DECENTRALIZED DATA-FREE UNLEARNING IN CONTINUAL LEARNING: TASK-CONDITIONED HYPERNETWORK APPROACH WITH THEORETICAL GUARANTEE

Anonymous authors
Paper under double-blind review

ABSTRACT

Unlearning in decentralized continual learning requires agents to unlearn previously acquired knowledge even when the corresponding historical data are unavailable. Performing unlearning without reconstructing historical inputs presents two challenges: (1) *Lack of Direct Unlearning Supervision*, as agents lack direct supervision for unlearning the target class while preserving the remaining classes of the same task; and (2) *Cross-Task Interference during Unlearning*, as local updates and neighbor aggregation may disrupt retained tasks through shared parameters, while broadly restricting parameter changes can hinder the requested unlearning. To address these challenges, we propose *Dec-UnHNet*, in which a hypernetwork generates task-specific classifier heads from descriptors built with diffusion-derived class prototypes. For a data-free unlearning request, surrogate features provide supervision for unlearning the target class, while hypernetwork-output regularization anchors generated classifiers for retained tasks without historical-sample replay. Theoretically, we rigorously prove that *Dec-UnHNet* achieves an $\mathcal{O}(1/\sqrt{nQ})$ convergence rate for the decentralized joint global empirical risk across all tasks. Extensive experiments on two benchmark datasets demonstrate improved continual-learning accuracy and zero target-class accuracy at all evaluated unlearning steps.

1 INTRODUCTION

Unlearning in decentralized continual learning (DCL) requires a network of agents to unlearn previously acquired knowledge while continuously learning from evolving private data and exchanging model information with neighboring agents. Such unlearning requests may arise from privacy or data-governance requirements (Voigt & Von dem Bussche, 2017; Bonta, 2022). Existing unlearning approaches include retraining and rehearsal (Xiong et al., 2023; Liu et al., 2022), gradient-ascent updates (Dhasade et al., 2024; Li et al., 2023), and fine-tuning (Shaik et al., 2024), with recent studies extending unlearning to decentralized systems (Qiao et al.; Wu et al., 2026). In DCL, however, an unlearning request may arrive after the corresponding historical samples have been discarded, while some agents may have acquired the relevant knowledge through neighbor exchange without ever possessing those samples. To recover supervision in this setting, DCU (Zhang et al., 2026) uses class prototypes to generate surrogate samples for continual learning and on-demand unlearning. Although this provides an alternative to accessing the original historical data, it still relies on generating and processing substitute samples to guide model updates. Consequently, avoiding original-data storage does not eliminate the computation required for sample synthesis and replay, or the dependence of the resulting supervision on how well the generated samples represent historical predictive behavior. This raises a further question: *how can a decentralized continual-learning system effectively unlearn historical knowledge and preserve other learned tasks without accessing the original data or reconstructing historical input samples?*

Addressing this problem requires resolving two key challenges. **(1) Lack of Direct Unlearning Supervision.** An unlearning request identifies the target class to be unlearned, but provides no direct supervision for updating the model. Historical data may become unavailable as tasks progress, while neighbor aggregation can transfer target-class knowledge to agents that never held the corresponding data. Without reconstructing historical inputs, agents must therefore derive unlearning updates from

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the available model and task information. The difficulty is to turn this information into effective class-specific updates that unlearn the target class while preserving the remaining classes of the same task. **(2) Cross-Task Interference during Unlearning.** When historical tasks share model parameters, local unlearning updates and neighbor aggregation may also change the predictions of tasks that should be retained. Without historical data or reconstructed inputs, agents cannot directly constrain these changes through losses or prediction consistency evaluated on the retained tasks. Restricting the overall change in shared parameters provides an alternative, but does not directly distinguish changes to target-task predictions from changes to retained-task predictions, and may hinder the requested unlearning. The difficulty is therefore to construct task-specific constraints from the available model and task information, so that changes to retained tasks can be limited while shared parameters adapt to the unlearning request.

To address these challenges, we propose *Dec-UnHNet*, a decentralized task-conditioned hypernetwork framework that represents historical knowledge through explicit task-to-parameter mappings rather than sample-level replay. Specifically, we employ a frozen diffusion model during learning to extract compact class prototypes, aggregate them among neighboring agents, and use them to construct task descriptors. The prototypes and descriptors are stored locally, while a shared hypernetwork maps each descriptor to the corresponding task-specific classifier parameters. Upon an unlearning request, we synthesize surrogate features using the pre-unlearning classifier to provide class-specific supervision without reconstructing historical inputs. These features guide the construction of target classifier parameters that unlearn the requested class while preserving the remaining classes of the same task. We then recompute the task descriptor without the target-class prototype and update the hypernetwork to align the generated classifier with these target parameters. Meanwhile, with the feature extractor fixed during unlearning, task-level output regularization constrains the classifier parameters generated for other historical tasks to remain close to their pre-unlearning values. Because each historical classifier can be generated from its stored task descriptor, these constraints can be evaluated without accessing or reconstructing historical inputs. Rather than directly restricting changes to the shared hypernetwork parameters, this regularization penalizes changes to individual retained classifiers, allowing the shared hypernetwork to adapt to the unlearning request while limiting cross-task interference. The resulting local hypernetwork updates are coordinated through neighbor aggregation to support collaborative data-free unlearning.

Theoretically, we bound the expected average gradient norm of the joint objective $F_{1:T}$ and make explicit how consensus drift, inter-task heterogeneity, and hypernetwork-specific approximation errors determine the resulting residual neighborhood. Extensive experiments on benchmark datasets demonstrate that our framework achieves superior performance in both sequential task acquisition and collaboratively targeted unlearning. Our main contributions can be summarized as follows:

- We propose *Dec-UnHNet*, a decentralized hypernetwork framework that realizes data-free unlearning in continual learning. It integrates hypernetworks with diffusion-based prototypes to decouple task knowledge and enable precise data-free unlearning without historical samples.
- We provide a convergence analysis for the decentralized joint objective that quantifies the residual effects of topology, task heterogeneity, and finite local updates.
- Extensive experiments on two datasets demonstrate that our model improves CIFAR-100 by 5.23% and Tiny-ImageNet by 4.91% over the best unlearning-capable baselines across all settings, while achieving zero target-class accuracy at all evaluated unlearning steps.

