A APPENDIX

In this section, we conduct additional analysis on the theory of gradient consistency in A.1. In Appendix A.2, we provide detailed information about five FL datasets. And we also introduce additional details about data processing in Appendix A.3. In addition, we discuss the privacy issues of the M^3Fed in Appendix A.4. Algorithm 1 gives the pseudo-code for the M^3Fed framework.

A.1 Analysis Of Angle Lower Bounds

Definition 3 builds on the Zoutendijk condition, which essentially requires that $\theta_{ij} < \frac{\pi}{2} - \gamma$. Let $f(x)$ be the objective function we aim to minimize, and let x_k be the current iterate point. The first-order Taylor approximation of $f(x)$ around x_k is given by:

$$f(x_k + d) \approx f(x_k) + \nabla f(x_k)^T d, \tag{10}$$

where $\nabla f(x_k)$ is the gradient of f at x_k, and d is the search direction.

Suppose d is the gradient update direction from another client. When the gradient update direction from another client d is orthogonal to the current client update direction $\nabla f(x_k)$, $\nabla f(x_k)^T d = 0$, then according to the above approximation, we have:

$$f(x_k + d) \approx f(x_k). \tag{11}$$

This implies that moving along the direction d, which is orthogonal to the gradient, results in an insignificant change in the objective function value $f(x)$ around x_k. In other words, such a move cannot effectively decrease the objective function value.

Therefore, in order to ensure that the objective function value is effectively reduced in each iteration, it is important to avoid choosing a search direction that is orthogonal or approximately orthogonal to the current gradient. This condition ensures that the angle between the gradient update directions has a certain lower bound $\theta_{ij} < \frac{\pi}{2} - \gamma$, which avoids possible orthogonality between the gradient update directions, and thus ensures an effective decrease of the objective function.

A.2 Datasets

AffectNet. The AffectNet dataset comprises over 1,000,000 facial images sourced from the internet, obtained through queries of 1250 emotion-related keywords in six different languages across three major search engines. Approximately half of the retrieved images are manually annotated to identify the presence of seven discrete facial expressions along with valence and arousal intensity. We exclusively utilize the subset of AffectNet containing images labeled with discrete facial expressions. This subset includes 146,198 images labeled as happy, 29,487 as sad, 16,288 as surprised, 8,191 as fearful, 5,264 as disgusted, 28,130 as angry, 5,135 as contemptuous, and 80,276 as neutral. Subsequently, these image data are partitioned into 20 clients for experimental purposes.

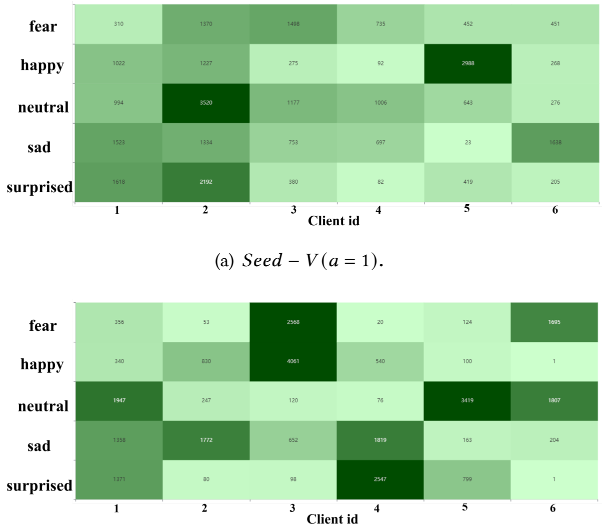
Seed-V. The SEED-V dataset, provided by BCMILab, comprises EEG (Electroencephalogram) and eye-tracking signals obtained from participants watching movie clips. A total of 16 participants (6 male and 10 female) are instructed to watch 15 movie clips, with each clip representing a specific emotion (three clips per emotion). Each participant underwent three experimental sessions. The SEED-V dataset consists of 29,167 samples, including 5,872 labeled as happy, 5,968 labeled as sad, 4,897 labeled as surprised, 4,815 labeled as fearful, and 7,615 labeled as neutral. Subsequently, these data samples are partitioned into six clients for experimental purposes.

