Supplementary Material for the Paper: "Wireless Hybrid Federated Learning for Task-Oriented Semantic Communication Systems"

Haofeng Sun, Wanli Ni, Hui Tian, Jingheng Zheng, Gaofeng Nie, and Dusit Niyato, *Fellow, IEEE*

APPENDIX A
PROOF OF LEMMA 1

Plugging $\boldsymbol{e}_{\boldsymbol{\phi},t,1}$, $\boldsymbol{e}_{\boldsymbol{\phi},t,2}$, $\boldsymbol{e}_{\boldsymbol{\phi},t,3}$ into (35), and $\boldsymbol{e}_{\boldsymbol{\theta},t,1}$, $\boldsymbol{e}_{\boldsymbol{\theta},t,2}$, $\boldsymbol{e}_{\boldsymbol{\theta},t,3}$ into (36), we can obtain

$$\boldsymbol{e}_{\boldsymbol{\phi},t} = \nabla_{\boldsymbol{\phi}} F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t) - \hat{\mathbf{g}}^F_{\boldsymbol{\phi},t}, \tag{65}$$

$$\begin{aligned}
\boldsymbol{e}_{\boldsymbol{\theta},t} &= \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t) - \sum_{k\in\mathcal{K}} \frac{D_k}{D} \sum_{l\in\boldsymbol{\xi}_{t,k}} \frac{\mathbf{g}_{\boldsymbol{\theta},t,k,l}}{\xi} \\
&\quad + \frac{\xi_{c,t}}{\xi}\left(\sum_{k\in\mathcal{K}} \frac{D_k}{D} \sum_{l\in\boldsymbol{\xi}_{c,t,k}} \frac{\mathbf{g}_{\boldsymbol{\theta},t,k,l}}{\xi_{c,t}} - \mathbf{g}^C_{\boldsymbol{\theta},t}\right) \\
&\quad + \frac{\xi_{f,t}}{\xi}\left(\sum_{k\in\mathcal{K}} \frac{D_k}{D} \mathbf{g}^F_{\boldsymbol{\theta},t,k} - \hat{\mathbf{g}}^F_{\boldsymbol{\theta},t}\right) \\
&\overset{(a)}{=} \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t) - \frac{\left(\xi_{f,t}\hat{\mathbf{g}}^F_{\boldsymbol{\theta},t} + \xi_{c,t}\mathbf{g}^C_{\boldsymbol{\theta},t}\right)}{\xi} \\
&\quad - \sum_{k\in\mathcal{K}} \frac{D_k}{D} \frac{\sum_{l\in\boldsymbol{\xi}_{t,k}} \mathbf{g}_{\boldsymbol{\theta},t,k,l}}{\xi} \\
&\quad + \sum_{k\in\mathcal{K}} \frac{D_k}{D} \frac{\sum_{l\in\boldsymbol{\xi}_{c,t,k}} \mathbf{g}_{\boldsymbol{\theta},t,k,l} + \sum_{l\in\boldsymbol{\xi}_{f,t,k}} \mathbf{g}_{\boldsymbol{\theta},t,k,l}}{\xi} \\
&\overset{(b)}{=} \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t) - \frac{\left(\xi_{f,t}\hat{\mathbf{g}}^F_{\boldsymbol{\theta},t} + \xi_{c,t}\mathbf{g}^C_{\boldsymbol{\theta},t}\right)}{\xi},
\end{aligned} \tag{66}$$

where (a) comes from the definition of $\mathbf{g}^C_{\boldsymbol{\theta},t}$ and $\mathbf{g}^F_{\boldsymbol{\theta},t,k}$, (b) comes from the definition of $\boldsymbol{\xi}_{t,k}$, $\boldsymbol{\xi}_{c,t,k}$ and $\boldsymbol{\xi}_{f,t,k}$. Plugging (65) and (66) into (9), the model update can be written as (34). This completes the proof.