2 Related Work

Distributed unlearning. Distributed unlearning aims to remove the influence of selected data or clients from collaboratively trained models. Exact approaches typically require costly retraining or carefully designed training procedures; for example, Exact-Fun (Xiong et al., 2023) uses quantized federated learning to support efficient exact deletion. Approximate methods instead trade strict retraining equivalence for efficiency. QuickDrop (Dhasade et al., 2024) combines dataset distillation with stochastic gradient ascent to reduce unlearning cost, while SFU (Li et al., 2023) performs gradient ascent in the orthogonal complement of subspaces constructed from retained-client gradients, avoiding the storage of historical updates. FedAU (Gu et al., 2024) introduces a lightweight auxiliary module during training to enable efficient unlearning. FedLU (Zhu et al., 2023) further

studies unlearning in heterogeneous federated knowledge-graph learning through knowledge distillation and forgetting mechanisms. Recent work extends unlearning to fully decentralized systems. PDUDT (Qiao et al.) studies provable client unlearning under dynamic communication topologies, while certified decentralized unlearning (Wu et al., 2026) uses second-order corrective updates to eliminate propagated influence. Despite these developments, most existing frameworks assume stationary environments and struggle when knowledge is updated sequentially.

Continual learning with unlearning. Continual learning with unlearning requires models to acquire new knowledge while selectively removing previously learned information. CLPU-DER++ (Liu et al., 2022) combines rehearsal with dynamic task structures to support private removal of previously learned tasks, while LSF (Shibata et al., 2021) embeds class-specific mnemonic codes during training, enabling later class removal without accessing the original data. FedAF (Li et al., 2025) leverages memory generation to overwrite historical knowledge with synthetic memories. Recent research has shifted toward unified frameworks that balance the trade-off between learning and forgetting. For instance, CLMUL (Chatterjee et al., 2024) utilizes a unified contrastive distillation framework, and UG-CLU (Huang et al., 2025) proposes a manifold-aware optimization framework to maintain the structural integrity of the remaining knowledge. These unified approaches aim to mitigate the forgetting-interference dilemma, ensuring that the removal of a specific task does not inadvertently degrade the ability of the model to retain unrelated tasks.

3 PRELIMINARIES

We model the decentralized learning system as a set of n agents, indexed by $\mathcal{V} = \{1, \ldots, n\}$. The interaction is characterized by a time-varying, symmetric doubly stochastic matrix $W^q \in \mathbb{R}^{n \times n}$. A non-zero entry $W^q_{ij} > 0$ indicates a direct communication link between agents i and j at round q. Considering the dynamic of decentralized environments, the topology may change between rounds.

In continual learning with unlearning, all agents receive a sequence of T requests. The t-th request is represented as $\mathcal{R}_t = (\varsigma_t, C_t, D_t)$, where $\varsigma_t \in \{\mathbf{L}, \mathbf{U}\}$ denotes either a learning or an unlearning instruction. C_t and D_t depend on the request type. For a learning request, C_t denotes the cumulative class space acquired up to stage t, and $D_t = \{D_{t,i}\}_{i=1}^n$ collects the local datasets available to the agents at the current stage. Due to intra-task data heterogeneity, agent i only observes a local class subset $C_{t,i} \subseteq C_t$. For an unlearning request, C_t specifies the target class or class set to be unlearned, while $D_t = \emptyset$, indicating that the corresponding original training data are unavailable.

4 METHODOLOGY

This section presents the proposed hypernetwork framework for decentralized continual learning and data-free targeted unlearning. Section 4.1 defines the task-conditioned model, while Sections 4.2 and 4.3 describe continual learning and data-free targeted knowledge unlearning. The complete algorithm and procedural functions are provided in Appendix A.

4.1 TASK-CONDITIONED HYPERNETWORK ARCHITECTURE

The core of our approach is to represent task-specific knowledge through a controllable task-to-parameter mapping rather than directly maintaining a single set of task-dependent model parameters. For each agent, the local task model consists of a feature extractor E_{θ^f} and a task-specific classifier $g_{\theta^c_t}$. The feature extractor learns representations shared across the task sequence, while a local hypernetwork $\mathcal{H}_\phi$ generates the classifier parameters conditioned on a task descriptor z_t:

$$\theta^c_t = \mathcal{H}_\phi(z_t), \qquad f_t(x) = g_{\theta^c_t}(E_{\theta^f}(x)) \,. \tag{1}$$

The feature extractor parameters θ^f are updated by standard stochastic gradient descent, whereas ϕ is optimized to capture the mapping from task descriptors to task-specific classifier parameters. This design separates task-specific decision mappings from shared representations, making each task mapping explicitly addressable through its descriptor and thereby enabling targeted modification during data-free unlearning. Generating only the classifier head also reduces hypernetwork complexity and communication overhead compared with generating the entire model. Accordingly,

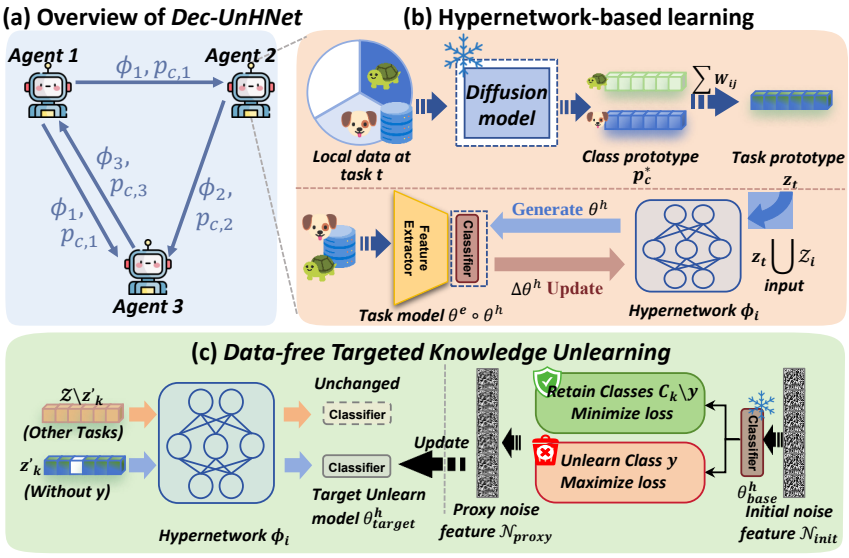

Figure 1: Overview of *Dec-UnHNet*. (a) In the decentralized setting, each agent communicates its hypernetwork and prototypes. (b) The hypernetwork produces task-specific classifier heads conditioned on diffusion-based prototypes. (c) Proxy-noise synthesis realigns the generated classifier mapping for data-free target-class unlearning. Embedded legacy notation uses θ^e for the current θ^f and θ^h for the current θ^c; "unlearn" in the artwork denotes the requested unlearning operation.

our unlearning mechanism operates on the target predictive mapping rather than claiming complete removal of target-related representations from the feature extractor. To provide a meaningful conditioning signal for this mapping, we next construct z_t by characterizing the task distribution through prototype extraction.

