I. SUPPLEMENTARY MATERIAL

A. *Proof of Theorem 1*

To enhance the clarity of our proof, we begin by presenting a comprehensive table of notations, as outlined in Table I-A.

TABLE I
NOTATIONS.

Notation	Definition
M	Number of horizontal silos.
N	Number of total samples, $N = \sum_{m=1}^{M} N_m$.
N_m	Number of samples within the m-th silo, $m \in [M]$.
J	Number of IoT sensors (modalities), $J \geq K$.
K	Number of IoT devices (vertical parties). When $K = J$, each device has one sensor.
$\mathbf{x}, \mathbf{y}$	Full dataset across all M silos, $\{\mathbf{x}_m, \mathbf{y}_m\}_{m=1}^{M}$.
$\mathbf{x}_m, \mathbf{y}_m$	m-th silo local (partial) dataset, $\{\mathbf{x}_m^k, \mathbf{y}_m\}_{k=1}^{K}$.
$\mathbf{x}_m^k$	k-th vertical party's data within the m-th silo.
$\mathcal{B}_m$	Randomly sampled mini-batch of size B_m.
η	Learning rate.
Q	VFL communication frequency (Q iterations).
RQ	HFL communication frequency (RQ iterations).
P	The number of total global rounds.
T	The number of total local iterations, $T = RQ \times P$.
Θ	Global model, $[\theta^0, \theta^1, \ldots, \theta^k, \ldots, \theta^K]$.
θ^k	k-th vertical party model, $\frac{1}{N}\sum_{m=1}^{M} N_m \theta_m^k$.
Θ_m	m-th silo model, $[\theta_m^0, \theta_m^1, \ldots, \theta_m^k, \ldots, \theta^K]$.
θ_m^0	m-th silo edge server model (head), $k = 0$.
θ_m^k	m-th silo k-th vertical party model.
h_m^k	m-th silo k-th vertical party embedding function.
$\Phi_m^{t_0}$	The set of embeddings that each party would receive at iteration t_0 within m-th silo.
Φ_m^{-k,t_0}	The set of embeddings from other parties $k' \neq k$ within m-th silo **(stale information at iteration** t).
$\Phi_m^{k,t}$	The set of embeddings used by k-th party within m-th silo, $\{\Phi_m^{-k,t_0}; h_m^k(\theta_m^{k,t}; \mathbf{x}_m^{k,\mathcal{B}_m^{t_0}})\}$.
$f(\Theta; \mathbf{x}; \mathbf{y})$	Global objective, $\frac{1}{N}\sum_{m=1}^{M} N_m f_m(\Theta)$.
$f_m(\Theta_m; \mathbf{x}_m; \mathbf{y}_m)$	m-th silo objective, $\frac{1}{N_m}\sum_{i=1}^{N_m} \mathcal{L}\left[\Theta_m; \mathbf{x}_m^i; y_m^i\right]$.
$\boldsymbol{G}^t$	The gradient at iteration t, $[(\boldsymbol{G}^{0,t})^\top, \cdots, (\boldsymbol{G}^{k,t})^\top, \cdots, (\boldsymbol{G}^{K,t})^\top]^\top$.
$\boldsymbol{G}^{k,t}$	The partial gradient at iteration t, $\frac{1}{N}\sum_{m=1}^{M} N_m \boldsymbol{G}_m^{k,t}$.
$\boldsymbol{G}_m^{k,t}$	The stochastic partial derivative at iteration t, $\nabla_k f_m(\theta_m^{0,t_0}, \theta_m^{1,t_0}, \cdots, \theta_m^{k,t}, \cdots, \theta_m^{K,t_0}; \mathcal{B}_m)$.
$\nabla_k f_m(\Theta_m)$	The partial derivative associated the coordinate partition θ_m^k within m-th silo.
$\nabla_k f_m(\Theta_m; \mathcal{B}_m)$	The stochastic partial derivative of the coordinate partition θ_m^k within m-th silo.
$\Gamma_m^{k,t}$	k-th party's view of the m-th silo-level model at iteration t, $[\theta_m^{0,t_0}, \theta_m^{1,t_0}, \cdots, \theta_m^{k,t}, \cdots, \theta_m^{K,t_0}]$.
$f_{\mathcal{B}_m}(\Gamma_m^{k,t})$	The stochastic loss for mini-batch $\mathcal{B}_m$ calculated by k-th party at iteration t within m-th silo.
L, L_k	Smoothness.
$\frac{\delta^2}{K}$	The expected squared euclidean norm of $\nabla_k f_m(\Theta; \mathcal{B}_m)$ is uniformly bounded.
$\frac{\sigma_k^2}{B}$	The variance of the stochastic partial derivatives.