APPENDIX B
PROOF OF THEOREM 1

From Assumption 1, due to the L-Lipschitz continuous of $\nabla_{\boldsymbol{\phi}} F(\boldsymbol{\phi},\boldsymbol{\theta})$ with $\boldsymbol{\phi}$ and $\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\phi},\boldsymbol{\theta})$ with $\boldsymbol{\theta}$, we have

$$\begin{aligned}
F(\boldsymbol{\phi}_{t+1},\boldsymbol{\theta}_{t+1}) &\le F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_{t+1}) + \frac{L}{2}\|\boldsymbol{\phi}_{t+1}-\boldsymbol{\phi}_t\|^2 \\
&\quad + (\boldsymbol{\phi}_{t+1}-\boldsymbol{\phi}_t)^T \nabla_{\boldsymbol{\phi}} F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_{t+1}),
\end{aligned} \tag{67}$$

$$\begin{aligned}
F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_{t+1}) &\le F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t) + \frac{L}{2}\|\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t\|^2 \\
&\quad + (\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t)^T \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t).
\end{aligned} \tag{68}$$

Mix (67) and (68) on both left and right sides, we can obtain

$$\begin{aligned}
F(\boldsymbol{\phi}_{t+1},\boldsymbol{\theta}_{t+1}) \le\ & F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t) + (\boldsymbol{\phi}_{t+1}-\boldsymbol{\phi}_t)^T \nabla_{\boldsymbol{\phi}} F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_{t+1}) \\
& + (\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t)^T \nabla_{\boldsymbol{\theta}} F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t) \\
& + \frac{L}{2}\|\boldsymbol{\phi}_{t+1}-\boldsymbol{\phi}_t\|^2 + \frac{L}{2}\|\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t\|^2,
\end{aligned} \tag{69}$$

where $(\boldsymbol{\phi}_{t+1}-\boldsymbol{\phi}_t)^T \nabla_{\boldsymbol{\phi}} F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_{t+1})$ can be reformulated as

$$\begin{aligned}
(\boldsymbol{\phi}_{t+1}-\boldsymbol{\phi}_t)^T \nabla_{\boldsymbol{\phi}} F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_{t+1}) &= (\boldsymbol{\phi}_{t+1}-\boldsymbol{\phi}_t)^T \nabla_{\boldsymbol{\phi}} F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t) \\
&\quad + \underbrace{(\boldsymbol{\phi}_{t+1}-\boldsymbol{\phi}_t)^T \left(\nabla_{\boldsymbol{\phi}} F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_{t+1}) - \nabla_{\boldsymbol{\phi}} F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t)\right)}_{\boldsymbol{c}_1}.
\end{aligned} \tag{70}$$

In the continue, $\boldsymbol{c}_1$ can be bounded as

$$\begin{aligned}
\boldsymbol{c}_1 &\overset{(a)}{\le} \|\nabla_{\boldsymbol{\phi}} F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_{t+1}) - \nabla_{\boldsymbol{\phi}} F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t)\| \left\|\boldsymbol{\phi}_{t+1}-\boldsymbol{\phi}_t\right\| \\
&\overset{(b)}{\le} L\left\|\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t\right\| \left\|\boldsymbol{\phi}_{t+1}-\boldsymbol{\phi}_t\right\| \\
&\overset{(c)}{\le} \frac{L}{2}\left\|\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t\right\|^2 + \frac{L}{2}\left\|\boldsymbol{\phi}_{t+1}-\boldsymbol{\phi}_t\right\|^2,
\end{aligned} \tag{71}$$

where (a) follows from the Cauthy-Schwarz inequality, (b) can be gotten by substituting Assumption 1 and (c) comes from the triangle inequality.

Plugging (71) into (69), while setting $\eta = \frac{1}{2L}$, we can achieve

$$\begin{aligned}
F(\boldsymbol{\phi}_{t+1},\boldsymbol{\theta}_{t+1}) &\overset{(a)}{\le} F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t) + \frac{1}{4L}\Big(-\|\nabla_{\boldsymbol{\phi}} F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t)\|^2 \\
&\quad - \|\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t)\|^2 + \|\boldsymbol{e}_{\boldsymbol{\phi},t}\|^2 + \|\boldsymbol{e}_{\boldsymbol{\theta},t}\|^2\Big).
\end{aligned} \tag{72}$$

where (a) comes from plugging (34). The expectation of $\left\|\boldsymbol{e}_{\boldsymbol{\phi},t}\right\|^2$ and $\left\|\boldsymbol{e}_{\boldsymbol{\theta},t}\right\|^2$ are reformulated in Lemma 5. With Lemma 5, we can directly obtain (37). This completes the proof.

