AFCL: Achieving Spatio-Temporal Invariance to Data Heterogeneity in Federated Continual Learning

Supplementary Material

A. Appendix for Validity Analyses

In this section, we present detailed validity analyses. Specifically, we first derive the closed-form solution to the linear regression problem discussed in this paper in **Lemma 1** within Appendix A.1. Subsequently, we prove that the global model recursively derived through AFCL is precisely equivalent to the optimal global model obtained via centralized joint training on the complete dataset in **Theorem 1** within Appendix A.2. Finally, we validate AFCL's ideal property of spatio-temporal invariance to non-IID data in **Theorem 2** within Appendix A.3. Detailed proofs are provided below.

A.1. Theoretical Analyses on Lemma 1

Here, employing the least-squares method with ℓ_2 regularization, we theoretically derive the closed-form solution to the linear regression problem discussed in this paper, as presented in Lemma 1. This result not only validates the analytic local training within AFCL but also yields the closed-form solution for the optimal global model. The detailed proof of Lemma 1 follows:
Lemma 1: For any linear regression problem of the form:

$$\underset{\mathbf{W}}{\operatorname{argmin}} \ \|\mathbf{Y}-\mathbf{F}\mathbf{W}\|_{\mathrm{F}}^2+\gamma\|\mathbf{W}\|_{\mathrm{F}}^2, \tag{14}$$

there exists a closed-form solution given by:

$$\mathbf{W}=\left(\mathbf{F}^{\top}\mathbf{F}+\gamma\mathbf{I}\right)^{-1}\mathbf{F}^{\top}\mathbf{Y}. \tag{15}$$

Proof. The objective function in (14) can be expanded as:

$$\begin{aligned}\|\mathbf{Y}-\mathbf{F}\mathbf{W}\|_F^2+\gamma\|\mathbf{W}\|_F^2 &= \operatorname{Tr}\left[(\mathbf{Y}-\mathbf{F}\mathbf{W})^{\top}(\mathbf{Y}-\mathbf{F}\mathbf{W})\right]+\gamma\operatorname{Tr}(\mathbf{W}^{\top}\mathbf{W})\\ &= \operatorname{Tr}(\mathbf{Y}^{\top}\mathbf{Y})-2\cdot\operatorname{Tr}(\mathbf{Y}^{\top}\mathbf{F}\mathbf{W})+\operatorname{Tr}(\mathbf{W}^{\top}\mathbf{F}^{\top}\mathbf{F}\mathbf{W})+\gamma\operatorname{Tr}(\mathbf{W}^{\top}\mathbf{W}).\end{aligned} \tag{16}$$

Subsequently, we further compute the derivative of the objective function with respect to $\mathbf{W}$ as follows:

$$\frac{\partial}{\partial\mathbf{W}}\left(\|\mathbf{Y}-\mathbf{F}\mathbf{W}\|_F^2+\gamma\|\mathbf{W}\|_{\mathrm{F}}^2\right)=-2\mathbf{F}^{\top}\mathbf{Y}+2\mathbf{F}^{\top}\mathbf{F}\mathbf{W}+2\gamma\mathbf{W}. \tag{17}$$

By setting the derivative to zero, we can obtain:

$$(\mathbf{F}^{\top}\mathbf{F}+\gamma\mathbf{I})\mathbf{W}=\mathbf{F}^{\top}\mathbf{Y}. \tag{18}$$

Since $(\mathbf{F}^{\top}\mathbf{F}+\gamma\mathbf{I})$ is positive-definite, we can obtain:

$$\mathbf{W}=(\mathbf{F}^{\top}\mathbf{F}+\gamma\mathbf{I})^{-1}\mathbf{F}^{\top}\mathbf{Y}. \tag{19}$$

■

A.2. Theoretical Analysis on Theorem 1

Here, we theoretically prove in Theorem 1 that the global model recursively derived through our AFCL is exactly equivalent to the optimal global model obtained via centralized joint learning over the complete dataset. Our proof involves two primary stages: (1) Given that the *Global Knowledge Matrix* serves as the fundamental to computing the global model in our AFCL, we inductively establish its closed-form expression using the *Woodbury Matrix Identity* in Lemma 3. (2) We substitute this closed-form expression into (10) to substantiate Theorem 1. To facilitate the foregoing proof, we first present the detailed *Woodbury Matrix Identity* in Lemma 2 as follows:

Lemma 2 (Woodbury Matrix Identity [16]): For matrices $\mathbf{D} \in \mathbb{R}^{n\times n}$, $\mathbf{U} \in \mathbb{R}^{n\times m}$, $\mathbf{E} \in \mathbb{R}^{m\times m}$, and $\mathbf{V} \in \mathbb{R}^{m\times n}$, if $\mathbf{D}$ and $\mathbf{E}$ are both reversible, we can derive the following expression:

$$(\mathbf{D}+\mathbf{UEV})^{-1} = \mathbf{D}^{-1} - \mathbf{D}^{-1}\mathbf{U}(\mathbf{E}^{-1}+\mathbf{VD}^{-1}\mathbf{U})^{-1}\mathbf{VD}^{-1}. \tag{20}$$

Proof. First, we multiply $\mathbf{D}+\mathbf{UEV}$ on the right by $\mathbf{D}^{-1}$ and obtain:

$$(\mathbf{D}+\mathbf{UEV})\mathbf{D}^{-1} = \mathbf{I}+\mathbf{UEVD}^{-1}. \tag{21}$$

Next, we right-multiply (21) by $\mathbf{U}$ and obtain:

$$(\mathbf{D}+\mathbf{UEV})\mathbf{D}^{-1}\mathbf{U} = \mathbf{U}+\mathbf{UEVD}^{-1}\mathbf{U}. \tag{22}$$

Since $\mathbf{E}$ is reversible, we can obtain:

$$(\mathbf{D}+\mathbf{UEV})\mathbf{A}^{-1}\mathbf{U} = \mathbf{UE}(\mathbf{E}^{-1}+\mathbf{VD}^{-1}\mathbf{U}). \tag{23}$$

Therefore, we can represent $\mathbf{UE}$ as

$$\mathbf{UE} = (\mathbf{D}+\mathbf{UEV})\mathbf{D}^{-1}\mathbf{U}(\mathbf{E}^{-1}+\mathbf{VD}^{-1}\mathbf{U})^{-1}. \tag{24}$$

Substituting it into $\mathbf{D}+\mathbf{UEV}$, we can obtain

$$\mathbf{D}+\mathbf{UEV} = \mathbf{D}+(\mathbf{D}+\mathbf{UEV})\mathbf{D}^{-1}\mathbf{U}(\mathbf{E}^{-1}+\mathbf{VD}^{-1}\mathbf{U})^{-1}\mathbf{V}. \tag{25}$$

So (21) can be rewritten as

$$(\mathbf{D}+\mathbf{UEV})\mathbf{D}^{-1} = \mathbf{I}+(\mathbf{D}+\mathbf{UEV})\mathbf{D}^{-1}\mathbf{U}(\mathbf{E}^{-1}+\mathbf{VD}^{-1}\mathbf{U})^{-1}\mathbf{VD}^{-1}. \tag{26}$$

Then we left-multiply it by $(\mathbf{D}+\mathbf{UEV})^{-1}$:

$$\mathbf{D}^{-1} = (\mathbf{D}+\mathbf{UEV})^{-1}+\mathbf{D}^{-1}\mathbf{U}(\mathbf{E}^{-1}+\mathbf{VD}^{-1}\mathbf{U})^{-1}\mathbf{VD}^{-1}. \tag{27}$$

Finally, we can transfer items to obtain

$$(\mathbf{D}+\mathbf{UEV})^{-1} = \mathbf{D}^{-1} - \mathbf{D}^{-1}\mathbf{U}(\mathbf{E}^{-1}+\mathbf{VD}^{-1}\mathbf{U})^{-1}\mathbf{VD}^{-1}. \tag{28}$$

■

Leveraging the *Woodbury Matrix Identity* presented in Lemma 2, we further establish the validity of the recursive update formulas of the *Global Knowledge Matrix* and derive its closed-form expressions in Lemma 3. The detailed proof follows:

Lemma 3: We consider the following recursive update formulation between $\mathbf{G}_{k-1}$ and $\mathbf{G}_k$:

$$\mathbf{G}_k = \left[\mathbf{A}_k\mathbf{G}_{k-1}+\mathbf{B}_k\hat{\mathbf{W}}_k \quad \mathbf{B}_k\check{\mathbf{W}}_k\right], k \in (1, K], \tag{29}$$

where

$$\begin{cases} \mathbf{A}_k = \mathbf{I}-(\tilde{\mathbf{R}}_{k-1})^{-1}\mathbf{R}_k(\mathbf{I}-(\tilde{\mathbf{R}}_k)^{-1}\mathbf{R}_k), & \tilde{\mathbf{R}}_k = \sum_{i=1}^{k}\mathbf{R}_i, \\ \mathbf{B}_k = \mathbf{I}-(\mathbf{R}_k)^{-1}\tilde{\mathbf{R}}_{k-1}(\mathbf{I}-(\tilde{\mathbf{R}}_k)^{-1}\tilde{\mathbf{R}}_{k-1}), & \mathbf{R}_i = \mathbf{F}_i^{\top}\mathbf{F}_i+\gamma\mathbf{I}. \end{cases} \tag{30}$$