We suppose HFM runs for a total of P global rounds, i.e., a total of $T = RQ \times P$ iterations. To offer a clearer illustration in the subsequent proof, we *tentatively* define the "virtual" global model to be updated as follows:

$$\Theta^{t+1} = \Theta^t - \eta \boldsymbol{G}^t. \tag{1}$$

Here, the definition of the "virtual" global model illustrates the ideal scenario where the global model is updated every iteration, although in practice, it is updated every RQ iterations. And, $\boldsymbol{G}^t$ denotes the gradient for the global model,

$$\begin{aligned}
\boldsymbol{G}^t &= \left[\left(\boldsymbol{G}^{0,t}\right)^\top, \left(\boldsymbol{G}^{1,t}\right)^\top, \cdots, \left(\boldsymbol{G}^{k,t}\right)^\top, \cdots, \left(\boldsymbol{G}^{K,t}\right)^\top\right]^\top \\
&= \begin{bmatrix} \frac{1}{N}\sum_{m=1}^{M} N_m \boldsymbol{G}_m^{0,t} \\ \frac{1}{N}\sum_{m=1}^{M} N_m \boldsymbol{G}_m^{1,t} \\ \cdots \\ \frac{1}{N}\sum_{m=1}^{M} N_m \boldsymbol{G}_m^{k,t} \\ \cdots \\ \frac{1}{N}\sum_{m=1}^{M} N_m \boldsymbol{G}_m^{K,t} \end{bmatrix},
\end{aligned} \tag{2}$$

where $\boldsymbol{G}_m^{k,t} = \nabla_k f_m\left(\theta_m^{0,t_0}, \theta_m^{1,t_0}, \cdots, \theta_m^{k,t}, \cdots, \theta_m^{K,t_0}; \mathcal{B}_m^{t_0}\right)$. Based on Assumption 1 and $N = \sum_{m=1}^{M} N_m$, we have

$$\begin{aligned}
&\left\|\nabla f(\Theta) - \nabla f(\Theta')\right\| \\
&= \left\|\nabla\Big(\frac{1}{N}\sum_{m=1}^{M} N_m f_m(\Theta)\Big) - \nabla\Big(\frac{1}{N}\sum_{m=1}^{M} N_m f_m(\Theta')\Big)\right\| \\
&= \left\|\frac{1}{N}\sum_{m=1}^{M} N_m \nabla f_m(\Theta) - \frac{1}{N}\sum_{m=1}^{M} N_m \nabla f_m(\Theta')\right\| && (3) \\
&= \frac{1}{N}\sum_{m=1}^{M} N_m \left\|\nabla f_m\left(\Theta\right) - \nabla f_m\left(\Theta'\right)\right\| && (4) \\
&\leq L\left\|\Theta - \Theta'\right\|. && (5)
\end{aligned}$$

Now, we can begin the proof of Theorem 1.

Proof.

$$\begin{aligned}
&\mathbb{E}\left[f\left(\Theta^{t+1}\right)\right] \\
&= \mathbb{E}\left[f\left(\Theta^t - \eta \nabla f\left(\Theta^t\right)\right)\right] && (6) \\
&\leq f\left(\Theta^t\right) - \mathbb{E}\left[\left\langle \nabla f\left(\Theta^t\right), \Theta^{t+1} - \Theta^t\right\rangle + \frac{L}{2}\left\|\Theta^{t+1} - \Theta^t\right\|^2\right] && (7) \\
&\leq f\left(\Theta^t\right) - \sum_{k=0}^{K} \mathbb{E}\left\langle \nabla_k f\left(\Theta^t\right), \eta \boldsymbol{G}^{k,t}\right\rangle + \frac{L\eta^2}{2}\sum_{k=0}^{K} \mathbb{E}\left\|\boldsymbol{G}^{k,t}\right\|^2. && (8)
\end{aligned}$$