Following textual inversion in diffusion models, we use a frozen pre-trained denoising network ϵ_ψ. For each locally observed class $c \in C_{t,i}$, agent i learns a condition embedding $p_{c,i}$, called a class prototype, by minimizing the denoising reconstruction loss

$$p^*_{c,i} := \arg\min_p \mathbb{E}_{\substack{x\sim D^c_{t,i},\ \tau\sim\mathcal{U}(1,T) \\ \epsilon\sim\mathcal{N}(0,\mathbf{I})}} \left[\|\epsilon - \epsilon_\psi(x_\tau, \tau, p)\|_2^2\right]. \tag{2}$$

Here x is a clean local image in class c, $\tau \sim \mathcal{U}(1,T)$ is a uniformly sampled diffusion time step, and x_τ is obtained by injecting Gaussian noise ϵ into x. Thus $p^*_{c,i}$ exists in the presented construction only for $c \in C_{t,i}$. The learned prototype encodes the class concept in the conditioning space of the frozen diffusion model.

Due to the inherent data heterogeneity in decentralized environments, local prototypes generated by a single agent may be biased by its non-Independently Identically Distributed (non-IID) data. To obtain more robust class representations, agents aggregate their local prototypes through decentralized consensus. Starting from the locally extracted prototype $p^*_{c,i}$, agent i iteratively mixes prototype information with its neighbors according to the communication matrices, and denotes the resulting aggregate for class c as $\bar{p}_{c,i}$. Based on the aggregated class prototypes, agent i constructs the task descriptor as

$$z_{t,i} = \frac{1}{|C_t|} \sum_{c\in C_t} \bar{p}_{c,i}. \tag{3}$$

We omit the agent index and write z_t when no ambiguity arises. The aggregated class prototypes and task descriptor are stored locally, enabling subsequent unlearning requests to reconstruct a modified task descriptor without accessing historical training data. The resulting z_t serves as the conditioning signal for the hypernetwork to generate the corresponding task-specific classifier parameters.

4.2 HYPERNETWORK-BASED CONTINUAL LEARNING

When agent i receives a learning request ($\varsigma_t = \mathbf{L}$), it optimizes its across-task feature extractor θ^f and the hypernetwork parameters ϕ. At communication round q, the generated head $\theta^c_{q,i,0} = \mathcal{H}_{\phi_{i,q}}(z_{t,i})$ is paired with the current extractor. At local step k, the agent samples $\xi_{q,i,k} \sim D_{t,i}$ and updates both components of the task model:

$$\theta_{q,i,k+1} = \theta_{q,i,k} - \beta \nabla_\theta g_i(\theta_{q,i,k}, \xi_{q,i,k}). \tag{4}$$

To update the hypernetwork, we measure the local adaptation of the classification head by $\Delta\theta^c_{q,i} = \theta^c_{q,i,K} - \theta^c_{q,i,0}$, where $\theta^c_{q,i,0} = \mathcal{H}_\phi(z_t)$ is generated by the current hypernetwork and $\theta^c_{q,i,K}$ is obtained after K local optimization steps. Leveraging the chain rule, we can approximate the gradient of the hypernetwork loss,

$$\nabla_\phi F_i(\phi) = \frac{\partial F_i}{\partial \theta^c_i} \frac{\partial \mathcal{H}_\phi(z_{t,i})}{\partial \phi} \approx (\Delta\theta^c_{q,i})^\top \nabla_\phi \mathcal{H}_\phi(z_{t,i}). \tag{5}$$

To mitigate catastrophic forgetting, we integrate task-level regularization into the optimization objective. Each agent stores a pre-update snapshot ϕ^* and applies a virtual update $\phi' = \phi - \alpha \nabla_\phi F_i$ to gain the newly current hypernetwork. Then the output-regularization loss penalizes changes to the generated classifier parameters for retained descriptors:

$$l(\phi') = \sum_{z_j \in \mathcal{Z}_i \setminus \{z_{t,i}\}} \|\mathcal{H}_{\phi'}(z_j) - \mathcal{H}_{\phi^*}(z_j)\|^2. \tag{6}$$

This process effectively performs knowledge replay at the task level, allowing the network to retain historical classification logic within the parameter space without the need for raw data rehearsal or storage. Thus, the whole local continual-learning update is

$$\phi_{i,q+\frac{1}{2}} = \phi_{i,q} - \alpha(\nabla_\phi F_i(\phi) + \gamma_{cl} \nabla_\phi l(\phi')), \tag{7}$$

where γ_{cl} controls retained-task regularization during learning. After this local update, agent i aggregates the hypernetwork weights received from its neighbors according to W^q,

$$\phi_i^{q+1} \leftarrow \sum_{j \in \mathcal{V}} W^q_{ij} \phi^a_j$$

4.3 DATA-FREE TARGETED KNOWLEDGE UNLEARNING

Upon receiving an unlearning request ($\varsigma_t = \mathbf{U}$) to unlearn class y from a previously learned task k, agent i no longer has access to the corresponding historical training samples. We therefore propose a Data-Free Targeted Knowledge Unlearning strategy that unlearns the target predictive mapping without revisiting the original data. Specifically, agent i retrieves the stored task descriptor $z_{k,i}$ and the corresponding class prototypes, and constructs a modified descriptor $z'_{k,i}$ by excluding the prototype associated with class y. The feature extractor remains unchanged during unlearning, while the task-specific mapping generated by the hypernetwork is selectively realigned.