The closed-form expression of $\mathbf{G}_k$ is given by:

$$\mathbf{G}_k = (\mathbf{F}_{1:k}^{\top}\mathbf{F}_{1:k}+k\gamma\mathbf{I})^{-1}(\mathbf{F}_{1:k}^{\top}\mathbf{Y}_{1:k}). \tag{31}$$

Proof. We employ mathematical induction to prove this lemma in detail. Specifically, we first establish the base case, i.e., for the first joined (virtual) client, the initial *Global Knowledge Matrix* $\mathbf{G}_1$ satisfies (31). Then, assuming that (31) holds for the arbitrary $\mathbf{G}_k$, we prove that the corresponding $\mathbf{G}_{k+1}$ obtained via (29) and (30), also satisfies (31). Through mathematical induction, we can thus prove that all *Global Knowledge Matrices* satisfy (31). The detailed proof is presented as follows.

(1) Base Case: Let's consider the initial case for the first joined (virtual) client. As detailed in Section 3.4, the initial value of the *Global Knowledge Matrix* is set to $\mathbf{G}_1 = \check{\mathbf{W}}_1$. Since no local model exists for known classes at this point, and based on (12), $\mathbf{G}_1$ can be numerically expressed as:

$$\mathbf{G}_1 = \check{\mathbf{W}}_1 = (\mathbf{F}_1^\top \mathbf{F}_1 + \gamma \mathbf{I})^{-1}(\mathbf{F}_1^\top \mathbf{Y}_1), \tag{32}$$

which clearly satisfies the form in (31).

(2) Inductive Hypothesis: Assume that the *Global Knowledge Matrix* $\mathbf{G}_k$ obtained after aggregating the first k clients satisfies (31), i.e.,

$$\mathbf{G}_k = (\mathbf{F}_{1:k}^\top \mathbf{F}_{1:k} + k\gamma \mathbf{I})^{-1}(\mathbf{F}_{1:k}^\top \mathbf{Y}_{1:k}). \tag{33}$$

(3) Inductive Step: Here, we further demonstrate that $\mathbf{G}_{k+1}$, derived from $\mathbf{G}_k$ using (29) and (30), satisfies (31). By substituting (31) into $\mathbf{A}_{k+1}\mathbf{G}_k$, we can obtain:

$$\begin{aligned}
\mathbf{A}_{k+1}\mathbf{G}_k &= [\mathbf{I} - (\tilde{\mathbf{R}}_k)^{-1}\mathbf{R}_{k+1}(\mathbf{I} - (\tilde{\mathbf{R}}_{k+1})^{-1}\mathbf{R}_{k+1})](\tilde{\mathbf{R}}_k)^{-1}\mathbf{F}_{1:k}^\top \mathbf{Y}_{1:k} \\
&= [\mathbf{I} - (\tilde{\mathbf{R}}_k)^{-1}\mathbf{R}_{k+1}(\mathbf{I} - (\mathbf{R}_{k+1} + \tilde{\mathbf{R}}_k)^{-1}\mathbf{R}_{k+1})](\tilde{\mathbf{R}}_k)^{-1}\mathbf{F}_{1:k}^\top \mathbf{Y}_{1:k} \\
&= [\mathbf{I} - (\tilde{\mathbf{R}}_k)^{-1}(\mathbf{R}_{k+1} - \mathbf{R}_{k+1}(\mathbf{R}_{k+1} + \tilde{\mathbf{R}}_k)^{-1}\mathbf{R}_{k+1})](\tilde{\mathbf{R}}_k)^{-1}\mathbf{F}_{1:k}^\top \mathbf{Y}_{1:k}.
\end{aligned} \tag{34}$$

According to the *Woodbury Matrix Identity* [16], for any matrices $\mathbf{D} \in \mathbb{R}^{n\times n}$, $\mathbf{U} \in \mathbb{R}^{n\times m}$, $\mathbf{E} \in \mathbb{R}^{m\times m}$, and $\mathbf{V} \in \mathbb{R}^{m\times n}$ presented in Lemma 2, the following equation holds consistently:

$$(\mathbf{D} + \mathbf{UEV})^{-1} = \mathbf{D}^{-1} - \mathbf{D}^{-1}\mathbf{U}(\mathbf{E}^{-1} + \mathbf{VD}^{-1}\mathbf{U})^{-1}\mathbf{VD}^{-1}. \tag{35}$$