Lemma 5: The expectation of $\left\|\boldsymbol{e}_{\boldsymbol{\phi},t}\right\|^2$ and $\left\|\boldsymbol{e}_{\boldsymbol{\theta},t}\right\|^2$ are given respectively by

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{e}_{\boldsymbol{\phi},t}\|^2\right] &= \mathbb{E}\left[\|\boldsymbol{e}_{\boldsymbol{\phi},t,1}\|^2\right] + \mathbb{E}\left[\|\boldsymbol{e}_{\boldsymbol{\phi},t,2}\|^2\right] \\
&\quad + \mathbb{E}\left[\|\boldsymbol{e}_{\boldsymbol{\phi},t,3}\|^2\right],
\end{aligned} \tag{73}$$

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{e}_{\boldsymbol{\theta},t}\|^2\right] &= \mathbb{E}\left[\|\boldsymbol{e}_{\boldsymbol{\theta},t,1}\|^2\right] + \frac{\xi^2_{c,t}}{\xi^2}\mathbb{E}\left[\|\boldsymbol{e}_{\boldsymbol{\theta},t,2}\|^2\right] \\
&\quad + \frac{\xi^2_{f,t}}{\xi^2}\mathbb{E}\left[\|\boldsymbol{e}_{\boldsymbol{\theta},t,3}\|^2\right].
\end{aligned} \tag{74}$$

Proof: See Appendix C. ∎

H. Sun, H. Tian, J. Zheng and G. Nie are with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: sunhaofeng@bupt.edu.cn; tianhui@bupt.edu.cn; zhengjh@bupt.edu.cn; niegaofeng@bupt.edu.cn).

W. Ni is with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: niwanli@tsinghua.edu.cn).

D. Niyato is with College of Computing and Data Science, Nanyang Technological University, Singapore 117583 (e-mail: dniyato@ntu.edu.sg).

Appendix C
Proof of Lemma 5

By expanding $\|\boldsymbol{e}_\phi\|^2$ and $\|\boldsymbol{e}_\theta\|^2$, we have

$$\begin{aligned}\|\boldsymbol{e}_{\boldsymbol{\phi},t}\|^2 =&\|\boldsymbol{e}_{\boldsymbol{\phi},t,1}+\boldsymbol{e}_{\boldsymbol{\phi},t,2}+\boldsymbol{e}_{\boldsymbol{\phi},t,3}\|^2\\ =&\|\boldsymbol{e}_{\boldsymbol{\phi},t,1}\|^2+\|\boldsymbol{e}_{\boldsymbol{\phi},t,2}\|^2+\|\boldsymbol{e}_{\boldsymbol{\phi},t,3}\|^2\\ &+2\boldsymbol{e}_{\boldsymbol{\phi},t,1}^H(\boldsymbol{e}_{\boldsymbol{\phi},t,2}+\boldsymbol{e}_{\boldsymbol{\phi},t,3})+2\boldsymbol{e}_{\boldsymbol{\phi},t,2}^H\boldsymbol{e}_{\boldsymbol{\phi},t,3}\\ \overset{(a)}{=}&\|\boldsymbol{e}_{\boldsymbol{\phi},t,1}\|^2+\|\boldsymbol{e}_{\boldsymbol{\phi},t,2}\|^2+\|\boldsymbol{e}_{\boldsymbol{\phi},t,3}\|^2,\end{aligned} \tag{75}$$

where (a) comes from (32) in Assumption 3 and $\mathbb{E}\left[\hat{\mathbf{g}}_{\boldsymbol{\phi},t}^F\right]=\mathbb{E}\left[\sum_{k\in\mathcal{K}}\frac{D_k}{D}\mathbf{g}_{\boldsymbol{\phi},t,k}^F\right]$. Similarly, we can obtain

$$\|\boldsymbol{e}_{\boldsymbol{\theta},t}\|^2=\|\boldsymbol{e}_{\boldsymbol{\phi},t,1}\|^2+\frac{\xi_{c,t}^2}{\xi^2}\|\boldsymbol{e}_{\boldsymbol{\phi},t,2}\|^2+\frac{\xi_{f,t}^2}{\xi^2}\|\boldsymbol{e}_{\boldsymbol{\phi},t,3}\|^2. \tag{76}$$

The proof completes.

