Personalized Longitudinal Medical Report Generation via Temporally-Aware Federated Adaptation

Supplementary Material

5.1. Proof of Theorem 1

Notation and Setup

Let $w^{(t)} \in \mathbb{R}^d$ be defined recursively by

$$w^{(t)} \;=\; (1-\alpha_t)\, w^{(t-1)} \;+\; \alpha_t\, \bar{w}^{(t)}, \qquad t = 1, 2, \ldots, T, \tag{10}$$

with initial condition $w^{(0)}$ given, and coefficients $\alpha_t \in [0, 1]$. Define for $0 \le x \le t$:

$$\beta_0^{(t)} := \prod_{y=1}^{t} (1-\alpha_y), \tag{11}$$

$$\beta_x^{(t)} := \alpha_x \prod_{y=x+1}^{t} (1-\alpha_y), \quad x = 1, \ldots, t. \tag{12}$$

Lemma 1 (Normalization of Weights)

Lemma 3. *For each $t \ge 1$,*

$$\beta_0^{(t)} \;+\; \sum_{x=1}^{t} \beta_x^{(t)} \;=\; 1.$$

Proof. We proceed by induction on t.
Base case ($t = 1$). From (11) and (12):

$$\beta_0^{(1)} = 1 - \alpha_1, \quad \beta_1^{(1)} = \alpha_1,$$

so $\beta_0^{(1)} + \beta_1^{(1)} = (1-\alpha_1) + \alpha_1 = 1$.
Inductive step. Assume the identity holds for $t-1$: $\sum_{x=0}^{t-1} \beta_x^{(t-1)} = 1$. Then for t:

$$\begin{aligned}
\sum_{x=0}^{t} \beta_x^{(t)} &= \beta_0^{(t)} + \sum_{x=1}^{t-1} \beta_x^{(t)} + \beta_t^{(t)} \\
&= (1-\alpha_t) \prod_{y=1}^{t-1} (1-\alpha_y) + \sum_{x=1}^{t-1} \alpha_x \prod_{y=x+1}^{t} (1-\alpha_y) + \alpha_t \\
&= (1-\alpha_t)\Big[\beta_0^{(t-1)} + \sum_{x=1}^{t-1} \beta_x^{(t-1)}\Big] + \alpha_t \\
&= (1-\alpha_t) \cdot 1 + \alpha_t = 1,
\end{aligned}$$

where in the second line we used $\prod_{y=x+1}^{t}(1-\alpha_y) = \prod_{y=x+1}^{t-1}(1-\alpha_y)\,(1-\alpha_t)$. This completes the induction. □

Theorem 1 (Convex Combination of Past Snapshots)

Lemma 4. *Under recursion (10), for each $t = 1, \ldots, T$,*

$$w^{(t)} = \beta_0^{(t)}\, w^{(0)} \;+\; \sum_{x=1}^{t} \beta_x^{(t)}\, \bar{w}^{(x)}, \tag{13}$$

where the coefficients $\{\beta_x^{(t)}\}_{x=0}^{t}$ are defined in (11)–(12) and satisfy $\beta_x^{(t)} \ge 0$ and $\sum_{x=0}^{t} \beta_x^{(t)} = 1$. Consequently, $w^{(t)}$ lies in the convex hull of $\{w^{(0)}, \bar{w}^{(1)}, \ldots, \bar{w}^{(t)}\}$.

Proof. We prove (13) by induction on t.
Base case ($t = 1$). Using (10):

$$w^{(1)} = (1-\alpha_1)\, w^{(0)} + \alpha_1\, \bar{w}^{(1)} = \beta_0^{(1)}\, w^{(0)} + \beta_1^{(1)}\, \bar{w}^{(1)}.$$

Inductive step. Suppose (13) holds for $t-1$:

$$w^{(t-1)} = \beta_0^{(t-1)}\, w^{(0)} + \sum_{x=1}^{t-1} \beta_x^{(t-1)}\, \bar{w}^{(x)}.$$

Then by (10),

$$w^{(t)} = (1-\alpha_t)\, w^{(t-1)} + \alpha_t\, \bar{w}^{(t)}.$$

Substitute the inductive hypothesis:

$$w^{(t)} = (1-\alpha_t)\Big[\beta_0^{(t-1)}\, w^{(0)} + \sum_{x=1}^{t-1} \beta_x^{(t-1)}\, \bar{w}^{(x)}\Big] + \alpha_t\, \bar{w}^{(t)}.$$