Next, based on Assumption 2 and Assumption 3, we can bound

$$\begin{aligned}
&\mathbb{E}\left\|\boldsymbol{G}^{k,t} - \mathbb{E}\left[\boldsymbol{G}^{k,t}\right]\right\|^2 \\
&= \mathbb{E}\Bigg\|\frac{1}{N}\sum_{m=1}^{M} N_m \nabla_k f_m\left(\Gamma_m^{k,t}; \mathcal{B}_m^{t_0}\right) \\
&\quad - \frac{1}{N}\sum_{m=1}^{M} N_m \mathbb{E}\Big[\nabla_k f_m(\Gamma_m^{k,t}; \mathcal{B}_m^{t_0})\Big]\Bigg\|^2 && (9) \\
&= \frac{1}{N^2}\mathbb{E}\Big[\sum_{m=1}^{M} N_m^2 \left\|\nabla_k f_m(\Gamma_m^{k,t}; \mathcal{B}_m^{t_0}) - \nabla_k f_m(\Gamma_m^{k,t})\right\|^2\Big] && (10) \\
&= \frac{1}{N^2}\sum_{m=1}^{M} N_m^2 \mathbb{E}\Big[\left\|\nabla_k f_m(\Gamma_m^{k,t}; \mathcal{B}_m^{t_0}) - \nabla_k f_m(\Gamma_m^{k,t})\right\|^2\Big] && (11)
\end{aligned}$$

$$\leq \frac{1}{N^2}\sum_{m=1}^{M} N_m^2 \frac{\sigma_k^2}{B} \tag{12}$$

$$\leq \frac{1}{M}\frac{\sigma_k^2}{B}. \tag{13}$$

Then, based on Assumption 4 and the Inequality (13), we can bound

$$\mathbb{E}\left\|\boldsymbol{G}^{k,t}\right\|^2$$

$$= \mathbb{E}\left\|\boldsymbol{G}^{k,t} - \mathbb{E}\left[\boldsymbol{G}^{k,t}\right]\right\|^2 + \left\|\mathbb{E}\left[\boldsymbol{G}^{k,t}\right]\right\|^2 \tag{14}$$

$$\leq \frac{1}{M}\frac{\sigma_k^2}{B} + \left\|\frac{1}{N}\sum_{m=1}^{M} N_m \mathbb{E}\left[\nabla_k f_m\left(\Gamma_m^{k,t};\mathcal{B}_m^{t_0}\right)\right]\right\|^2 \tag{15}$$

$$\leq \frac{1}{M}\frac{\sigma_k^2}{B} + \frac{1}{N^2}\sum_{m=1}^{M} N_m^2 \mathbb{E}\left\|\nabla_k f_m\left(\Gamma_m^{k,t};\mathcal{B}_m^{t_0}\right)\right\|^2 \tag{16}$$

$$\leq \frac{1}{M}\left(\frac{\sigma_k^2}{B} + \frac{\delta^2}{K}\right). \tag{17}$$

Since $\langle a,b\rangle = \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2 - \frac{1}{2}\|a-b\|^2$, we can bound

$$\mathbb{E}\left\langle \nabla_k f\left(\Theta^t\right), \eta\boldsymbol{G}^{k,t}\right\rangle$$

$$= \left\langle \nabla_k f\left(\Theta^t\right), \mathbb{E}\left[\eta\boldsymbol{G}^{k,t}\right]\right\rangle \tag{18}$$

$$= \frac{\eta}{N}\sum_{m=1}^{M} N_m \left\langle \nabla_k f\left(\Theta^t\right), \nabla_k f_m\left(\Gamma_m^{k,t}\right)\right\rangle \tag{19}$$