Appendix D
Proof of Lemma 2

Considering the fact that $\mathbf{g}_{t,k,l}=[\mathbf{g}_{\boldsymbol{\phi},t,k,l},\mathbf{g}_{\boldsymbol{\theta},t,k,l}]$, we have $\boldsymbol{e}_{t,1}=[\boldsymbol{e}_{\boldsymbol{\phi},t,1},\boldsymbol{e}_{\boldsymbol{\theta},t,1}]$. Based on Assumption 3, $\mathbb{E}\left[\|\boldsymbol{e}_{\boldsymbol{\phi},t,1}\|^2+\|\boldsymbol{e}_{\boldsymbol{\theta},t,1}\|^2\right]$ can be bounded by

$$\begin{aligned}&\mathbb{E}\left[\|\boldsymbol{e}_{t,1}\|^2\right]=\mathbb{E}\left[\|\boldsymbol{e}_{\boldsymbol{\phi},t,1}\|^2+\|\boldsymbol{e}_{\boldsymbol{\theta},t,1}\|^2\right]\\ &=\mathbb{E}\left[\left\|\sum_{k\in\mathcal{K}}\frac{D_k}{D}\sum_{l\in\boldsymbol{\xi}_{t,k}}\frac{\mathbf{g}_{t,k,l}}{\xi}-\sum_{k\in\mathcal{K}}\frac{D_k}{D}\sum_{l\in\mathcal{D}_k}\frac{\mathbf{g}_{t,k,l}}{D_k}\right\|^2\right]\\ &=\mathbb{E}\left[\left\|\sum_{k\in\mathcal{K}}\frac{D_k}{D}\left(\sum_{l\in\boldsymbol{\xi}_{t,k}}\frac{\mathbf{g}_{t,k,l}}{\xi}-\sum_{l\in\mathcal{D}_k}\frac{\mathbf{g}_{t,k,l}}{D_k}\right)\right\|^2\right]\\ &\overset{(a)}{\le}\mathbb{E}\left[\left\|\sum_{k\in\mathcal{K}}\frac{D_k}{D}\left\|\sum_{l\in\boldsymbol{\xi}_{t,k}}\frac{\mathbf{g}_{t,k,l}}{\xi}-\sum_{l\in\mathcal{D}_k}\frac{\mathbf{g}_{t,k,l}}{D_k}\right\|\right\|^2\right]\\ &\overset{(b)}{\le}\mathbb{E}\left[\left\|\sum\nolimits_{k\in\mathcal{K}}\frac{D_k}{D}\sigma\right\|^2\right]\\ &=\sigma^2.\end{aligned} \tag{77}$$

where (a) comes from the triangle-inequality, and (b) comes from (33) in Assumption 3.

In the continue, based on [1], the expectation of $\|\boldsymbol{e}_{\boldsymbol{\phi},t,2}\|^2$ can be bounded as follows:

$$\begin{aligned}&\mathbb{E}\left[\|\boldsymbol{e}_{\boldsymbol{\phi},t,2}\|^2\right]=\mathbb{E}\left[\left\|\sum_{k\in\mathcal{K}}\frac{D_k}{D}\left(\sum_{l\in\boldsymbol{\xi}_{t,k}}\frac{\mathbf{g}_{\boldsymbol{\phi},t,k,l}}{\xi}-\mathbf{g}_{\boldsymbol{\phi},t,k}^F\right)\right\|^2\right]\\ &\overset{(a)}{\le}\mathbb{E}\left[\left(\sum_{k\in\mathcal{K}}\frac{D_k}{D}\left\|\sum_{l\in\boldsymbol{\xi}_{t,k}}\frac{\mathbf{g}_{\boldsymbol{\phi},t,k,l}}{\xi}-\sum_{l\in\boldsymbol{\xi}_{f,t,k}}\frac{\mathbf{g}_{\boldsymbol{\phi},t,k,l}}{\xi_{f,t}}\right\|\right)^2\right]\\ &=\mathbb{E}\left[\left(\sum_{k\in\mathcal{K}}\frac{D_k}{D}\left\|\sum_{l\in\boldsymbol{\xi}_{c,t,k}}\frac{\mathbf{g}_{\boldsymbol{\phi},t,k,l}}{\xi}-\sum_{l\in\boldsymbol{\xi}_{f,t,k}}\frac{\xi_{c,t}\mathbf{g}_{\boldsymbol{\phi},t,k,l}}{\xi_{f,t}\xi}\right\|\right)^2\right]\\ &\overset{(b)}{\le}\mathbb{E}\left[\left(\sum_{k\in\mathcal{K}}\frac{D_k}{D}\left\|\sum_{l\in\boldsymbol{\xi}_{c,t,k}}\frac{\mathbf{g}_{\boldsymbol{\phi},t,k,l}}{\xi}\right\|+\left\|\sum_{l\in\boldsymbol{\xi}_{f,t,k}}\frac{\xi_{c,t}\mathbf{g}_{\boldsymbol{\phi},t,k,l}}{\xi_{f,t}\xi}\right\|\right)^2\right]\\ &\overset{(c)}{\le}\mathbb{E}\left[\left(\sum_{k\in\mathcal{K}}\frac{D_k}{D}\sum_{l\in\boldsymbol{\xi}_{c,t,k}}\left\|\frac{\mathbf{g}_{\boldsymbol{\phi},t,k,l}}{\xi}\right\|+\sum_{l\in\boldsymbol{\xi}_{f,t,k}}\left\|\frac{\xi_{c,t}\mathbf{g}_{\boldsymbol{\phi},t,k,l}}{\xi_{f,t}\xi}\right\|\right)^2\right]\\ &\overset{(d)}{\le}\mathbb{E}\left[\left(\sum_{k\in\mathcal{K}}\frac{D_k}{D}\frac{2\xi_{c,t}}{\xi}\sqrt{\left(\beta_1+\beta_2\|\nabla_{\boldsymbol{\phi}}F(\boldsymbol{\phi},\boldsymbol{\theta})\|^2\right)}\right)^2\right]\\ &=\frac{4\xi_{c,t}^2}{\xi^2}\left(\beta_1+\beta_2\|\nabla_{\boldsymbol{\phi}}F(\boldsymbol{\phi},\boldsymbol{\theta})\|^2\right).\end{aligned} \tag{78}$$

where (a), (b) and (c) follow from the triangle-inequality, (d) can be easily obtained by plugging (31).

Considering the fact that the SNR (21) of semantic symbols transmitted to the BS is equal to the SNR γ_{SC} for local training, the centralized SR model training and local SR model training are equivalent. Thus, the ℓ_2 norm expectation of training error $\boldsymbol{e}_{\boldsymbol{\theta},t,2}$ with respect to $\boldsymbol{\theta}$ caused by the aggregation method is

$$\begin{aligned}&\mathbb{E}\left[\|\boldsymbol{e}_{\boldsymbol{\theta},t,2}\|^2\right]=\mathbb{E}\left[\left\|\sum_{k\in\mathcal{K}}\frac{D_k}{D}\sum_{l\in\boldsymbol{\xi}_{c,t,k}}\frac{\mathbf{g}_{\boldsymbol{\theta},t,k,l}}{\xi_{c,t}}-\mathbf{g}_{\boldsymbol{\theta},t}^C\right\|^2\right]\\ &=\mathbb{E}\left[\left\|\sum_{k\in\mathcal{K}}\frac{D_k}{D}\left(\sum_{l\in\boldsymbol{\xi}_{c,t,k}}\frac{\mathbf{g}_{\boldsymbol{\theta},t,k,l}}{\xi_{c,t}}-\sum_{l\in\boldsymbol{\xi}_{c,t,k}}\frac{\mathbf{g}_{\boldsymbol{\theta},t,k,l}}{\xi_{c,t}}\right)\right\|^2\right]\\ &=0.\end{aligned} \tag{79}$$

Based on the $\mathrm{MSE}_{t,n}$ in (18), the expected upper bound of $\|\boldsymbol{e}_{\boldsymbol{\phi},t,3}\|^2$ can be derived by

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{e}_{\boldsymbol{\phi},t,3}\|^2\right]&=\mathbb{E}\left[\left\|\mathbf{g}_{\boldsymbol{\phi},t}^F-\hat{\mathbf{g}}_{\boldsymbol{\phi},t}^F\right\|^2\right]\\ &=\mathbb{E}\left[\sum\nolimits_{q=1}^{Q}\left(g_{\boldsymbol{\phi},t,q}^F-\hat{g}_{\boldsymbol{\phi},t,q}^F\right)^2\right]\\ &=\tilde{\sigma}_t^2\frac{Q}{A_t}\sum\nolimits_{n=1}^{N}\alpha_{t,n}\mathrm{MSE}_{t,n}\\ &\le\tilde{\sigma}_t^2\frac{Q}{A_t}\max_n\{\alpha_{t,n}\mathrm{MSE}_{t,n}\}\sum\nolimits_{n=1}^{N}\alpha_{t,n}\\ &=\tilde{\sigma}_t^2Q\max_n\{\alpha_{t,n}\mathrm{MSE}_{t,n}\}.\end{aligned} \tag{80}$$