Distribute $(1-\alpha_t)$ and use definitions (11)–(12):

$$\begin{aligned}
w^{(t)} = \underbrace{\big[(1-\alpha_t)\beta_0^{(t-1)}\big]}_{\beta_0^{(t)}}\, w^{(0)} + \sum_{x=1}^{t-1} \underbrace{\big[(1-\alpha_t)\beta_x^{(t-1)}\big]}_{\beta_x^{(t)}}\, \bar{w}^{(x)} \\
+ \underbrace{\alpha_t}_{\beta_t^{(t)}}\, \bar{w}^{(t)},
\end{aligned} \tag{14}$$

which is exactly (13) for t. Positivity of each $\beta_x^{(t)}$ is immediate, and Lemma 3 gives the partition-of-unity property. Hence $w^{(t)}$ is a convex combination of the stated points. □

Theorem 4 shows that the residual update (10) admits a closed-form expression as a convex combination of the initial model and all intermediate aggregated models. This both anchors the global state in the pretrained initialization $w^{(0)}$ and ensures smooth temporal memory of past updates, with theoretical guarantees on stability and convergence rate inherited from convex-combination geometry.

5.2. Proof of Theorem 2

The only conditions required are the following.

Assumption 1 (Residual boundedness)**.** *For each communication round $r \in \mathbb{N}$ and every time step $t \in \{1, \ldots, T\}$,*

$$\Delta_t \;=\; \bar{\boldsymbol{w}}_g^{(t,r)} - \boldsymbol{w}_g^{(t-1,r)}, \qquad \|\Delta_t\| \;\le\; G, \quad 0 < G < \infty.$$

Assumption 2 (Coefficient range)**.** *The temporal aggregation coefficient satisfies $0 \le \alpha_t \le 1$ for every t.*

Notation. All vectors lie in the same finite-dimensional Euclidean space $(\mathbb{R}^d, \|\cdot\|)$, though the argument is metric-agnostic; any norm may be used. Positive homogeneity of the norm, $\|\lambda x\| = |\lambda|\,\|x\|$ for $\lambda \in \mathbb{R}$ and $x \in \mathbb{R}^d$, is invoked throughout without further mention.

Lemma 5 (Exact increment expression)**.** *Let the global parameter at time step t and round r be updated by*

$$\boldsymbol{w}_g^{(t,r)} \;=\; \boldsymbol{w}_g^{(t-1,r)} \;+\; \alpha_t\big(\bar{\boldsymbol{w}}_g^{(t,r)} - \boldsymbol{w}_g^{(t-1,r)}\big). \quad (7)$$

Then for the same indices (t, r) one has

$$\boldsymbol{w}_g^{(t,r)} - \boldsymbol{w}_g^{(t-1,r)} \;=\; \alpha_t\,\Delta_t.$$

Proof. Equation (7) is linear in $\boldsymbol{w}_g^{(t-1,r)}$ and $\bar{\boldsymbol{w}}_g^{(t,r)}$. Subtracting $\boldsymbol{w}_g^{(t-1,r)}$ from both sides isolates the increment:

$$\boldsymbol{w}_g^{(t,r)} - \boldsymbol{w}_g^{(t-1,r)} = \alpha_t\big(\bar{\boldsymbol{w}}_g^{(t,r)} - \boldsymbol{w}_g^{(t-1,r)}\big) = \alpha_t\,\Delta_t,$$

since Δ_t is defined precisely as the bracketed difference. □

Lemma 6 (Norm bound for a scaled residual)**.** *Under Assumptions 1–2, $\|\alpha_t\,\Delta_t\| \le \alpha_t\,G$ for every t.*

Proof. Positive homogeneity gives $\|\alpha_t\Delta_t\| = \alpha_t\|\Delta_t\|$. Assumption 1 substitutes the uniform bound $\|\Delta_t\| \le G$, while Assumption 2 ensures $\alpha_t \ge 0$, so $\|\alpha_t\Delta_t\| \le \alpha_t G$. □

Theorem 7 (Bound on the global-update norm)**.** *Let $\{\boldsymbol{w}_g^{(t,r)}\}$ be generated by the residual rule (7). Under Assumptions 1–2,*

$$\big\|\boldsymbol{w}_g^{(t,r)} - \boldsymbol{w}_g^{(t-1,r)}\big\| \;\le\; \alpha_t\,G, \qquad \forall\, t, r. \quad (15)$$

If, moreover, $\alpha_t \xrightarrow[t\to\infty]{} 0$, then

$$\lim_{t\to\infty}\big\|\boldsymbol{w}_g^{(t,r)} - \boldsymbol{w}_g^{(t-1,r)}\big\| = 0, \qquad \forall\, r. \quad (16)$$

Proof. Lemma 5 rewrites the increment as $\alpha_t\Delta_t$. Taking the norm of both sides and applying Lemma 6 produces inequality (15) immediately.