Subsequently, based on (35), we substitute both $\mathbf{U}$ and $\mathbf{V}$ with $\mathbf{I}$. This adjustment allows us to derive:

$$(\mathbf{D} + \mathbf{E})^{-1} = \mathbf{D}^{-1} - \mathbf{D}^{-1}(\mathbf{D}^{-1} + \mathbf{E}^{-1})^{-1}\mathbf{D}^{-1}, \tag{36}$$

$$\mathbf{D}^{-1}(\mathbf{D}^{-1} + \mathbf{E}^{-1})^{-1}\mathbf{D}^{-1} = \mathbf{D}^{-1} - (\mathbf{D} + \mathbf{E})^{-1}. \tag{37}$$

By replacing $\mathbf{D}^{-1}$ and $\mathbf{E}^{-1}$ in (37) with $\mathbf{R}_{k+1}$ and $\tilde{\mathbf{R}}_k$ respectively, yields:

$$\mathbf{R}_{k+1}(\mathbf{R}_{k+1} + \tilde{\mathbf{R}}_k)^{-1}\mathbf{R}_{k+1} = \mathbf{R}_{k+1} - ((\tilde{\mathbf{R}}_k)^{-1} + (\mathbf{R}_{k+1})^{-1})^{-1}. \tag{38}$$

Incorporating (38) into (34), we can find:

$$\begin{aligned}
\mathbf{A}_{k+1}\mathbf{G}_k &= [\mathbf{I} - (\tilde{\mathbf{R}}_k)^{-1}(\mathbf{R}_{k+1} - \mathbf{R}_{k+1}(\mathbf{R}_{k+1} + \tilde{\mathbf{R}}_k)^{-1}\mathbf{R}_{k+1})](\tilde{\mathbf{R}}_k)^{-1}\mathbf{F}_{1:k}^\top \mathbf{Y}_{1:k} \\
&= [\mathbf{I} - (\tilde{\mathbf{R}}_k)^{-1}(\mathbf{R}_{k+1} - \mathbf{R}_{k+1} + ((\tilde{\mathbf{R}}_k)^{-1} + (\mathbf{R}_{k+1})^{-1})^{-1})](\tilde{\mathbf{R}}_k)^{-1}\mathbf{F}_{1:k}^\top \mathbf{Y}_{1:k} \\
&= [(\tilde{\mathbf{R}}_k)^{-1} - (\tilde{\mathbf{R}}_k)^{-1}((\tilde{\mathbf{R}}_k)^{-1} + (\mathbf{R}_{k+1})^{-1})^{-1}(\tilde{\mathbf{R}}_k)^{-1}]\mathbf{F}_{1:k}^\top \mathbf{Y}_{1:k}.
\end{aligned} \tag{39}$$

Subsequently, replacing $\mathbf{D}$ and $\mathbf{E}$ in (37) with $\tilde{\mathbf{R}}_k$ and $\mathbf{R}_{k+1}$ respectively, yields:

$$(\tilde{\mathbf{R}}_k)^{-1}((\tilde{\mathbf{R}}_k)^{-1} + (\mathbf{R}_{k+1})^{-1})^{-1}(\tilde{\mathbf{R}}_k)^{-1} = (\tilde{\mathbf{R}}_k)^{-1} - (\tilde{\mathbf{R}}_k + \mathbf{R}_{k+1})^{-1}. \tag{40}$$

By incorporating (40) into (39), we can obtain:

$$\begin{aligned}
\mathbf{A}_{k+1}\mathbf{G}_k &= [(\tilde{\mathbf{R}}_k)^{-1} - (\tilde{\mathbf{R}}_k)^{-1}((\tilde{\mathbf{R}}_k)^{-1} + (\mathbf{R}_{k+1})^{-1})^{-1}(\tilde{\mathbf{R}}_k)^{-1}]\mathbf{F}_{1:k}^\top \mathbf{Y}_{1:k} \\
&= [(\tilde{\mathbf{R}}_k)^{-1} - (\tilde{\mathbf{R}}_k)^{-1} + (\tilde{\mathbf{R}}_k + \mathbf{R}_{k+1})^{-1}]\mathbf{F}_{1:k}^\top \mathbf{Y}_{1:k} \\
&= (\tilde{\mathbf{R}}_{k+1})^{-1}\mathbf{F}_{1:k}^\top \mathbf{Y}_{1:k}.
\end{aligned} \tag{41}$$

Similarly, we can obtain:

$$\mathbf{B}_{k+1}\hat{\mathbf{W}}_{k+1} = (\tilde{\mathbf{R}}_{k+1})^{-1}\mathbf{F}_{k+1}^\top \hat{\mathbf{Y}}_{k+1}, \quad \mathbf{B}_{k+1}\check{\mathbf{W}}_{k+1} = (\tilde{\mathbf{R}}_{k+1})^{-1}\mathbf{F}_{k+1}^\top \check{\mathbf{Y}}_{k+1}. \tag{42}$$