$$= \frac{\eta}{2N}\sum_{m=1}^{M} N_m \left[\left\|\nabla_k f\left(\Theta^t\right)\right\|^2 + \left\|\nabla_k f_m\left(\Gamma_m^{k,t}\right)\right\|^2 - \left\|\nabla_k f\left(\Theta^t\right) - \nabla_k f_m\left(\Gamma_m^{k,t}\right)\right\|^2\right] \tag{20}$$

$$= \frac{\eta}{2}\left\|\nabla_k f\left(\Theta^t\right)\right\|^2 + \frac{\eta}{2N}\sum_{m=1}^{M} N_m \left\|\nabla_k f_m\left(\Gamma_m^{k,t}\right)\right\|^2 - \frac{\eta}{2N}\sum_{m=1}^{M} N_m \left\|\nabla_k f\left(\Theta^t\right) - \nabla_k f_m\left(\Gamma_m^{k,t}\right)\right\|^2 \tag{21}$$

$$= \frac{\eta}{2}\left\|\nabla_k f\left(\Theta^t\right)\right\|^2 + \frac{\eta}{2N}\sum_{m=1}^{M} N_m \left\|\nabla_k f_m\left(\Gamma_m^{k,t}\right)\right\|^2 - \frac{\eta}{2N}\sum_{m=1}^{M} N_m \left\|\nabla_k f\left(\Theta^t\right) - \nabla_k f_m\left(\Theta^t\right) + \nabla_k f_m\left(\Theta^t\right) - \nabla_k f_m\left(\Gamma_m^{k,t}\right)\right\|^2 \tag{22}$$

$$\geq \frac{\eta}{2}\left\|\nabla_k f(\Theta^t)\right\|^2 + \frac{\eta}{2N}\sum_{m=1}^{M} N_m \left\|\nabla_k f_m(\Gamma_m^{k,t})\right\|^2 - \frac{\eta}{N}\sum_{m=1}^{M} N_m \left\|\nabla_k f\left(\Theta^t\right) - \nabla_k f_m\left(\Theta^t\right)\right\|^2 - \frac{\eta}{N}\sum_{m=1}^{M} N_m \left\|\nabla_k f_m\left(\Theta^t\right) - \nabla_k f_m\left(\Gamma_m^{k,t}\right)\right\|^2 \tag{23}$$

$$\geq \frac{\eta}{2}\left\|\nabla_k f\left(\Theta^t\right)\right\|^2 + \frac{\eta}{2N}\sum_{m=1}^{M} N_m \left\|\nabla_k f_m\left(\Gamma_m^{k,t}\right)\right\|^2 - \frac{\eta}{N}\sum_{m=1}^{M} N_m \left\|\nabla_k f_m\left(\Theta^t\right) - \nabla_k f_m\left(\Gamma_m^{k,t}\right)\right\|^2 \tag{24}$$

$$\geq \frac{\eta}{2}\left\|\nabla_k f\left(\Theta^t\right)\right\|^2 + \frac{\eta}{2N}\sum_{m=1}^{M} N_m \left\|\nabla_k f_m\left(\Gamma_m^{k,t}\right)\right\|^2 - \frac{\eta L_k^2}{N}\sum_{m=1}^{M} N_m \left\|\Theta^t - \Gamma_m^{k,t}\right\|^2. \tag{25}$$

The first Inequality (23) above arises from the relationship $|a-b|^2 \leq 2|a-c|^2 + 2|c-b|^2$.

Then, based on the previous Inequality (8), we have

$$\mathbb{E}\left[f\left(\Theta^{t+1}\right)\right] \leq f\left(\Theta^t\right) - \sum_{k=0}^{K}\frac{\eta}{2}\left\|\nabla_k f\left(\Theta^t\right)\right\|^2 - \sum_{k=0}^{K}\frac{\eta}{2N}\sum_{m=1}^{M} N_m \left\|\nabla_k f_m\left(\Gamma_m^{k,t}\right)\right\|^2 + \sum_{k=0}^{K}\frac{\eta L_k^2}{N}\sum_{m=1}^{M} N_m \left\|\Theta^t - \Gamma_m^{k,t}\right\|^2 + \frac{L\eta^2}{2}\sum_{k=0}^{K}\mathbb{E}\left\|\boldsymbol{G}^{k,t}\right\|^2 \tag{26}$$