Algorithm 1 M_3Fed

```
Input: Communication rounds $C$, number of client K, local datasets
    $\{D_i\}_{i=1}^K$, learning rate $\eta_p, \eta_g$, locat step $E$, meta learner $f_\theta$,
    personalized operator $\{T_i\}_{i=1}^K$ and shared consensus operator
    $\{G_i\}_{i=1}^K$, personalized model $\{\psi_i\}_{i=1}^K$
Server broadcasts $\{f_\theta, G_i\}$ to all clients
for $t$=0 to $C-1$ do
    for Client $i = 1, 2 \ldots K$ in parallel do
        Client Dual-level Optimization $(\{f_\theta, G_i, T_i, \psi_i\}, D_i)$
    end for
    Server executes:
    //Gradient Consistency-based Clustering.
    Calculated to the similarity of meta learner gradient direction
        $\sigma$ according to the Pearson correlation Eq.(5)
    Computing the angular distance $\{A_{ij}\}_{i,j=1}^K$ from Eq.(6)
    Cluster diagonal distances through spectral clustering and
        aggregate within groups $(f_{\theta,a}, f_{\theta,b}, f_{\theta,c} \cdots)$
    //Global Consensus Collaboration Matrix.
    Collaborative correlation matrix $S$ computation for shared
        consensus operators $\{G_i\}_{i=1}^K$ according to Eq.(7)
    Then aggregate the shared consensus operators $G : G_i^{t+1} =$
        $\sum_j^K S_{ij} \cdot G_j^t$
    Server sends $(f_{\theta,a}, f_{\theta,b}, f_{\theta,c} \cdots)$ and $\{G_i^{t+1}\}_{i=1}^K$ to clients
end for
Client Dual-level Optimization $(\{f_\theta, G_i, T_i, \psi_i\}, D_i)$:
for $e = 1$ to $E$ do
    //Personalized Optimization.
    Client-side personalized optimization update: $\psi_i^\star = \psi_i -$
        $\eta_p \nabla(\mathcal{L}_i + \frac{\lambda}{2}\varrho^2), T_i^\star = T_i - \eta_p \nabla(\mathcal{L}_i + \frac{\lambda}{2}\varrho^2)$
    //Global Optimization.
    Client-side global optimization update: $f_\theta^{t+1} = f_\theta^t - \eta_g \nabla(\mathcal{L}_i +$
        $\frac{\lambda}{2}\varrho^2) \langle \psi_i^\star, T_i^\star \rangle, T_i^{t+1} = T_i^t - \eta_g \nabla(\mathcal{L}_i + \frac{\lambda}{2}\varrho^2) \langle \psi_i^\star, T_i^\star \rangle$
end for
Upload the meta learners and shared consensus operators to
    the server
```

UCF-101. The UCF101 dataset, a popular action recognition dataset, is collected from YouTube videos. It comprises 13,320 video clips spanning 101 human action categories, encompassing daily activities to sports, with a total duration of 27 hours. The video lengths vary from a few seconds to over 20 seconds. We utilized only the video modality data, partitioned into eight clients for experimental setup. Ultimately, we obtain an action recognition task across eight clients in the video modality, consisting of a total of 13,320 video instances.