By substituting (41) and (42), into (29), we can further derive:

$$\begin{aligned}\mathbf{G}_{k+1} &= \begin{bmatrix}\mathbf{A}_{k+1}\mathbf{G}_k + \mathbf{B}_{k+1}\hat{\mathbf{W}}_{k+1} & \mathbf{B}_{k+1}\check{\mathbf{W}}_{k+1}\end{bmatrix} \\ &= \begin{bmatrix}(\tilde{\mathbf{R}}_{k+1})^{-1}\mathbf{F}_{1:k}^{\top}\mathbf{Y}_{1:k} + (\tilde{\mathbf{R}}_{k+1})^{-1}\mathbf{F}_{k+1}^{\top}\hat{\mathbf{Y}}_{k+1} & (\tilde{\mathbf{R}}_{k+1})^{-1}\mathbf{F}_{k+1}^{\top}\check{\mathbf{Y}}_{k+1}\end{bmatrix} \\ &= \left[\sum_{i=1}^{k+1}(\mathbf{F}_i^{\top}\mathbf{F}_i + \gamma\mathbf{I})\right]^{-1}\begin{bmatrix}\mathbf{F}_{1:k}^{\top}\mathbf{Y}_{1:k} + \mathbf{F}_{k+1}^{\top}\hat{\mathbf{Y}}_{k+1} & \mathbf{F}_{k+1}^{\top}\check{\mathbf{Y}}_{k+1}\end{bmatrix}.\end{aligned} \tag{43}$$

Considering the presence of known and unknown classes in a class-incremental scenario, the complete feature matrix $\mathbf{F}_{1:k+1}$ and the corresponding label matrix $\mathbf{Y}_{1:k+1}$ satisfy:

$$\mathbf{F}_{1:k+1} = \begin{bmatrix}\mathbf{F}_{1:k} \\ \mathbf{F}_{k+1}\end{bmatrix}, \quad \mathbf{Y}_{1:k+1} = \begin{bmatrix}\mathbf{Y}_{1:k} & \mathbf{0} \\ \hat{\mathbf{Y}}_{k+1} & \check{\mathbf{Y}}_{k+1}\end{bmatrix}. \tag{44}$$

Therefore, we can obtain:

$$\begin{aligned}\mathbf{F}_{1:k+1}^{\top}\mathbf{F}_{1:k+1} + (k+1)\gamma\mathbf{I} &= \begin{bmatrix}\mathbf{F}_{1:k}^{\top} & \mathbf{F}_{k+1}^{\top}\end{bmatrix}\begin{bmatrix}\mathbf{F}_{1:k} \\ \mathbf{F}_{k+1}\end{bmatrix} + (k+1)\gamma\mathbf{I} \\ &= \mathbf{F}_{1:k}^{\top}\mathbf{F}_{1:k} + \mathbf{F}_{k+1}^{\top}\mathbf{F}_{k+1} + (k+1)\gamma\mathbf{I} \\ &= \sum\nolimits_{i=1}^{k+1}(\mathbf{F}_i^{\top}\mathbf{F}_i + \gamma\mathbf{I}),\end{aligned} \tag{45}$$

$$\begin{aligned}\mathbf{F}_{1:k+1}^{\top}\mathbf{Y}_{1:k+1} &= \begin{bmatrix}\mathbf{F}_{1:k}^{\top} & \mathbf{F}_{k+1}^{\top}\end{bmatrix}\begin{bmatrix}\mathbf{Y}_{1:k} & \mathbf{0} \\ \hat{\mathbf{Y}}_{k+1} & \check{\mathbf{Y}}_{k+1}\end{bmatrix} \\ &= \begin{bmatrix}\mathbf{F}_{1:k}^{\top}\mathbf{Y}_{1:k} + \mathbf{F}_{k+1}^{\top}\hat{\mathbf{Y}}_{k+1} & \mathbf{F}_{k+1}^{\top}\check{\mathbf{Y}}_{k+1}\end{bmatrix}.\end{aligned} \tag{46}$$

By substituting (45) and (46) into (43), we can further derive:

$$\begin{aligned}\mathbf{G}_{k+1} &= \left[\sum_{i=1}^{k+1}(\mathbf{F}_i^{\top}\mathbf{F}_i + \gamma\mathbf{I})\right]^{-1}\begin{bmatrix}\mathbf{F}_{1:k}^{\top}\mathbf{Y}_{1:k} + \mathbf{F}_{k+1}^{\top}\hat{\mathbf{Y}}_{k+1} & \mathbf{F}_{k+1}^{\top}\check{\mathbf{Y}}_{k+1}\end{bmatrix} \\ &= \left[\mathbf{F}_{1:k+1}^{\top}\mathbf{F}_{1:k+1} + (k+1)\gamma\mathbf{I}\right]^{-1}(\mathbf{F}_{1:k+1}^{\top}\mathbf{Y}_{1:k+1}).\end{aligned} \tag{47}$$

The result shows that if $\mathbf{G}_k$ satisfies (31), then the corresponding $\mathbf{G}_{k+1}$ obtained from (29) and (30) also satisfies (31).