$$= f\left(\Theta^t\right) - \frac{\eta}{2}\left\|\nabla f\left(\Theta^t\right)\right\|^2 - \sum_{k=0}^{K}\frac{\eta}{2N}\sum_{m=1}^{M} N_m \left\|\nabla_k f_m\left(\Gamma_m^{k,t}\right)\right\|^2 + \sum_{k=0}^{K}\frac{\eta L_k^2}{N}\sum_{m=1}^{M} N_m \left\|\Theta^t - \Gamma_m^{k,t}\right\|^2 + \frac{L\eta^2}{2}\sum_{k=0}^{K}\mathbb{E}\left\|\boldsymbol{G}^{k,t}\right\|^2 \tag{27}$$

$$\leq f\left(\Theta^t\right) - \frac{\eta}{2}\left\|\nabla f\left(\Theta^t\right)\right\|^2 - \sum_{k=0}^{K}\frac{\eta}{2N}\sum_{m=1}^{M} N_m \left\|\nabla_k f_m\left(\Gamma_m^{k,t}\right)\right\|^2 + \sum_{k=0}^{K}\frac{\eta L_k^2}{N}\sum_{m=1}^{M} N_m \left\|\Theta^t - \Gamma_m^{k,t}\right\|^2 + \frac{L\eta^2}{2M}\sum_{k=0}^{K}\left(\frac{\sigma_k^2}{B} + \frac{\delta^2}{K}\right). \tag{28}$$

Rearranging the Inequality (28) above, we have

$$\left\|\nabla f\left(\Theta^t\right)\right\|^2 \leq \frac{2}{\eta}\left[f\left(\Theta^t\right) - \mathbb{E}\left[f\left(\Theta^{t+1}\right)\right]\right] - \sum_{k=0}^{K}\frac{1}{N}\sum_{m=1}^{M} N_m \left\|\nabla_k f_m\left(\Gamma_m^{k,t}\right)\right\|^2 + \sum_{k=0}^{K}\frac{2L_k^2}{N}\sum_{m=1}^{M} N_m \left\|\Theta^t - \Gamma_m^{k,t}\right\|^2 + \frac{L\eta}{M}\sum_{k=0}^{K}\left(\frac{\sigma_k^2}{B} + \frac{\delta^2}{K}\right). \tag{29}$$

Now averaging over all P global rounds, i.e., all the intervals of RQ iterations over $t = 0, 1, \ldots, P-1$, such that $T = RQ \times P$ in Inequality (29) above, we have

$$\frac{1}{P}\sum_{t=0}^{P-1}\mathbb{E}\left[\left\|\nabla f\left(\Theta^t\right)\right\|^2\right] \leq \frac{2\left[f\left(\Theta^0\right) - f\left(\Theta^*\right)\right]}{\eta P}$$

$$
\begin{aligned}
&- \sum_{k=0}^{K} \sum_{t=0}^{P-1} \frac{1}{NP} \sum_{m=1}^{M} N_m \left\| \nabla_k f_m \left(\Gamma_m^{k,t} \right) \right\|^2 \\
&+ \sum_{k=0}^{K} \sum_{t=0}^{P-1} \frac{2L_k^2}{NP} \sum_{m=1}^{M} N_m \left\| \Theta^t - \Gamma_m^{k,t} \right\|^2 + \frac{L\eta}{M} \sum_{k=0}^{K} (\frac{\sigma_k^2}{B} + \frac{\delta^2}{K}),
\end{aligned} \tag{30}
$$

where $f(\Theta^*)$ is the optimal value of the global objective. It is noted here that we employ the actual (instead of "virtual") global model updated every RQ iterations.