To derive the desired classifier state without historical data, we synthesize surrogate features through an error min–max optimization. Starting from random features $\varepsilon \sim \mathcal{N}(0, \mathbf{I})$, we optimize them against the base classifier $\theta^c_{\text{base}} = \mathcal{H}_{\phi^*}(z_{k,i})$ within the bounded feature space $\mathcal{B}_R = \{\varepsilon : \|\varepsilon\|_2 \leq R\}$. For each retained class c, the corresponding surrogate feature is optimized to minimize the classification loss, whereas the surrogate feature associated with the target class y is optimized to maximize the loss:

$$\varepsilon_c^* = \arg\min_{\varepsilon_c \in \mathcal{B}_R} g_i(\theta^c_{\text{base}}, (\varepsilon_c, c)), \quad c \neq y, \qquad \varepsilon_y^* = \arg\max_{\varepsilon_y \in \mathcal{B}_R} \left[g_i(\theta^c_{\text{base}}, (\varepsilon_y, y)) - \lambda \|\varepsilon_y\|_2 \right]. \tag{8}$$

The retained-class proxies characterize predictive responses that should be preserved, while the target-class proxy provides an opposing signal for unlearning the requested class. These surrogate features are subsequently used to locally adapt the base classifier, producing target parameters θ^c_{target} that unlearn the target class while maintaining the predictive mappings of the retained classes. The detailed iterative optimization procedure is provided in Appendix Algorithm 2.

We then realign the hypernetwork so that the modified descriptor $z'_{k,i}$ generates the target classifier θ^c_{target}. Specifically, we define the mapping-alignment objective as

$$F_{\text{UL}}(\phi) = \frac{1}{2} \left\| \mathcal{H}_\phi(z'_{k,i}) - \theta^c_{\text{target}} \right\|_2^2, \tag{9}$$

whose gradient is

$$\nabla_\phi F_{\text{UL}} = \left[\mathcal{H}_\phi(z'_{k,i}) - \theta^c_{\text{target}} \right]^\top \nabla_\phi \mathcal{H}_\phi(z'_{k,i}). \tag{10}$$

Minimizing this objective explicitly moves the classifier generated from the modified descriptor toward the desired unlearned state, thereby realigning the task-to-parameter mapping rather than directly modifying historical model parameters.

To prevent the unlearning update from interfering with other previously learned tasks, we further employ the same task-level meta-regularization used during continual learning. After the virtual update $\phi' = \phi - \eta \nabla_\phi F_{\text{UL}}$, the generated classifiers of unaffected task descriptors are constrained to remain close to those produced by the fixed pre-unlearning hypernetwork:

$$l(\phi') = \sum_{z_j \in \mathcal{Z}_i \setminus \{z_{k,i}\}} \left\| \mathcal{H}_{\phi'}(z_j) - \mathcal{H}_{\phi^*}(z_j) \right\|_2^2. \tag{11}$$

The hypernetwork parameters are then updated by

$$\phi \leftarrow \phi - \eta \left(\nabla_\phi F_{\text{UL}} + \gamma_{\text{ul}} \nabla_\phi l(\phi') \right), \tag{12}$$

where η is the unlearning step size and γ_{ul} controls the preservation of unaffected task mappings. The updated hypernetwork is subsequently propagated through decentralized consensus, enabling agents to consistently unlearn the requested predictive mapping across the network. After unlearning, the modified descriptor and retained class prototypes replace the original task record for subsequent requests. Since the shared feature extractor is not modified during this process, our unlearning objective specifically targets the deployed predictive mapping and does not claim complete removal of all target-correlated representations from the feature extractor.

5 Convergence Analysis

To analyze the algorithm from a global perspective, we define $\Phi_{t,q} := [\phi_{t,1,q} \cdots \phi_{t,n,q}]$ as the concatenation of the hypernetwork parameters of all agents at communication round q of task t, and denote their average by $\bar{\phi}_{t,q} = \Phi_{t,q} \mathbf{1}_n / n$. For agent i, the local objective on task t is defined as

$$F_{t,i}(\phi) := \mathbb{E}_{(x,y)\sim D_{t,i}} \left[\ell \Big(g_{\mathcal{H}_\phi(z_t)} \Big(E_{\theta_i^f}(x) \Big), y \Big) \right], \tag{13}$$

where the feature extractor parameters θ_i^f are treated as fixed during the hypernetwork update. Accordingly, we define the stacked local gradients as

$$\partial F_t(\Phi_{t,q}) := [\nabla_\phi F_{t,1}(\phi_{t,1,q}) \cdots \nabla_\phi F_{t,n}(\phi_{t,n,q})]. \tag{14}$$

The global objective for task t is the average of the local objectives, i.e., $F_t(\phi) = \frac{1}{n} \sum_{i=1}^n F_{t,i}(\phi)$. Therefore, after sequentially learning T tasks, the decentralized continual learning problem is formulated as

$$\min_{\phi_T} F_{1:T}(\phi_T) := \min_{\phi_T} \sum_{t=1}^T F_t(\phi_T) = \min_{\phi_T} \frac{1}{n} \sum_{t=1}^T \sum_{i=1}^n F_{t,i}(\phi_T). \tag{15}$$

We next bound the expected average squared gradient norm of the global objective. The assumptions and complete proof are provided in Appendix B.