(4) Conclusion: By the principle of mathematical induction, for any *Global Knowledge Matrix* $\mathbf{G}_k$ derived through our AFCL, the corresponding closed-form expression is given by (31). This also confirms the validity of our recursive update method for the *Global Knowledge Matrix*.

■

Based on Lemma 3, we can obtain the analytical expression of the *Global Knowledge Matrix*. Building on this result, we further analyze the validity of our AFCL and theoretically prove that the global model recursively derived through (10) is exactly equivalent to the optimal global model obtained via centralized joint learning over the complete dataset in Theorem 1.

Theorem 1: We consider the following computational formulation:

$$\mathbf{W}_k = \mathbf{G}_k + [(k-1)\gamma(\tilde{\mathbf{R}}_k - (k-1)\gamma\mathbf{I})^{-1}]\mathbf{G}_k, \tag{48}$$

where $\mathbf{G}_k$ and $\tilde{\mathbf{R}}_k$ are obtained from (8) and (9), respectively. The computation result of (48) is exactly equivalent to (13), which corresponds to the optimal solution (i.e., the centralized joint learning) of empirical risk minimization over the full datasets $\mathcal{D}_{1:k}$ from the first k clients.

Proof. By further transforming the analytical expression of $\mathbf{G}_k$ derived in Lemma 2, we can obtain:

$$\begin{aligned}\mathbf{G}_k &= (\mathbf{F}_{1:k}^{\top}\mathbf{F}_{1:k} + k\gamma\mathbf{I})^{-1}(\mathbf{F}_{1:k}^{\top}\mathbf{Y}_{1:k}) \\ &= (\sum\nolimits_{i=1}^{k}\mathbf{F}_k^{\top}\mathbf{F}_k + k\gamma\mathbf{I})^{-1}(\mathbf{F}_{1:k}^{\top}\mathbf{Y}_{1:k}) \\ &= (\tilde{\mathbf{R}}_k)^{-1}(\mathbf{F}_{1:k}^{\top}\mathbf{Y}_{1:k}).\end{aligned} \tag{49}$$

By substituting (49) into (48), we can further derive:

$$\begin{aligned}\mathbf{W}_k &= \mathbf{G}_k + [(k-1)\gamma(\tilde{\mathbf{R}}_k - (k-1)\gamma\mathbf{I})^{-1}]\mathbf{G}_k \\ &= [(\tilde{\mathbf{R}}_k - (k-1)\gamma\mathbf{I})^{-1}(\tilde{\mathbf{R}}_k - (k-1)\gamma\mathbf{I}) + (\tilde{\mathbf{R}}_k - (k-1)\gamma\mathbf{I})^{-1}(k-1)\gamma\mathbf{I}]\mathbf{G}_k \\ &= [(\tilde{\mathbf{R}}_k - (k-1)\gamma\mathbf{I})^{-1}]\tilde{\mathbf{R}}_k\mathbf{G}_k \\ &= (\sum\nolimits_{i=1}^{k}\mathbf{F}_k^\top\mathbf{F}_k + \gamma\mathbf{I})^{-1}\tilde{\mathbf{R}}_k(\tilde{\mathbf{R}}_k)^{-1}(\mathbf{F}_{1:k}^\top\mathbf{Y}_{1:k}) \\ &= (\mathbf{F}_{1:k}^\top\mathbf{F}_{1:k} + \gamma\mathbf{I})^{-1}(\mathbf{F}_{1:k}^\top\mathbf{Y}_{1:k}).\end{aligned} \tag{50}$$

It is evident that the computation result of (48) is exactly equivalent to the result derived from (13). In other words, the global model derived from our AFCL is identical to the optimal result obtained via centralized joint learning.

■

A.3. Theoretical Analysis on AFCL's Spatio-temporal Invariance

Here, building upon the previously established proofs, we further provide a detailed theoretical validation in Theorem 2 for the *spatio-temporal invariance to Non-IID data* and *order invariance across local clients* inherent in our AFCL.