Next, we consider $t_0 \leq t$ as the final iteration in which each silo conducts VFL across K parties in every round, and $T_0 \leq t_0$ as the final iteration in which the global server conducts HFL across M silos in each round. Therefore, $t-1-t_0 \leq Q$, $t-1-T_0 \leq RQ$, $t_0-1-T_0 \leq RQ-Q$. We can then establish the following bound

$$
\begin{aligned}
&\sum_{m=1}^{M} N_m \mathbb{E} \left\| \Theta^t - \Gamma_m^{k,t} \right\|^2 \\
&= \sum_{m=1}^{M} N_m \Big(\sum_{k'=0}^{K-1} \mathbb{E} \left\| \Theta^{k',t} - \Theta_m^{k',t_0} \right\|^2 + \mathbb{E} \left\| \Theta^{k,t} - \Theta_m^{k,t} \right\|^2 \Big) \\
&= \sum_{m=1}^{M} N_m \Big(\sum_{k'=0}^{K-1} \mathbb{E} \left\| \Theta_m^{k',t_0} - \Theta^{k',t} \right\|^2 + \mathbb{E} \left\| \Theta_m^{k,t} - \Theta^{k,t} \right\|^2 \Big) \\
&= \sum_{m=1}^{M} N_m \Big(\sum_{k'=0}^{K-1} \mathbb{E} \Big\| \Theta_m^{k',t_0} - \Big(\Theta^{k',T_0} - \sum_{\tau=T_0}^{t-1} \eta \boldsymbol{G}^{k',\tau} \Big) \Big\|^2 \\
&\quad + \mathbb{E} \Big\| \Theta_m^{k,t} - \Big(\Theta^{k,T_0} - \sum_{\tau=T_0}^{t-1} \eta \boldsymbol{G}^{k,\tau} \Big) \Big\|^2 \Big)
\end{aligned} \tag{31}
$$

$$
\begin{aligned}
&= \sum_{m=1}^{M} N_m \sum_{k'=0}^{K-1} \mathbb{E} \Big\| - \sum_{\tau=T_0}^{t_0-1} \eta \nabla_{k'} f \Big(\Gamma_m^{k',\tau}; \mathcal{B}_m^{t_0} \Big) + \sum_{\tau=T_0}^{t-1} \eta \boldsymbol{G}^{k',\tau} \Big\|^2 \\
&\quad + \sum_{m=1}^{M} N_m \mathbb{E} \Big\| - \sum_{\tau=T_0}^{t-1} \eta \nabla_k f \Big(\Gamma_m^{k,\tau}; \mathcal{B}_m^{t_0} \Big) + \sum_{\tau=T_0}^{t-1} \eta \boldsymbol{G}^{k,\tau} \Big\|^2 \\
&\leq \sum_{m=1}^{M} N_m \sum_{k'=0}^{K-1} \mathbb{E} \Big\| \sum_{\tau=T_0}^{t_0-1} \eta \nabla_{k'} f \Big(\Gamma_m^{k',\tau}; \mathcal{B}_m^{t_0} \Big) \Big\|^2 \\
&\quad + \sum_{k'=0}^{K-1} \sum_{m=1}^{M} N_m \mathbb{E} \Big\| \sum_{\tau=T_0}^{t-1} \eta \boldsymbol{G}^{k',\tau} \Big\|^2 \\
&\quad + \sum_{m=1}^{M} N_m \mathbb{E} \Big\| \sum_{\tau=T_0}^{t-1} \eta \nabla_k f \Big(\Gamma_m^{k,\tau}; \mathcal{B}_m^{t_0} \Big) \Big\|^2 \\
&\quad + \sum_{m=1}^{M} N_m \mathbb{E} \Big\| \sum_{\tau=T_0}^{t-1} \eta \boldsymbol{G}^{k,\tau} \Big\|^2
\end{aligned} \tag{32}
$$