Theorem 1 *Under Assumptions 1–6 stated in Appendix B, let* $A_1 = 1 - \frac{27\alpha^2 n L_\phi^2}{(1-\sqrt{\rho})^2} > 0$ *and* $\Delta_f = \mathbb{E}[F_{1:T}(\frac{\Phi_{T,0}\mathbf{1}_n}{n})] - \mathbb{E}[F_{1:T}(\frac{\Phi_{T,Q}\mathbf{1}_n}{n})]$. *The expected average squared gradient norm satisfies*

$$\frac{1}{Q} \sum_{q=0}^{Q-1} \mathbb{E} \left\| \nabla F_{1:T} \left(\frac{\Phi_{T,q} \mathbf{1}_n}{n} \right) \right\|^2 \le \frac{2\Delta_f}{\alpha Q} + 3\mathbb{E} \left\| \nabla F_{1:T-1} \left(\frac{\Phi_{T,0} \mathbf{1}_n}{n} \right) \right\|^2 \tag{16}$$

$$+\left((T-1)+\frac{27\alpha^2(T+1)L_\phi^2 n}{(1-\sqrt{\rho})^2 A_1}\right)\frac{1}{Q}\sum_{q=0}^{Q-1}\mathbb{E}\left\|\nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\right\|^2$$

$$+3\alpha L_\phi K^2\frac{1}{Q}\sum_{q=0}^{Q-1}\mathbb{E}\left\|\frac{\partial F_T(\Phi_{T,q})\mathbf{1}_n}{n}\right\|^2+\frac{3\alpha^2 L_\phi^2(T-1)^2K^2\Gamma}{n}+\frac{3\alpha L_\phi K^2\beta^2 b_2^2\delta^2}{n}$$

$$+\frac{3K^2\alpha L_\phi\Gamma}{n}+\frac{3\alpha^2(T+1)L_\phi^2 nK^2\Gamma}{(1-\rho)A_1}+\frac{27\alpha^2(T+1)L_\phi^2 n}{(1-\sqrt{\rho})^2A_1}\sigma_g^2+T^2\sigma_t^2.$$

Here n, T, and Q denote the numbers of agents, sequential tasks, and communication rounds, respectively. Consistent with the algorithm, α is the decentralized hypernetwork update step size, β is the local update step size, and K is the number of local steps. The topology factor is $\rho := \max_q\big(\max\{|\lambda_2(W^q)|,|\lambda_n(W^q)|\}\big)^2$, where $\lambda_j(W^q)$ is the j-th largest eigenvalue of W^q. The constant L_ϕ is the smoothness parameter, b_2 bounds the hypernetwork gradients, and δ^2, σ_g^2, and σ_t^2 denote stochastic-gradient variance, intra-task gradient heterogeneity, and inter-task gradient heterogeneity, respectively. Moreover, Γ is the finite-local-step approximation-error bound from Appendix Lemma 2.

Remark 1 (Consensus Drift) *The convergence bound in Theorem 1 is significantly influenced by the spectral gap of the time-varying communication matrix W^q, represented by terms involving ρ. This illustrates the impact of consensus drift in a decentralized setting. As the topology becomes sparser, that is spectral gap ρ nears to 1, the information propagation among agents slows down, leading to a larger discrepancy between local hypernetwork parameters $\phi_{T,q,i}$ and the global average. This theoretically justifies that maintaining a well-connected agent graph is important for ensuring that the hypernetwork remains consistent across the entire system.*

Remark 2 (Cumulative Task Interference in Continual Learning) *A distinct feature of our result is the dependence on the number of tasks T and the inter-task gradient heterogeneity σ_t^2. The term $T^2\sigma_t^2$ highlights that as the agent sequentially learns more tasks, the cumulative discrepancy between task-specific decision boundaries can hinder the convergence of the shared hypernetwork parameters Φ. Furthermore, the presence of the historical gradient term $\|\nabla F_{1:T-1}\left(\frac{\Phi_{T,0}\boldsymbol{I}_n}{n}\right)\|^2$ indicates that as preceding tasks fit well, this residual error diminishes, thereby effectively tightening the convergence upper bound for the current task T.*

Remark 3 (Hypernetwork Generation Error) *Different from the convergence result of the classical decentralized parallel stochastic gradient descent algorithm, our Dec-UnHNet uses multiple local gradient steps on the agents and applies the chain rule to approximate the local hypernetwork gradient, which introduces an additional error Γ.*

Corollary 1 *Under Assumptions 1-5, let $\alpha=\sqrt{n/Q}$. When $Q \geq \frac{54L_\phi^2 n^2}{(1-\sqrt{\rho})^2}$ is sufficient large, the convergence rate of Dec-UnHNetcan be expressed as follows:*

$$\frac{1}{Q}\sum_{q=0}^{Q-1}\mathbb{E}\left\|\nabla F_{1:T}\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\right\|^2 \leq \mathcal{O}\left(\frac{1}{\sqrt{nQ}}+\frac{1}{Q}\right) \tag{17}$$

Remark 4 *Corollary 1 provides the simplified convergence rate of $\mathcal{O}(\frac{1}{\sqrt{nQ}}+\frac{1}{Q})$, demonstrating that our proposed algorithm achieves a linear speedup with respect to the number of participating agents n. This is a significant result for decentralized systems, as it proves that despite the complexities of continual task growth and the approximation overhead of the hypernetwork, the system effectively aggregates the computational power of all agents. To achieve an ϵ-stationary point, the total communication complexity scales at $\mathcal{O}(1/n\epsilon^2)$.*

6 Experiment

6.1 Setup

We evaluate CIFAR-100 (Krizhevsky et al., 2009) with VGG16 (Simonyan & Zisserman, 2014) and Tiny-ImageNet (Le & Yang, 2015) with ResNet18 (He et al., 2016). A Dirichlet split $Dir(\lambda)$ with

Table 1: Accuracy on *CIFAR-100* with *VGG16* for the request sequence ① $cl_{0:9}$ (learn classes 0-9) → ② ul_2 (unlearn class 2) → ③ $cl_{10:19}$ (learn classes 10-19) → ④ ul_0 (unlearn class 0) → ⑤ $cl_{20:29}$ (learn classes 20-29) → ⑥ ul_{29} (unlearn class 29).