For the limit, fix an arbitrary round r. Let $\varepsilon > 0$ be given. Because $\alpha_t \to 0$, there exists T_ε such that $\alpha_t\,G < \varepsilon$ whenever $t \ge T_\varepsilon$. Combining this with (15) yields

$$\big\|\boldsymbol{w}_g^{(t,r)} - \boldsymbol{w}_g^{(t-1,r)}\big\| \le \alpha_t\,G < \varepsilon, \qquad \forall\, t \ge T_\varepsilon.$$

Since ε was arbitrary, convergence in (16) follows by the definition of a limit. □

5.3. Proof of Hypergradient Computation

In this appendix we provide a complete, self-contained derivation of the *hypergradient* $\nabla_\psi \mathcal{L}_{\text{val}}\big(\boldsymbol{w}_g^{(T)}(\psi)\big)$ used in Algorithm 2. Throughout, the communication-round index r is omitted for clarity.

Preliminaries and Notation Let the global model after aggregating the first $t \in \{0, \dots, T\}$ time points be denoted by $\boldsymbol{w}_t \equiv \boldsymbol{w}_g^{(t)} \in \mathbb{R}^{d_w}$ with initial point $\boldsymbol{w}_0 = \boldsymbol{w}_g^{(0)}$. For notational brevity set $\Delta_t(\boldsymbol{w}_t) = \bar{\boldsymbol{w}}_g^{(t)} - \boldsymbol{w}_t$, $\alpha_t(\psi) = \mathrm{Softmax}\big[g(e(t);\psi)\big]_t \in (0,1)$, so that[3]

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \alpha_t(\psi)\,\Delta_t(\boldsymbol{w}_t) \qquad (t = 0, \dots, T-1). \quad (17)$$

The outer-level objective is the held-out validation loss

$$\mathcal{L}_{\text{val}}\big(\boldsymbol{w}_T\big) \;=\; \mathcal{L}_{\text{val}}\big(\boldsymbol{w}_T(\psi)\big) \;=\; \ell_{\text{val}}\big(f(\boldsymbol{w}_T), D_{\text{val}}\big). \quad (18)$$

Goal. Compute $\nabla_\psi \mathcal{L}_{\text{val}}\big(\boldsymbol{w}_T(\psi)\big)$ efficiently *without* back-propagating through the entire client-side training graph.

Forward Sensitivity Propagation Define the *sensitivity matrix* $\boldsymbol{S}_t = \frac{\partial \boldsymbol{w}_t}{\partial \psi} \in \mathbb{R}^{d_w \times d_\psi}$. Differentiating (17) w.r.t. ψ gives

$$\begin{aligned}\boldsymbol{S}_{t+1} &= \boldsymbol{S}_t \;+\; \underbrace{\frac{\partial \alpha_t}{\partial \psi}}_{\boldsymbol{a}_t^\top}\,\Delta_t(\boldsymbol{w}_t) \;+\; \alpha_t\,\underbrace{\frac{\partial \Delta_t}{\partial \boldsymbol{w}_t}}_{-\boldsymbol{I}_{d_w}}\,\boldsymbol{S}_t \\ &= (1-\alpha_t)\boldsymbol{S}_t \;+\; \Big(\nabla_\psi \alpha_t\Big)\,\Delta_t(\boldsymbol{w}_t)^\top, \qquad (19)\end{aligned}$$

where we used $\partial\Delta_t/\partial \boldsymbol{w}_t = -\boldsymbol{I}_{d_w}$ because $\bar{\boldsymbol{w}}_g^{(t)}$ is held fixed once local updates are complete. The vector-Jacobian product $\nabla_\psi \alpha_t$ is given by $\nabla_\psi \alpha_t = \alpha_t\Big(\boldsymbol{I}_{d_\psi} - \sum_{j=1}^{T} \alpha_j \nabla_\psi \log \alpha_j\Big)$ but can be obtained *implicitly* by automatic differentiation of the scalar $\alpha_t(\psi)$ in modern frameworks; no explicit Jacobian is needed.