$$
\begin{aligned}
&\leq \sum_{m=1}^{M} N_m \sum_{k'=0}^{K-1} \eta^2 \left(t_0 - 1 - T_0 \right) \sum_{\tau=T_0}^{t_0-1} \mathbb{E} \left\| \nabla_{k'} f \left(\Gamma_m^{k',\tau}; \mathcal{B}_m^{t_0} \right) \right\|^2 \\
&\quad + \sum_{m=1}^{M} N_m \eta^2 \left(t - 1 - T_0 \right) \sum_{\tau=T_0}^{t-1} \sum_{k'=0}^{K-1} \mathbb{E} \Big\| \boldsymbol{G}^{k',\tau} \Big\|^2 \\
&\quad + \sum_{m=1}^{M} N_m \eta^2 (t-1-T_0) \sum_{\tau=T_0}^{t-1} \mathbb{E} \Big\| \nabla_k f \Big(\Gamma_m^{k,\tau}; \mathcal{B}_m^{t_0} \Big) \Big\|^2 \\
&\quad + \sum_{m=1}^{M} N_m \eta^2 (t-1-T_0) \sum_{\tau=T_0}^{t-1} \mathbb{E} \Big\| \boldsymbol{G}^{k,\tau} \Big\|^2
\end{aligned} \tag{33}
$$

$$
\begin{aligned}
&\leq K \Big[\eta^2 (t_0 - 1 - T_0)^2 N \frac{\delta^2}{K} \\
&\quad + \eta^2 (t-1-T_0)^2 N \frac{1}{M} \Big(\frac{\sigma_k^2}{B} + \frac{\delta^2}{K} \Big) \Big] \\
&\quad + \eta^2 (t-1-T_0)^2 N \frac{\delta^2}{K} \\
&\quad + \eta^2 (t-1-T_0)^2 N \frac{1}{M} \Big(\frac{\sigma_k^2}{B} + \frac{\delta^2}{K} \Big)
\end{aligned} \tag{34}
$$

$$
\begin{aligned}
&\leq K \Big[\eta^2 (RQ-Q)^2 N \frac{\delta^2}{K} + \eta^2 (RQ)^2 N \frac{1}{M} \Big(\frac{\sigma_k^2}{B} + \frac{\delta^2}{K} \Big) \Big] \\
&\quad + \eta^2 (RQ)^2 N \frac{\delta^2}{K} + \eta^2 (RQ)^2 N \frac{1}{M} \Big(\frac{\sigma_k^2}{B} + \frac{\delta^2}{K} \Big)
\end{aligned} \tag{35}
$$

$$
\begin{aligned}
&= K N \eta^2 \Big[(RQ-Q)^2 \frac{\delta^2}{K} + (RQ)^2 \frac{1}{M} \Big(\frac{\sigma_k^2}{B} + \frac{\delta^2}{K} \Big) \Big] \\
&\quad + \eta^2 (RQ)^2 N \frac{\delta^2}{K} + \eta^2 (RQ)^2 N \frac{1}{M} \Big(\frac{\sigma_k^2}{B} + \frac{\delta^2}{K} \Big)
\end{aligned} \tag{36}
$$

$$
\begin{aligned}
&= N \eta^2 \Big(\Big[K (RQ-Q)^2 + (\frac{K+1}{M} + 1)(RQ)^2 \Big] \frac{\delta^2}{K} \\
&\quad + \frac{K+1}{M} (RQ)^2 \frac{\sigma_k^2}{B} \Big).
\end{aligned} \tag{37}
$$

Finally, based on the previous inequality (30), we have

$$
\begin{aligned}
&\frac{1}{P} \sum_{t=0}^{P-1} \mathbb{E} \left[\left\| \nabla f \left(\Theta^t \right) \right\|^2 \right] \\
&\leq \frac{2 \left[f \left(\Theta^0 \right) - \mathbb{E} \left[f \left(\Theta^* \right) \right] \right]}{\eta P} \\
&\quad + 2\eta^2 \sum_{k=0}^{K} L_k^2 \Big[\Big(K (RQ-Q)^2 \\
&\quad + (\frac{K+1}{M} + 1) R^2 Q^2 \Big) \frac{\delta^2}{K} + \frac{K+1}{M} R^2 Q^2 \frac{\sigma_k^2}{B} \Big] \\
&\quad + \frac{\eta L}{M} \sum_{k=0}^{K} (\frac{\sigma_k^2}{B} + \frac{\delta^2}{K}),
\end{aligned} \tag{38}
$$

where $f(\Theta^*)$ is the optimal value of the global objective. It is noted here that we employ the actual (instead of "virtual") global model updated every RQ iterations.

Now we conclude the proof of Theorem 1. ■