Algorithm	$Dir(\lambda = 0.5)$						$Dir(\lambda = 1.0)$					
	$cl_{0:9}$ r↑%	ul_2 r↑/u↓%	$cl_{10:19}$ r↑%	ul_0 r↑/u↓%	$cl_{20:29}$ r↑%	ul_{29} r↑/u↓%	$cl_{0:9}$ r↑%	ul_2 r↑/u↓%	$cl_{10:19}$ r↑%	ul_0 r↑/u↓%	$cl_{20:29}$ r↑%	ul_{29} r↑/u↓%
continual learning												
FT	42.91	-	29.44	-	24.17	-	67.15	-	36.37	-	26.88	-
EWC	43.16	-	28.82	-	23.68	-	69.53	-	35.89	-	28.39	-
LwF	40.47	-	29.55	-	26.12	-	69.36	-	33.90	-	31.01	-
HNET	69.68	-	41.03	-	31.01	-	71.85	-	40.28	-	34.50	-
continual learning with unlearning												
FedAF	43.16	30.24/0.70	30.53	33.68/0.00	25.06	20.92/0.00	69.53	45.38/0.00	36.09	39.13/0.00	31.62	30.09/0.00
EWC												
+FedAU	43.16	31.57/0.00	29.35	29.98/0.00	25.20	26.13/0.00	69.53	62.32/0.00	40.12	39.08/0.00	30.90	28.92/0.00
+QuickDropSGA	43.16	29.01/0.00	31.66	20.51/0.00	24.90	12.58/0.00	69.53	63.80/0.00	39.12	31.25/0.00	33.88	30.70/0.00
+BadTeacher	43.16	38.28/1.20	18.98	11.29/0.00	14.55	16.77/10.90	69.74	50.20/12.50	26.58	24.94/0.70	22.93	30.23/6.60
LwF												
+FedAU	40.47	34.73/0.00	29.62	28.33/0.00	26.77	26.75/0.00	69.36	64.67/0.00	37.36	30.99/0.00	30.62	28.08/0.00
+QuickDropSGA	40.47	33.26/0.00	26.16	18.54/0.00	17.91	10.57/0.00	69.36	66.32/3.10	30.00	21.87/0.00	17.56	13.78/0.00
+BadTeacher	40.47	19.54/0.00	15.02	10.94/3.50	14.95	11.90/0.20	69.36	56.84/9.90	30.05	26.04/0.00	23.85	28.22/11.00
UniCLUN	38.42	12.09/0.00	33.25	11.27/0.00	20.85	6.96/0.00	68.31	20.98/2.40	40.74	13.79/0.00	38.13	18.33/0.00
UG-CLU	55.95	52.89/0.00	39.06	27.42/0.00	29.95	30.13/0.00	65.72	64.62/0.00	**45.56**	36.62/0.00	38.88	29.15/0.00
Ours	**69.68**	**60.97/0.00**	**40.88**	**39.22/0.00**	**38.35**	**36.29/0.00**	**71.85**	**67.82/0.00**	43.67	**43.50/0.00**	**41.85**	**40.64/0.00**

$\lambda \in \{0.5, 1.0\}$ partitions each task non-IID across $n = 10$ agents. Communication uses a doubly stochastic mixing matrix constructed from an Erdős–Rényi graph $G(n, p)$ with $p = 0.5$. Experiments are conducted using momentum SGD with a learning rate of 0.01, a momentum parameter of 0.9, a weight decay of 5e-4. We use 50 communication rounds, 5 local epochs per round, a batch size of 128 per agent, and $\gamma_{cl} = \gamma_{ul} = 1$.

For learning requests, we compare fine-tuning (FT) (Lian et al., 2018), EWC (Kirkpatrick et al., 2017), LwF (Li & Hoiem, 2017) and HNET (von Oswald et al., 2020). For unlearning requests, we evaluate FedAF (Li et al., 2025) and combine EWC or LwF with FedAU (Gu et al., 2024), QuickDropSGA (Dhasade et al., 2024), and BadTeacher (Chundawat et al., 2023), as indicated by the indented rows in Tables 1 and 2. We additionally compare UniCLUN and UG-CLU (Huang et al., 2025), with the latter adapted to decentralized execution using D-PSGD (Lian et al., 2018). For data-dependent baselines requiring retained or target-class historical samples, we provide a historical-data buffer under a common protocol. Other details are given in Appendix C.

6.2 Evaluation

Tables 1 and 2 compare *Dec-UnHNet* with the baselines on CIFAR-100 and Tiny-ImageNet across both learning and unlearning requests under two non-IID settings. Averaged over the reported settings, *Dec-UnHNet* improves accuracy over the best unlearning-capable baselines by 5.23% and 4.91%, respectively. At the final unlearning phase ul_{29}, it outperforms the best baselines by up to 9.94% on CIFAR-100 and 8.20% on Tiny-ImageNet.

Dec-UnHNet also achieves a favorable unlearning–retention trade-off without historical samples. Its target-class accuracy is 0.00% at every evaluated unlearning step; at ul_{29} with $Dir(\lambda = 0.5)$, it retains 36.29% accuracy on CIFAR-100 versus 6.96% for UniCLUN, and 26.04% on Tiny-ImageNet versus 17.84% for LwF+BadTeacher. This behavior is consistent with task-conditioned classifier generation concentrating the update on the requested mapping while regularization limits changes to retained mappings. Unlike data-dependent baselines, *Dec-UnHNet* uses synthesized features to perform this update after historical data have been discarded. The reported target-class results characterize empirical unlearning performance.

Across the two non-IID settings, *Dec-UnHNet* further exhibits stable retained accuracy. On CIFAR-100 at ul_{29}, its gap between $Dir(\lambda = 0.5)$ and $Dir(\lambda = 1.0)$ is 4.35 percentage points, compared with more than 9 points for FedAF. Under the common request sequence and topology, this smaller cross-split gap suggests lower sensitivity to the tested degree of agent heterogeneity. This observa-

Table 2: Accuracy on *Tiny-ImageNet* with *ResNet18* for the request sequence ① $cl_{0:9}$ (learn classes 0-9) → ② $cl_{10:19}$ (learn classes 10-19) → ③ ul_0 (unlearn class 0) → ④ $cl_{20:29}$ (learn classes 20-29) → ⑤ ul_{20} (unlearn class 20) → ⑥ ul_{29} (unlearn class 29).