Theorem 2: The final global model obtained by our AFCL is independent of the spatio-temporal data heterogeneity and client registration order, being identical to that obtained by centralized joint learning over the full dataset $\mathcal{D}_{1:K}$.

Proof. First of all, let's analyze the fundamental impact of different cases of spatio-temporal data heterogeneity and client registration order, as outlined below:

(1) Spatio-temporal data heterogeneity reflects the Non-IID characteristics of local datasets across (virtual) clients. Therefore, different cases of spatio-temporal data heterogeneity fundamentally correspond to the reassignment of a subset of samples among clients' local datasets $\{\mathcal{D}_k\}_{k=1}^K$.

(2) Client registration order determines the sequencing of the clients' local datasets within the complete dataset. Therefore, different cases of client registration order fundamentally correspond to reordering the clients' local datasets $\{\mathcal{D}_k\}_{k=1}^K$ within the complete dataset $\mathcal{D}_{1:K}$. Furthermore, although the client registration order also affects the generated one-hot encoding functions, this influence is limited to inducing a consistent permutation of the columns of the resulting global model parameters, without altering their substantive values. In other words, the weights learned for each class remain invariant, and thus changes in the one-hot encoding functions do not impact the final global model.

In summary, different cases of spatio-temporal data heterogeneity and client registration order can be interpreted as reordering the samples within the complete dataset $\mathcal{D}_{1:K}$.

Subsequently, we further demonstrate that reordering the samples does not affect either the optimal model or the model obtained through our AFCL. Specifically, we denote the reordered feature matrix and label matrix as $\mathbf{F}_{1:K}^*$ and $\mathbf{Y}_{1:K}^*$, respectively, which can be represented as:

$$\mathbf{F}_{1:K}^* = \mathbf{\Pi}\mathbf{F}_{1:K}, \quad \mathbf{Y}_{1:K}^* = \mathbf{\Pi}\mathbf{Y}_{1:K}, \tag{51}$$

where $\mathbf{\Pi}$ is the corresponding permutation matrix. Due to the orthogonality of the permutation matrix, it satisfies $\mathbf{\Pi}^{-1} = \mathbf{\Pi}^\top$. According to Lemma 2 and Theorem 1, the global model obtained recursively and distributively via our AFCL is fully equivalent to the optimal result derived from centralized joint learning on the full dataset. We denote the global model obtained by AFCL based on the reordered dataset as $\mathbf{W}_K^*$. Therefore, the global model $\mathbf{W}_K^*$ is fully equivalent to the optimal result derived from centralized joint learning, and its analytical expression can be represented as:

$$\mathbf{W}_K^* = ({\mathbf{F}_{1:K}^*}^\top\mathbf{F}_{1:K}^* + \gamma\mathbf{I})^{-1}{\mathbf{F}_{1:K}^*}^\top\mathbf{Y}_{1:K}^*. \tag{52}$$

By substituting (51) into (52), we can further obtain:

$$\begin{aligned}\mathbf{W}_K^* &= ({\mathbf{F}_{1:K}}^\top\mathbf{\Pi}^\top\mathbf{\Pi}\mathbf{F}_{1:K} + \gamma\mathbf{I})^{-1}{\mathbf{F}_{1:K}}^\top\mathbf{\Pi}^\top\mathbf{\Pi}\mathbf{Y}_{1:K} \\ &= ({\mathbf{F}_{1:K}}^\top\mathbf{\Pi}^{-1}\mathbf{\Pi}\mathbf{F}_{1:K} + \gamma\mathbf{I})^{-1}{\mathbf{F}_{1:K}}^\top\mathbf{\Pi}^\top\mathbf{\Pi}\mathbf{Y}_{1:K} \\ &= ({\mathbf{F}_{1:K}}^\top\mathbf{F}_{1:K} + \gamma\mathbf{I})^{-1}{\mathbf{F}_{1:K}}^\top\mathbf{Y}_{1:K} = \mathbf{W}_K.\end{aligned} \tag{53}$$

It is evident that, regardless of variations in sample ordering, the global model recursively derived by our AFCL is exactly equivalent to the optimal result obtained through centralized joint learning over the complete dataset $\mathcal{D}_{1:K}$. In other words, final global model obtained by our AFCL is independent of spatio-temporal data heterogeneity and client registration order. ■

Notably, the final global model obtained by our AFCL is also independent of the number of (virtual) clients, as long as the complete dataset remains the same. Specifically, let us consider the extreme case with the maximum possible number of clients, where each client holds a unique, non-repetitive sample. Any other cases with fewer clients can be regarded as aggregating all samples from the aforementioned extreme case into the local datasets of a smaller subset of clients, leaving the remaining clients without any data. The transformation of the clients' local datasets can essentially be regarded as a combination of reordering the clients and altering the spatio-temporal data heterogeneity. According to Theorem 2, neither spatio-temporal data heterogeneity nor client registration order impacts the result of our AFCL, implying that results across varying number of clients align precisely with those of the extreme case. Consequently, the final global model obtained by our AFCL remains invariant to the number of clients. Furthermore, in our experimental results presented in Table 1, our AFCL achieves identical performance under settings with 25 or 50 virtual clients (comprising 5 or 10 tasks each distributed across 5 actual clients), thereby further validating the invariance to the number of (virtual) clients.

