robust and communication-efficient federated learning from non-iid data. IEEE Trans Neural Netw Learn Syst 31:3400–3413
22. Kang M, Kim S, Jin KH, Adeli E, Pohl KM, Park SH (2024) Fednn: federated learning on concept drift data using weight and adaptive group normalizations. Pattern Recogn 149:110230
23. Karimireddy SP, Kale S, Mohri M, Reddi S, Stich S, Suresh AT (2020) Scaffold: Stochastic controlled averaging for federated learning. International Conference on Machine Learning, 5132–5143
24. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008
25. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C (2018) Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520
26. Mehta S, Rastegari M (2022) Separable self-attention for mobile vision transformers. arXiv preprint arXiv:2206.02680
27. Woo S, Debnath S, Hu R, Chen X, Liu Z, Kweon IS, Xie S (2023) Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16133–16142
28. He C, Li S, So J, Zhang M, Wang H, Wang X, Vepakomma P, Singh A, Qiu H, Shen L, et al. (2020) Fedml: A research library and benchmark for federated machine learning. arXiv preprint arXiv:2007.13518
29. Reddi S, Charles Z, Zaheer M, Garrett Z, Rush K, Konečnỳ J, Kumar S, McMahan HB (2021) Adaptive federated optimization. In: International Conference on Learning Representations
30. Zhao S, Gao Y, Jiang X, Yao H, Chua T-S, Sun X (2014) Exploring principles-of-art features for image emotion recognition. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 47–56
31. You Q, Jin H, Luo J (2017) Visual sentiment analysis by attending on local image regions. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31
32. Peng K-C, Chen T, Sadovnik A, Gallagher AC (2016) Emotion-aware image classification based on affective cues. In: Proceedings of the 24th ACM International Conference on Multimedia, pp. 1172–1176
33. Zhao S, Jia Z, Chen H, Li L, Ding G, Keutzer K (2019) Polarity-consistent deep attention network for fine-grained visual emotion

regression. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 192–201
34. Joshi D, Datta R, Fedorovskaya E, Luong Q-T, Wang JZ, Li J, Luo J (2011) Aesthetics and emotions in images. IEEE Signal Process Mag 28(5):94–115
35. Acar DAE, Zhao Y, Matas R, Mattina M, Whatmough P, Saligrama V (2021) Federated learning based on dynamic regularization. In: International Conference on Learning Representations
36. Li Q, He B, Song D (2021) Model-contrastive federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10713–10722
37. Lee G, Jeong M, Shin Y, Bae S, Yun S-Y (2022) Preservation of the global knowledge by not-true distillation in federated learning. Adv Neural Inf Process Syst 35:38461–38474
38. Liu X, Fang X, Li C, Guo J (2021) An efficient approach for cross-silo federated learning to rank. In: Proceedings of the 37th IEEE International Conference on Data Engineering (ICDE), pp. 1366–1377
39. Yang H, Zhang W, Wang F, Du J (2022) Fed-ltd: Towards cross-platform ride hailing via federated learning to dispatch. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), pp. 1700–1708
40. Zhang Y, Li J, He X, Gu J (2022) Hu-fu: Efficient and secure spatial queries over data federation. In: Proceedings of the 48th International Conference on Very Large Data Bases (VLDB), pp. 12–23
41. Chen H-Y, Chao W-L (2022) On bridging generic and personalized federated learning for image classification. In: International Conference on Learning Representations
42. Collins L, Hassani H, Mokhtari A, Shakkottai S (2021) Exploiting shared representations for personalized federated learning. In: International Conference on Machine Learning, pp. 2089–2099. PMLR
43. Huang Y, Chu L, Zhou Z, Wang L, Liu J, Pei J, Zhang Y (2021) Personalized cross-silo federated learning on non-iid data. Proceedings of the AAAI Conference on Artificial Intelligence 35:7865–7873
44. Li X, Jiang M, Zhang X, Kamp M, Dou Q (2021) Fedbn: Federated learning on non-iid features via local batch normalization. In: International Conference on Learning Representations
45. OH JH, Kim S, Yun S (2022) Fedbabu: Toward enhanced representation for federated image classification. In: 10th International Conference on Learning Representations, ICLR 2022. International Conference on Learning Representations (ICLR)
46. Arivazhagan MG, Aggarwal V, Singh AK, Choudhary S (2019) Federated learning with personalization layers. arXiv preprint arXiv:1912.00818
47. Hu M, Zhou P, Yue Z, Ling Z, Huang Y, Li A, Liu Y, Lian X, Chen M (2024) Fedcross: Towards accurate federated learning via multi-model cross-aggregation. In: 2024 IEEE 40th International Conference on Data Engineering (ICDE), pp. 2137–2150
48. Xu J, Tong X, Huang S-L (2023) Personalized federated learning with feature alignment and classifier collaboration. In: The Eleventh International Conference on Learning Representations
49. Li X-C, Zhan D-C, Shao Y, Li B, Song S (2021) Fedphp: Federated personalization with inherited private models. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 587–602. Springer
50. Yi L, Wang G, Liu X, Shi Z, Yu H (2023) Fedgh: Heterogeneous federated learning with generalized global header. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 8686–8696
51. Zhang J, Hua Y, Wang H, Song T, Xue Z, Ma R, Guan H (2023) Fedcp: Separating feature information for personalized federated learning via conditional policy. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 3249–3261
52. Liu C, Luo Y, Xu Y, Du B (2024) Foundation models matter: federated learning for multi-center tuberculosis diagnosis via adaptive regularization and model-contrastive learning. World Wide Web 27(3):30
53. Nemirovski A, Juditsky A, Lan G, Shapiro A (2009) Robust stochastic approximation approach to stochastic programming. SIAM J Optim 19(4):1574–1609
54. Fallah A, Mokhtari A, Ozdaglar A (2020) Personalized federated learning with theoretical guarantees: a model-agnostic meta-learning approach. Adv Neural Inf Process Syst 33:3557–3568
55. Zhu H, Li L, Wu J, Zhao S, Ding G, Shi G (2020) Personalized image aesthetics assessment via meta-learning with bilevel gradient optimization. IEEE Trans Cybern 52:1798–1811
56. Howard A, Sandler M, Chu G, Chen L-C, Chen B, Tan M, Wang W, Zhu Y, Pang R, Vasudevan V et al. (2019) Searching for mobilenetv3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1314–1324