Algorithm	$Dir(\lambda = 0.5)$						$Dir(\lambda = 1.0)$					
	$cl_{0:9}$ $r\uparrow\%$	$cl_{10:19}$ $r\uparrow\%$	ul_0 $r\uparrow/u\downarrow\%$	$cl_{20:29}$ $r\uparrow\%$	ul_{20} $r\uparrow/u\downarrow\%$	ul_{29} $r\uparrow/u\downarrow\%$	$cl_{0:9}$ $r\uparrow\%$	$cl_{10:19}$ $r\uparrow\%$	ul_0 $r\uparrow/u\downarrow\%$	$cl_{20:29}$ $r\uparrow\%$	ul_{20} $r\uparrow/u\downarrow\%$	ul_{29} $r\uparrow/u\downarrow\%$
continual learning												
FT	38.80	24.24	-	13.12	-	-	46.28	27.72	-	16.51	-	-
EWC	35.92	24.00	-	15.49	-	-	48.70	28.04	-	21.15	-	-
LwF	38.52	29.33	-	17.99	-	-	48.94	33.54	-	25.41	-	-
HNET	40.32	34.87	-	27.23	-	-	50.42	41.35	-	30.81	-	-
continual learning with unlearning												
FedAF	35.92	24.00	28.28/0.00	16.48	12.27/13.00	13.01/0.00	48.70	28.04	29.98/0.60	21.92	18.40/16.60	19.58/0.60
EWC												
+FedAU	35.92	24.00	27.68/0.00	16.06	17.02/0.00	17.07/0.00	48.70	28.04	28.81/0.00	21.58	21.06/0.00	20.33/0.00
+QuickDropSGA	35.92	24.00	24.36/0.00	15.08	14.16/0.00	12.12/0.00	48.70	28.04	26.66/0.00	22.52	19.98/0.00	17.07/0.00
+BadTeacher	37.84	22.57	16.24/1.20	10.20	12.43/11.60	10.73/0.00	47.38	28.20	20.10/4.80	21.75	20.95/3.40	23.85/0.60
LwF												
+FedAU	38.52	29.33	30.13/0.00	17.09	16.81/0.00	16.40/0.00	48.94	33.54	31.11/0.00	24.87	23.94/0.00	22.61/0.00
+QuickDropSGA	38.52	29.33	24.09/0.00	19.48	19.45/0.00	17.30/0.00	48.94	33.54	26.32/0.00	24.95	22.74/0.00	17.41/0.00
+BadTeacher	38.52	29.33	21.96/7.40	15.08	15.94/7.40	17.84/1.60	48.94	33.54	27.06/6.20	20.42	20.84/4.60	23.27/0.40
UniCLUN	38.52	28.79	17.69/0.00	22.00	16.39/1.00	13.52/0.00	48.94	34.37	26.82/0.00	29.84	24.11/14.00	23.12/0.00
UG-CLU	34.36	24.61	20.04/0.00	15.22	15.43/0.00	17.67/0.00	40.36	28.25	27.22/0.00	19.30	20.48/0.00	22.56/0.00
Ours	**40.32**	**34.87**	**31.16/0.00**	**28.05**	**27.11/0.00**	**26.04/0.00**	**50.42**	**41.35**	**37.53/0.00**	**31.03**	**31.69/0.00**	**28.82/0.00**

tion is consistent with the shared task embedding enabling the hypernetwork to integrate information across agents when generating task-specific parameters.

Hyperparameter ablation. In *Dec-UnHNet*, γ_{cl} weights the regularization that preserves retained-task mappings during sequential learning, whereas γ_{ul} protects unaffected mappings during unlearning. We evaluate both parameters on CIFAR-100 with $Dir(\lambda = 0.5)$ following the request sequence in Table 1. Figure 2 reports retained-class accuracy after ③ $cl_{10:19}$ for γ_{cl} and after ④ ul_0 for γ_{ul}. As γ_{cl} increases from 0.0 to 1.0, accuracy rises from 34.75% to 40.88%, a gain of 6.13 percentage points. Increasing γ_{ul} over the same range yields a smaller but monotonic gain from 37.44% to 39.22%. Thus, retention is more sensitive to γ_{cl} in this configuration, while γ_{ul} consistently reduces interference with unaffected mappings during unlearning. These configuration-specific trends do not establish a causal mechanism.

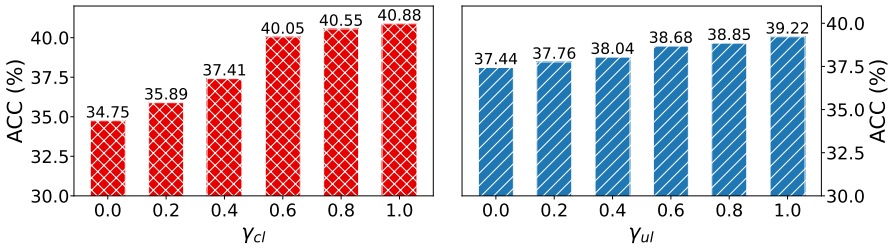

Figure 2: Hyperparameter analysis of γ_{cl} and γ_{ul}.

7 Conclusion

In this paper, we addressed the challenge of data-free unlearning in decentralized continual learning environments with *Dec-UnHNet*. By integrating diffusion-based prototypes and a Data-Free Targeted Knowledge Unlearning strategy, our approach enables agents to unlearn specified target classes without historical samples. Furthermore, we theoretically established a convergence rate of $\mathcal{O}(1/\sqrt{nQ})$, quantifying the impacts of consensus drift and task heterogeneity. Extensive experiments demonstrated that *Dec-UnHNet*achieved strong performance in both collaborative task acquisition and targeted unlearning, providing a scalable data-free solution for evolving distributed intelligence systems.

AI USE STATEMENT

Generative AI tools were used for language editing, structural revision, and literature cross-checking. No experimental values were generated by these tools. The authors reviewed the final text and take responsibility for its content.

REPRODUCIBILITY STATEMENT

We have made efforts to support the reproducibility of both the proposed method and the reported experimental results. The main paper provides the complete methodological description, including the overall framework, learning and unlearning procedures, and the definitions of the optimization objectives. The appendix provides detailed algorithms and implementation procedures, together with the assumptions and complete proofs for the theoretical results. The experimental section and appendix describe the datasets, data processing procedures, model architectures, baseline methods, evaluation protocols, and training configurations used in our experiments.

REFERENCES

Rob Bonta. California consumer privacy act (ccpa). *Retrieved from State of California Department of Justice: https://oag. ca. gov/privacy/ccpa*, 2022.

Romit Chatterjee, Vikram Chundawat, Ayush Tarun, Ankur Mali, and Murari Mandal. A unified framework for continual learning and machine unlearning. *arXiv e-prints*, pp. arXiv–2408, 2024.

Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli. Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7210–7217, 2023.

Akash Dhasade, Yaohong Ding, Song Guo, Anne-Marie Kermarrec, Martijn de Vos, and Leijie Wu. Quickdrop: Efficient federated unlearning via synthetic data generation. In *Proceedings of the 25th International Middleware Conference*, pp. 266–278, 2024.