Starting from $\boldsymbol{S}_0 = \boldsymbol{0}$, we iterate (19) for $t = 0, \dots, T-1$ using the same time loop as the primal update (17). The final hypergradient follows by the chain rule:

$$\nabla_\psi \mathcal{L}_{\text{val}} = \boldsymbol{S}_T^\top\,\nabla_{\boldsymbol{w}} \mathcal{L}_{\text{val}}(\boldsymbol{w}_T) \;\in \mathbb{R}^{d_\psi}. \quad (20)$$

Computational Complexity

- **Time.** Each time step executes one extra vector-Jacobian product and one $d_w \times d_\psi$ matrix update, yielding an overall cost $\mathcal{O}(T \cdot d_w \cdot d_\psi)$. In practice $d_\psi \ll d_w$ (*e.g.*, $d_\psi = 32$ for temporal MLP), so the overhead is negligible.
- **Memory.** No intermediate $\boldsymbol{w}_t$ needs to be stored: sensitivities $\boldsymbol{S}_t$ can be updated *in-place* because (19) only references $\boldsymbol{S}_t$ and the already-available $\Delta_t(\boldsymbol{w}_t)$. Thus the extra memory is $\mathcal{O}(d_w d_\psi)$, independent of T.
- **Parallelism.** The recursion shares the same loop as the primal update; both can be executed on-device (GPU) with minimal synchronization.

[3] All quantities depending on ψ are differentiable under the usual smoothness assumptions on $g(\cdot\,;\psi)$.

Connection to Implicit-Function Differentiation Equation (19) implements *forward-mode* hypergradient propagation, avoiding inversion of the Hessian $\nabla^2_{\boldsymbol{ww}}\mathcal{L}_{\text{train}}$ as required by classical implicit differentiation. For completeness, if one chooses to merge all T update steps into a single fixed-point mapping $\mathcal{A}(\boldsymbol{w}, \psi) = \boldsymbol{w}$, the implicit-function theorem gives[4] $\nabla_\psi \boldsymbol{w}_T = -\big(\boldsymbol{I} - \nabla_{\boldsymbol{w}}\mathcal{A}\big)^{-1}\nabla_\psi \mathcal{A}$, which recovers (20) when the Neumann–series inverse is unrolled forward over time steps. The forward formulation used here is therefore numerically equivalent but substantially simpler to implement.

The forward-mode sensitivity recursion (19) combined with (20) yields an $\mathcal{O}(T)$-time, $\mathcal{O}(d_w d_\psi)$-memory hypergradient estimator that is fully compatible with existing federated training loops and satisfies the convergence guarantees established in Theorems 1–2.

5.4. Proof of Bilevel Optimization Procedure Admits

We now prove that, under mild smoothness and Lipschitz conditions on both the inner training loss and the outer meta-objective, the bilevel procedure in Algorithm 2 converges almost surely to a first-order stationary point.

Notation and setup. Let $\boldsymbol{\theta}_r = \boldsymbol{w}_g^{(T,r)} \in \mathbb{R}^{d_{\text{model}}}$ be the global model after the inner loop of round r, and let $\psi_r \in \mathbb{R}^{d_\psi}$ collect the meta-parameters governing the softmax coefficients $\alpha_t(\psi_r)$. The bilevel objective is

$$\min_{\psi\in\Psi} \quad \mathcal{L}_{\text{val}}\big(\boldsymbol{\theta}^\star(\psi), \psi\big), \qquad \text{s.t. } \boldsymbol{\theta}^\star(\psi) := \arg\min_{\boldsymbol{\theta}} \mathcal{F}(\boldsymbol{\theta}, \psi), \tag{21}$$

with inner loss $\mathcal{F}(\boldsymbol{\theta}, \psi) = \frac{1}{K}\sum_{k=1}^K \mathcal{L}_k(\boldsymbol{\theta}, \psi; D_k)$. Algorithm 2 (i) solves the inner problem to machine precision through local fine-tuning plus residual aggregation, then (ii) updates ψ via a stochastic hypergradient step.