Similarly, we can obtain

$$\mathbb{E}\left[\|\boldsymbol{e}_{\boldsymbol{\theta},t,3}\|^2\right]=\tilde{\sigma}_t^2R\max_n\{\alpha_{t,n}\mathrm{MSE}_{t,n}\}. \tag{81}$$

The proof completes.

Appendix E
Proof of Theorem 2

The expected upper bound of the ℓ_2 norm of the global gradients with respect to $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ can be obtained by substituting $\lambda=\min_t\{\lambda_t\}$, which can by expressed as

$$\begin{aligned}&\mathbb{E}\left[\|\nabla_{\boldsymbol{\phi}}F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t)\|^2+\frac{1}{\lambda}\|\nabla_{\boldsymbol{\theta}}F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t)\|^2\right]\\ &\le\frac{4L}{\lambda}\mathbb{E}\left[F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t)-F(\boldsymbol{\phi}_{t+1},\boldsymbol{\theta}_{t+1})+\frac{\sigma^2}{4L}+\frac{\Lambda_t}{4L}\right].\end{aligned} \tag{82}$$

The upper bound of the average ℓ_2 norm of global gradients in Hybrid-FL can be obtained by summing up (82) from $t=1$ to $t=T$ and dividing it by T, as illustrated in (44). This completes the proof.

Appendix F
Proof of Lemma 4

Based on Assumption 1 and Theorem 1, while setting the learning rate $\eta = \frac{1}{2L}$. In the t-th training round, the convergence behavior of FL method can be characterized as

$$\begin{aligned}\mathbb{E}\left[F(\boldsymbol{\phi}_{t+1},\boldsymbol{\theta}_{t+1})\right] \leq\ & \mathbb{E}\Bigg[F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t) - \frac{1}{4L}\Big(\|\nabla_{\boldsymbol{\phi}}F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t)\|^2 \\ & + \|\nabla_{\boldsymbol{\theta}}F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t)\|^2 - \|\boldsymbol{e}_{\boldsymbol{\phi},t,1}\|^2 - \|\boldsymbol{e}_{\boldsymbol{\phi},t,2}\|^2 - \|\boldsymbol{e}_{\boldsymbol{\phi},t,3}\|^2 \\ & - \|\boldsymbol{e}_{\boldsymbol{\theta},t,1,\mathrm{FL}}\|^2 - \|\boldsymbol{e}_{\boldsymbol{\theta},t,2,\mathrm{FL}}\|^2 - \|\boldsymbol{e}_{\boldsymbol{\theta},t,3,\mathrm{FL}}\|^2\Big)\Bigg]. \end{aligned} \tag{83}$$

Similar to Lemma 2, the expected ℓ_2 norm upper bound of $\boldsymbol{e}_{\boldsymbol{\theta},t,2,\mathrm{FL}}$ and $\boldsymbol{e}_{\boldsymbol{\theta},t,3,\mathrm{FL}}$ can be respectively given by

$$\mathbb{E}\left[\|\boldsymbol{e}_{\boldsymbol{\theta},t,2,FL}\|^2\right] \leq \frac{4\xi_{c,t}^2}{\xi^2}\left(\beta_1 + \beta_2\|\nabla_{\boldsymbol{\theta}}F(\boldsymbol{\phi},\boldsymbol{\theta})\|^2\right), \tag{84}$$

$$\mathbb{E}\left[\|\boldsymbol{e}_{\boldsymbol{\theta},t,3,\mathrm{FL}}\|^2\right] \leq \tilde{\sigma}_t^2 R \max_n\{\alpha_{t,n}\mathrm{MSE}_{t,n}\}. \tag{85}$$