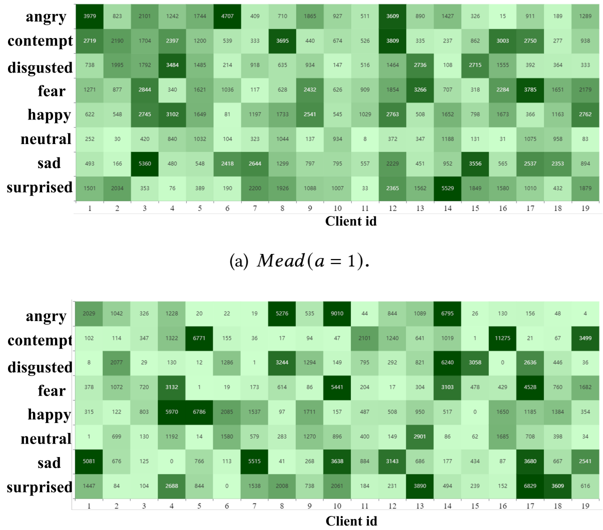
Epic-Kitchens. The dataset is a large-scale egocentric video dataset collected from daily kitchen activities. We conduct the experiment on the version Epic-Kitchens-100, which has 89977 video segments of human-object interaction captured by 37 participants.Sixteen participants also contribute audio data. For the task, we use the unique 97 verb labels as activity classes and only consider audio modality. We partition the speech data accordingly to form



(a) $Seed - V(a = 1)$.

(b) $Seed - V(a = 0.4)$.

Figure 6: Illustration of Seed-V Non-IID data distributions over 6 clients with $\alpha = 1$ and $\alpha = 0.4$. The x-axes represents the client IDs. The y-axes represents the emotion labels. The depth of color represents the amount of data.



(a) $Mead(a = 1)$.

(b) $Mead(a = 0.4)$.

Figure 7: Illustration of MEAD Non-IID data distributions over 20 clients with $\alpha = 1$ and $\alpha = 0.4$.

ten clients. Eventually, we obtain a kitchen behavior recognition task across ten clients in the audio modality, comprising a total of 34,018 instances.

MEAD. The MEAD is a talking-face video corpus featuring 60 actors expressing eight different emotions at three intensity levels (excluding neutral). The videos are recorded from seven different angles in a controlled environment to provide high-quality details of facial expressions. Approximately 40 hours of audiovisual segments are recorded for each actor and viewer. Due to some damaged videos, we utilized data from 47 released actors, totaling 28,749 videos labeled as angry, 28,890 videos labeled as contemptuous, 29,357 videos labeled as disgusted, 29,105 videos labeled as fearful, 29,428 videos labeled as happy, 29,609 videos labeled as sad, 29,349 videos labeled as surprised, and 13,172 videos labeled as neutral. Finally, the data is partitioned into twenty clients for experimentation purposes.

A.3 Additional details

For the five datasets in our experiment, there are 4 unique modalities (i.e., video, audio, eeg, and image). To facilitate the fair comparison with existing methods, we first extract the raw features for different modalities. In particular, we employ the following approach to extract features from different modalities. For video data, a pretrained network of ResNet-3D[24] is used to extract 2048-dimensional visual features of all frames. For audio data, we extract audio 1024-dimensional representations using the current widely used Wav2Vec 2.0[13] speech recognition model. For EEG data, we utilize a pre-trained DCCA[33] model to extract features of 320 dimensions. For image data, the MT-EmotiEffNet[42] model is employed to obtain a 1408-dimensional feature representation for images through pre-training.

Data Partition. Each client is allocated a proportion of the samples of each label according to Dirichlet distribution. In detail, we sample the data by simulating $m_j \sim Dir(\alpha)$ and allocate a portion of $m_{j,i}$ of the samples in class j to client i. Here α controls the degree of skewness. Note that when using this partitioning strategy, the training data of each client may have majority classes, minority classes, or even some missing classes, which is more practical in real-world applications. See Fig.6 for the detailed two Non-IID ($\alpha = 0.4, \alpha = 1$) data partitions on Seed-V datasets.In Fig.7, two non-iid. data partitions on the MEAD dataset are visualized.

A.4 Privacy-Preserving Discussion

In our M^3Fed, we introduce a shared consistency operator to learn the shared feature space projection. This operator primarily focuses on the relationship of feature space projection, without involving customer's private data. We believe this does not pose a risk of privacy leakage. Additionally, addressing the privacy protection issue of model communication, we can adopt techniques such as differential privacy and homomorphic encryption at any time, as used in previous studies[43, 52], to protect model parameters, thereby safeguarding customer data privacy to a certain extent.