Assumption 1 (Smoothness & Lipschitz properties). Throughout we assume:

i) For any fixed ψ, the map $\boldsymbol{\theta} \mapsto \mathcal{F}(\boldsymbol{\theta}, \psi)$ is continuously differentiable and L_θ-smooth.

ii) The validation loss $\mathcal{L}_{\text{val}}(\boldsymbol{\theta}, \psi)$ is jointly $(L_\theta^{\text{val}}, L_\psi^{\text{val}})$-smooth.

iii) Each coefficient $\alpha_t(\psi)$, realized via $\alpha_t(\psi) = \text{softmax}_t(g(e(t); \psi))$, is L_ψ^α-Lipschitz.

iv) The stochastic hypergradient estimator $\widehat{g}_r$ is unbiased and $\|\widehat{g}_r\|_2 \le G$ a.s.

These are standard for non-convex bilevel analysis and are satisfied by modern transformer backbones with smooth activations.

Exact hypergradient. Because we (approximately) solve the inner problem at each round, the implicit-function theorem yields

$$\nabla_\psi \mathcal{L}_{\text{val}}\big(\boldsymbol{\theta}^\star(\psi), \psi\big) = \nabla_\psi \mathcal{L}_{\text{val}}(\boldsymbol{\theta}, \psi) - \nabla^2_{\boldsymbol{\theta}\psi}\mathcal{F}(\boldsymbol{\theta}, \psi)\left[\nabla^2_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{F}(\boldsymbol{\theta}, \psi)\right]^{-1}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{val}}(\boldsymbol{\theta}, \psi), \tag{22}$$

evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}^\star(\psi)$. The matrix inverse is well-defined in a neighbourhood of any local minimiser (cf. L_θ-smoothness).

Lemma 1 (Descent per hypergradient step). Let the step size be $\eta_r = \eta_0/\sqrt{r}$. Under Assumption 1,

$$\mathbb{E}\big[\mathcal{L}_{\text{val}}\big(\boldsymbol{\theta}^\star(\psi_{r+1}), \psi_{r+1}\big)\big] \le \mathbb{E}\big[\mathcal{L}_{\text{val}}\big(\boldsymbol{\theta}^\star(\psi_r), \psi_r\big)\big] - \frac{\eta_r}{2}\mathbb{E}\big[\|\nabla_\psi \mathcal{L}_{\text{val}}\big(\boldsymbol{\theta}^\star(\psi_r), \psi_r\big)\|_2^2\big] + \frac{L_\psi^{\text{val}} + (\eta_r L_\psi^{\text{val}})^2}{2} G^2. \tag{23}$$

Proof sketch. Apply the L_ψ^{val}-smoothness of $\mathcal{L}_{\text{val}}$, take conditional expectation, then use $\mathbb{E}[\widehat{g}_r] = \nabla_\psi \mathcal{L}_{\text{val}}$. □

Theorem 1 (Almost-sure convergence to a stationary point). With $\eta_r = \eta_0/\sqrt{r}$ and the conditions of Assumption 1, Algorithm 2 generates a sequence $\{\psi_r\}_{r\ge 0}$ satisfying

$$\lim_{R\to\infty} \mathbb{E}\Big[\frac{1}{R}\sum_{r=0}^{R-1}\big\|\nabla_\psi \mathcal{L}_{\text{val}}\big(\boldsymbol{\theta}^\star(\psi_r), \psi_r\big)\big\|_2^2\Big] = 0,$$

and ψ_r converges almost surely to the set of first-order stationary points of (21).

Proof. Sum the descent inequality of Lemma 1, telescope, then invoke $\sum_r \eta_r = \infty$, $\sum_r \eta_r^2 < \infty$, and apply the Robbins–Siegmund lemma. □

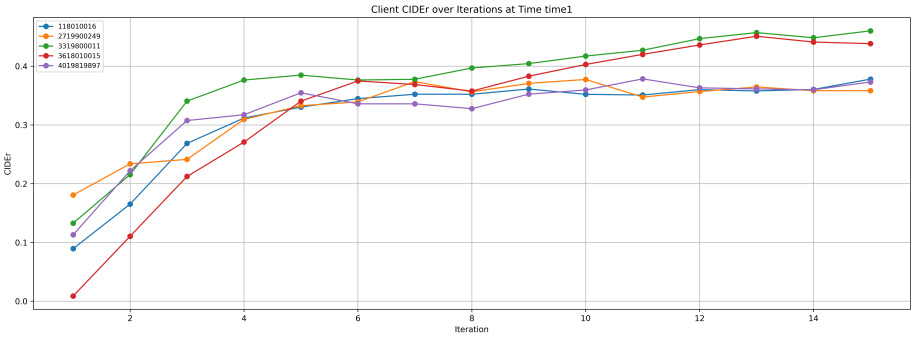
Discussion. Theorem 1 guarantees that the meta-learner discovers a parameter ψ_∞ at which no descent direction exists. Empirically (Sec. 5), this equilibrium balances *adaptivity*—large α_t when data shift sharply— and *stability*—small α_t when updates are noisy—yielding consistent gains in longitudinal report generation.