B. Appendix for Experimental Details

B.1. Details on Datasets & Settings

Here, we detail the dataset partitioning strategy used in our experiments. To emulate realistic FCL scenarios, we used a mixed splitting strategy (***Dirichlet + Si-Blurry***) to model ***both spatial and temporal heterogeneity***. Spatially, we used the standard Dirichlet distribution in FL [38]. Temporally, instead of *disjoint-class splitting* (where new data contains solely new classes), we used *Si-Blurry* (where new data contains mixed classes) [46], as it is more ***practical*** and ***challenging***. Specifically, the entire dataset is first divided into tasks using the Si-Blurry setting, followed by distributing each task's data across multiple clients using the Dirichlet distribution. The detailed process is provided as follows.

First, we adopt the well-established Si-Blurry setting [46] to partition the entire dataset into 5 or 10 tasks. Specifically, in the Si-Blurry setting, all classes are categorized into two groups: disjoint classes (with data exclusive to a specific task) and blurry classes (with data distributed across multiple tasks). The proportion of disjoint classes among the total classes is denoted as r_{D}. Subsequently, a proportion r_{B} of the data from each blurry class is randomly reassigned to other blurry classes. Each task's data consisted of a mix of disjoint and blurry classes, with parameters set as $r_{\mathrm{B}} = 10\%$ and $r_{\mathrm{D}} = 50\%$.

Second, to simulate spatial data heterogeneity, we employ the Dirichlet distribution [38] to divide each task's dataset among 5 clients. Specifically, the data allocated to each client has an equal number of samples and follows the Dirichlet distribution, where the degree of spatial data heterogeneity among clients is controlled by the Dirichlet parameter α. For CIFAR-100 and Tiny-ImageNet, we set $\alpha \in \{0.1, 0.2\}$. Given the higher complexity of ImageNet-R, baselines perform poorly on this dataset, making meaningful comparisons difficult. Therefore, we use $\alpha \in \{0.5, 1.0\}$ to reduce spatial heterogeneity and enhance the baselines' performance. Notably, this adjustment does not affect our AFCL due to our inspiring property of spatio-temporal invariance, but merely brings baseline performance closer to ours for more meaningful comparisons. Moreover, this setup aligns with the experimental settings adopted in existing studies [28, 34, 53, 56, 66].

Since the design of FedMGP is not well suited to a ResNet backbone, directly applying ResNet-18 would be unreasonable. Therefore, for FedMGP, we follow its original paper and use a stronger ViT backbone with a larger parameter scale. Although this stronger backbone may improve the performance of FedMGP, its overestimated results are still markedly worse than those of our AFCL. All of the experiments are conducted on Nvidia RTX 4090D GPUs with 15 vCPUs Intel(R) Xeon(R).

B.2. Details on Evaluation Metrics

Here, we provide a detailed description of the metrics employed in our experiments. For convenience, let $\mathcal{A}_j^i$ denote the accuracy of the global model in the i-round on the test set of the j-th task. The metrics are defined in detail as follows:

- **Average Accuracy** is used to evaluate the overall performance of the method, computed as the average accuracy of the current global model on the test sets of all previously learned tasks. The average accuracy $\mathcal{A}_i$ in the i-round is calculated as:

$$\mathcal{A}_i = \frac{1}{i} \sum\nolimits_{j=1}^{i} \mathcal{A}_j^i \times 100\%. \tag{54}$$

- **Average Knowledge Retention** is used to evaluate the retention of previously learned knowledge as tasks progress, computed as the average ratio of accuracy on each previously learned task's test set between the global model when that task is first learned and the current global model. The average knowledge retention $\mathcal{F}_i$ in the i-round is calculated by:

$$\mathcal{F}_i = \frac{1}{i-1} \sum\nolimits_{j=1}^{i-1} (\mathcal{A}_j^i / \mathcal{A}_j^j) \times 100\%. \tag{55}$$

- **Cumulative Runtime** is used to evaluate the efficiency, computed as the total runtime required to complete training across all tasks, including both client-side local training and server-side global aggregation.