Plugging (38), (40), (42), (84), (85) into (83), we finally obtain (47). Denote $\lambda = \min_t\{\lambda_{t,\mathrm{FL}}\}$, the ℓ_2 norm upper bound of the global gradients in FL method with respect to $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ can be given by

$$\begin{aligned}&\mathbb{E}\left[\|\nabla_{\boldsymbol{\phi}}F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t)\|^2 + \|\nabla_{\boldsymbol{\theta}}F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t)\|^2\right] \\ &\quad\leq \frac{4L}{\lambda}\mathbb{E}\left[F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t) - F(\boldsymbol{\phi}_{t+1},\boldsymbol{\theta}_{t+1}) + \frac{\sigma^2}{4L} + \frac{\Lambda_{t,\mathrm{FL}}}{4L}\right].\end{aligned} \tag{86}$$

Due to the split structure of ST and SR models, the joint global gradient $\nabla F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t)$ can be decomposed into $\nabla_{\boldsymbol{\phi}}F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t)$ and $\nabla_{\boldsymbol{\theta}}F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t)$. This decomposition can be formulated as $\nabla F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t) = [\nabla_{\boldsymbol{\phi}}F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t), \nabla_{\boldsymbol{\theta}}F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t)]$. Consequently, we have $\|\nabla F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t)\|^2 = \|[\nabla_{\boldsymbol{\phi}}F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t), \nabla_{\boldsymbol{\theta}}F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t)]\|^2$, which implies that $\|\nabla F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t)\|^2 = \|\nabla_{\boldsymbol{\phi}}F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t)\|^2 + \|\nabla_{\boldsymbol{\theta}}F(\boldsymbol{\phi}_t,\boldsymbol{\theta}_t)\|^2$. The average upper bound of the ℓ_2 norm of global gradients (48) in FL method can be obtained by summing the term in (86) from $t = 1$ to $t = T$. The proof completes.

Appendix G
Proof of Proposition 1

Firstly, denote $\{\{\vartheta_{t,k}\}|\{P_{t,n}^{\mathrm{AirComp}}\}\}$ as the feasible region of $\{\vartheta_{t,k}\}$ with variables $\{P_{t,n}^{\mathrm{AirComp}}\}$. Denote $\{\{\vartheta_{t,k}\}|\{P_{t,n}^{\mathrm{AirComp}*}\}\}$ as the feasible region of $\{\vartheta_{t,k}\}$ with fixed $\{P_{t,n}^{\mathrm{AirComp}*}\}$ in (59). It is worth noting that the original constraints (49c), which have been reformulated as (58), allow us to obtain the minimum AirComp power coefficients based on equation (58).

From (59), we can find the co-existence of $\{P_{t,n}^{\mathrm{AirComp}}\}$ and $\{\vartheta_{t,k}\}$ only lies in constraint (49h). Moreover, as the value of $\{P_{t,n}^{\mathrm{AirComp}}\}$ increases, the energy $E_{t,k}^{U}$ utilized for transmission also increases, potentially resulting in a decrease in the energy $E_{t,k}^{C}$ used for computation and a reduction for the feasible region of $\{\vartheta_{t,k}\}$. Thus, easily acquired, due to the minimal property of $\{P_{t,n}^{\mathrm{AirComp}*}\}$ and decoupled form of $\{P_{t,n}^{\mathrm{AirComp}}\}$ and $\{\vartheta_{t,k}\}$ in (49h), $\forall P_{t,n}^{\mathrm{AirComp}} \geq P_{t,n}^{\mathrm{AirComp}*}$, $\{\{\vartheta_{t,k}\}|\{P_{t,n}^{\mathrm{AirComp}}\}\} \in \{\{\vartheta_{t,k}\}|\{P_{t,n}^{\mathrm{AirComp}*}\}\}$. Thus, the feasible region of $\{\vartheta_{t,k}\}$ remains unchanged with fixed AirComp power coefficients given in (60). Furthermore, by considering the fixed AirComp power coefficients in (60), the min-max problem (59) can be transformed into a standard GP problem (61). Notably, problem (61) exhibits convexity with respect to $\tau_{t,\max}^{C}$ and $\{\vartheta_{t,k}\}$, allowing us to employ disciplined GP [2] for obtaining the optimal CPU frequency.

References

[1] M. P. Friedlander and M. Schmidt, "Hybrid deterministic-stochastic methods for data fitting," *SIAM J. Sci. Comput.*, vol. 34, no. 3, pp. A1380–A1405, 2012.

[2] A. Agrawal, S. Diamond, and S. Boyd, "Disciplined geometric programming," *Optim. Lett.*, vol. 13, no. 5, pp. 961–976, Jul. 2019.