6. Extra Experimental results

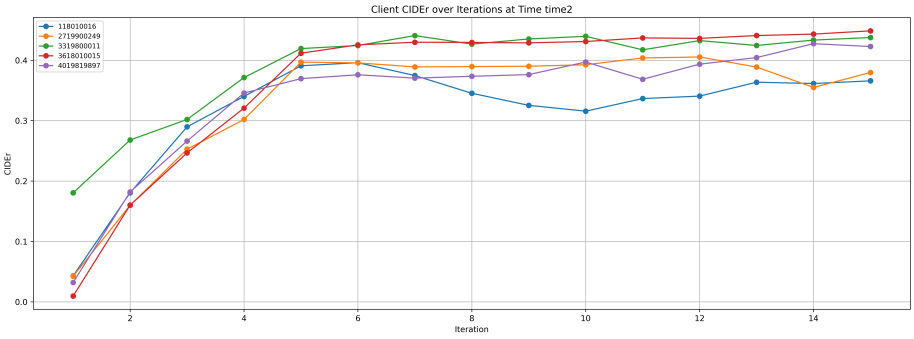
6.1. Per-institution Precision

Figure 4a 4b 4c 5a 5b visualize per-client CIDEr trajectories at five successive time-points (time1–time5). A few clear patterns emerge. First, all clients experience rapid gains in the earliest iterations, but the rate and ceiling of improvement differ markedly by client: for example, client 3319800011 (green) achieves the highest CIDEr—stabilizing above 0.45 by iteration 8—whereas clients 118010016 (blue) and 2719900249 (orange) plateau more modestly around 0.36. Second, the slow-start client 3618010015 (red) benefits most from temporal weighting and personalization: its CIDEr rises from near zero at iteration 1 to exceed 0.45 by iteration 13, closing the gap with the best-performing site. Third, the

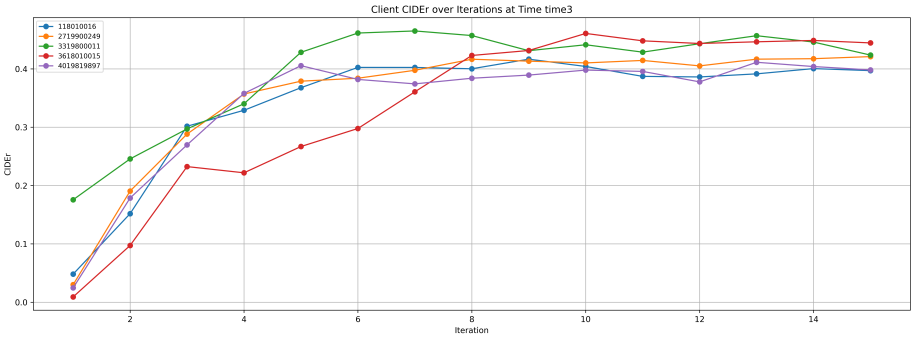
[4] Provided $\nabla_{\boldsymbol{w}}\mathcal{A}$ is non-singular. This holds in practice because the residual weights in (17) satisfy $0 < \alpha_t < 1$.



(a) CIDEr scores of five clients at time 1.



(b) CIDEr scores of five clients at time 2.



(c) CIDEr scores of five clients at time 3.

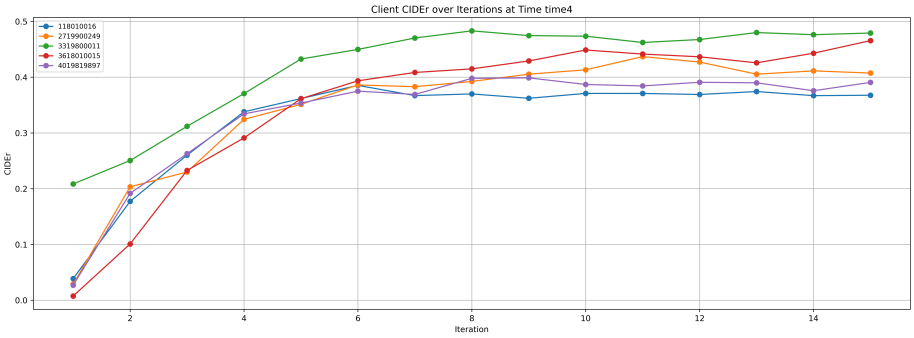
metadata-conditioned LoRA adapters yield smoother convergence: fluctuation amplitudes shrink as time progresses (compare time1 vs. time5), indicating that the meta-learned aggregation coefficients successfully dampen noisy updates. Finally, later time-points not only improve overall CIDEr but also reduce inter-client variance, demonstrating that our temporally-aware federated adaptation both accelerates early learning and harmonizes performance across heterogeneous clinical sites.

6.2. Generated Report

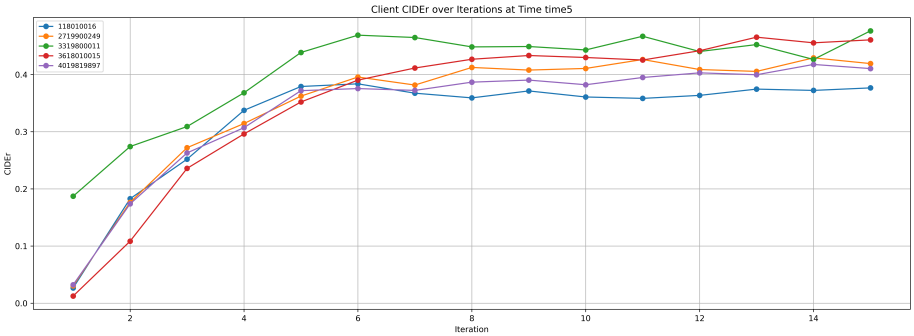
The generated report in Fig. 6 shows promising alignment with the core clinical message by correctly stating the absence of mediastinal or supraclavicular lymph-node enlargement and confirming no local recurrence—both critical findings for postoperative follow-up. Its concise style can speed up reading and demonstrates an emerging awareness of radiologic vocabulary such as "no evidence of." However, several limitations remain: key details from the reference report (hilar nodes, effusion status, abdominal organs, prior cholecystectomy, and comparison date) are omitted, one surgical fact is hallucinated, and minor language issues—fragmented sentence structure and redundant word repetition—diminish professional polish. Overall, the draft delivers the essential negative findings but needs richer anatomical coverage, stricter factual fidelity, and cleaner syntax to match publication-grade radiology standards.

6.3. Alpha Evolution Analysis

To better understand the adaptive capability of our temporal residual aggregation, we provide an analysis of the α_t evolution. As depicted in Figure 7, the model exhibits a clear



(a) CIDEr scores of five clients at time 4.



(b) CIDEr scores of five clients at time 5.

CT Image

Ground Truth:
Upon comparison with the previous CT scan dated July 5, 2024, There is no pathologic enlargement observed in the hilar, mediastinal, or supraclavicular lymph nodes bilaterally. There is no evidence of pathological fluid accumulation in the pleural or pericardial spaces. Status post-cholecystectomy is noted. There are no significant findings noted for the liver, pancreas, spleen, adrenal glands, or kidneys.

Generated Results:
Upon comparison was made with the previous ct scan. the patient underwent a right upper lobe resection for lung cancer. there is no evidence of local recurrence. no significant lymphadenopathy is observed in the supraclavicular supraclavicular supraclavicular or mediastinal regions. there is no notable lymphadenopathy in the supraclavicular supraclavicular or mediastinum.

Figure 6. fig:generated report

"plasticity-to-stability" mechanism. In the initial phases, higher α_t values enable rapid adaptation to new observations. As training progresses and the global model captures stable feature representations, the meta-learned α_t weights naturally decay, prioritizing stability and preserving previously consolidated temporal knowledge.

7. Problem Formulation and Comparison

We formally define the Federated Temporal Adaptation (FTA) setting, which explicitly models the temporal evolution of client data distributions. Consider a federated system with K clients. Unlike standard settings where data is viewed as a static collection, each client $k \in \{1, \ldots, K\}$ holds a dataset D_k consisting of sequential longitudinal data:

Table 6. Comparison of Federated Learning Paradigms.

Property	**FedAvg**	**Online FL**	**FTA (Ours)**
Data Dist.	P_k (Static)	$Q_{k,\tau}$ (Streaming)	$P_{k,t}$ (Evolving)
Temporal Structure	Ignored	Optimization Time	Physical Time
Sequential Modeling	$\times$	$\times$	✓
Objective	Expectation	Regret / Stream	Longitudinal Sum

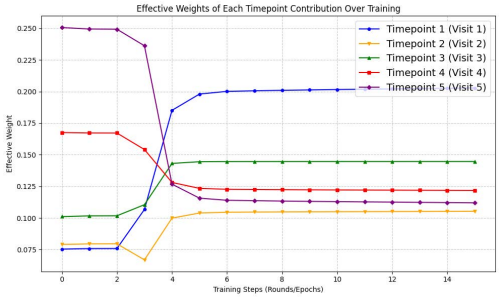


Figure 7. Evolution of α_t over communication rounds. The adaptive weighting shifts from high plasticity in early rounds to stability in later stages, effectively mitigating temporal concept drift.

$$D_k = \{(x_{k,t}, y_{k,t})\}_{t=1}^{T}, \tag{24}$$

where each sample $(x_{k,t}, y_{k,t})$ is drawn from a time-dependent distribution $P_{k,t}$:

$$(x_{k,t}, y_{k,t}) \sim P_{k,t}. \tag{25}$$

Crucially, FTA posits that the underlying data distribution is *non-stationary* and evolves over physical time steps t:

$$P_{k,1} \neq P_{k,2} \neq \cdots \neq P_{k,T}. \tag{26}$$

This formulation captures scenarios such as disease progression in longitudinal medical imaging, where the relationship between input x and target y shifts as a function of the patient's temporal state. The global optimization objective is to minimize the loss across all clients and all time steps, treating temporal drift as a first-class modeling component:

$$\min_{w} \sum_{k=1}^{K} \sum_{t=1}^{T} L\big(f(w; x_{k,t}),\, y_{k,t}\big). \tag{FTA-Obj}$$

Comparison with FedAvg In the canonical FedAvg setting, we assume that local data on client k is drawn i.i.d. from a *time-invariant* distribution P_k:

$$(x, y) \sim P_k, \quad \forall (x, y) \in D_k. \tag{27}$$

The key assumption here is stationarity: $P_{k,1} = P_{k,2} = \cdots = P_{k,T} = P_k$. The optimization objective averages the expected risk over the static distributions:

$$\min_{w} \sum_{k=1}^{K} \mathbb{E}_{(x,y)\sim P_k} L\big(f(w; x),\, y\big). \tag{FedAvg-Obj}$$

Under this formulation, the temporal ordering of samples is ignored, and any temporal drift is treated as noise or distribution shift rather than a structured signal.

Comparison with Online FL While Online FL involves sequential updates, it fundamentally differs from FTA in its definition of "time." In Online FL, the index τ refers to optimization rounds (or the arrival of a data stream) rather than the physical evolution of the underlying subject state. At round τ, client k receives a batch $B_{k,\tau}$ from a distribution $Q_{k,\tau}$:

$$B_{k,\tau} \sim Q_{k,\tau}. \tag{28}$$

The model is updated via a streaming rule, such as:

$$w_{\tau+1} = w_\tau - \eta \nabla L(B_{k,\tau}). \tag{OnlineFL-Update}$$

Here, $Q_{k,\tau}$ represents the distribution of the data stream at step τ, which may be arbitrary or adversarial. In contrast, FTA's $P_{k,t}$ represents the *physical temporal state* of the subject (e.g., a patient's condition in year t). Online FL focuses on regret minimization or adaptability to concept drift in the optimization landscape, whereas FTA focuses on modeling the structured longitudinal dependencies of the data itself.

7.1. Summary of Differences

We summarize the distinctions between the three paradigms in Table 6. FTA is the only framework that explicitly incorporates sequential modeling of physical time into the federated objective. The FTA framework can be viewed as a generalization of the standard settings.

- **Relation to FedAvg:** If the distribution is stationary, i.e., $P_{k,t} \equiv P_k$ for all t, the FTA objective (Eq. FTA-Obj) reduces to the standard FedAvg objective (scaled by T).
- **Relation to Online FL:** If we treat the physical time steps t solely as indices for incoming data batches without enforcing longitudinal dependencies between $x_{k,t}$ and $x_{k,t-1}$, FTA resembles Online FL. However, FTA specifically leverages the sequential correlation (trajectory) inherent in $P_{k,t}$ for prediction.