Hanlin Gu, Gongxi Zhu, Jie Zhang, Xinyuan Zhao, Yuxing Han, Lixin Fan, and Qiang Yang. Unlearning during learning: an efficient federated machine unlearning method. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 4035–4043, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Zhehao Huang, Xinwen Cheng, Jie Zhang, Jinghao Zheng, Haoran Wang, Zhengbao He, Tao Li, and Xiaolin Huang. A unified gradient-based framework for task-agnostic continual learning-unlearning. *arXiv preprint arXiv:2505.15178*, 2025.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Guanghao Li, Li Shen, Yan Sun, Yue Hu, Han Hu, and Dacheng Tao. Subspace based federated unlearning. *arXiv preprint arXiv:2302.12448*, 2023.

Yuyuan Li, Jiaming Zhang, Yixiu Liu, and Chaochao Chen. Class-wise federated unlearning: Harnessing active forgetting with teacher–student memory generation. *Knowledge-Based Systems*, 316:113353, 2025.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in Neural Information Processing Systems 30*, 8:5331–5341, 2018.

Bo Liu, Qiang Liu, and Peter Stone. Continual learning and private unlearning. In *Conference on Lifelong Learning Agents*, pp. 243–254. PMLR, 2022.

Jing Qiao, Yu Liu, Zengzhe Chen, Mingyi Li, Yuan Yuan, Xiao Zhang, and Dongxiao Yu. Pdudt: Provable decentralized unlearning under dynamic topologies. In *Forty-second International Conference on Machine Learning*.

Thanveer Shaik, Xiaohui Tao, Lin Li, Haoran Xie, Taotao Cai, Xiaofeng Zhu, and Qing Li. Framu: Attention-based machine unlearning using federated reinforcement learning. *IEEE Transactions on Knowledge and Data Engineering*, 36(10):5153–5167, 2024.

Takashi Shibata, Go Irie, Daiki Ikami, and Yu Mitsuzumi. Learning with selective forgetting. In *IJCAI*, volume 3, pp. 4, 2021.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A practical guide, 1st ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

Johannes von Oswald, Christian Henning, Benjamin F. Grewe, and João Sacramento. Continual learning with hypernetworks. In *International Conference on Learning Representations*, 2020.

Hengliang Wu, Youming Tao, Anhao Zhou, Shuzhen Chen, Falko Dressler, and Dongxiao Yu. Certified unlearning in decentralized federated learning. *arXiv preprint arXiv:2601.06436*, 2026.

Zuobin Xiong, Wei Li, Yingshu Li, and Zhipeng Cai. Exact-fun: an exact and efficient federated unlearning approach. In *2023 IEEE International Conference on Data Mining (ICDM)*, pp. 1439–1444. IEEE, 2023.

Xiao Zhang, Zengzhe Chen, Mingyi Li, Jing Qiao, Fuzhen Zhuang, YUAN YUAN, and Dongxiao Yu. Forgetting whenever you want: A decentralized continual learning framework with on-demand unlearning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=pSOiiUSOzp.

Xiangrong Zhu, Guangyao Li, and Wei Hu. Heterogeneous federated knowledge graph embedding learning and unlearning. In *Proceedings of the ACM web conference 2023*, pp. 2444–2454, 2023.

A ALGORITHMIC DETAILS

Algorithm 1: *Dec-UnHNet*: Global Decentralized Orchestration

Input: Request sequence $\mathcal{R} * t * t = 1^T$, where $\mathcal{R} * t = (\varsigma_t, C_t, D * t, i * i \in \mathcal{V})$; mixing matrices $W^q * q = 0^{Q-1}$.

Output: Final collaborative hypernetwork weights ${\phi_{T,i}}_{i=1}^{n}$.

Initialize $\phi_{0,i}$, θ_i^f, $\mathcal{I}_i$, and $\mathcal{Z}_i$ for all $i \in \mathcal{V}$;
for $t = 1$ **to** T **do**
  Receive request $\mathcal{R}_t = (\varsigma_t, C_t, D_t)$;
  if $\varsigma_t = \mathbf{L}$ **then**
    $(\mathcal{I}_i[t], z_{t,i}) \leftarrow \text{TrainingPrototype}(D_{t,i}, C_t)$ for all i in parallel
    Store $\mathcal{Z}_i[t] \leftarrow z_{t,i}$
  else
    Identify target task k and class y
    Retrieve $z_{k,i}$ and construct $z'_{k,i}$ by excluding the prototype of class y
  end
  for $q = 0$ **to** $Q - 1$ **do**
    if $\varsigma_t = \mathbf{L}$ **then**
      $(\phi_{i,q+\frac{1}{2}}, \theta_i^f) \leftarrow \text{CL}(\phi_{i,q}, \theta_i^f, D_{t,i}, z_{t,i}, \mathcal{Z}_i)$ for all i in parallel
    else
      $\phi_{i,q+\frac{1}{2}} \leftarrow \text{UL}(\phi_{i,q}, z_{k,i}, z'_{k,i}, y, \mathcal{Z}_i)$ for all i in parallel
    end
    $\phi_{i,q+1} \leftarrow \sum_{j \in \mathcal{V}} W_{ij}^q \phi_{j,q+\frac{1}{2}}$ for all i in parallel
  end
  $\phi_{t,i} \leftarrow \phi_{i,Q}$ for all i
  if $\varsigma_t = \mathbf{U}$ **then**
    Update $\mathcal{I}_i[k]$ by removing class y and set $\mathcal{Z}_i[k] \leftarrow z'_{k,i}$
  end
end

Algorithm 2: *Dec-UnHNet*: Procedural Functions

Function *CL*$(\phi, \phi^*, \theta^f, D, z_t, \mathcal{Z})$
  $\theta_0^c \leftarrow \mathcal{H}_\phi(z_t)$
  $\theta_0 \leftarrow (\theta^f, \theta_0^c)$
  for $k = 0$ **to** $K - 1$ **do**
    Sample $\xi_k \sim D$
    $\theta_{k+1} \leftarrow \theta_k - \beta \nabla_\theta g_i(\theta_k, \xi_k)$
  $\delta\theta^c \leftarrow \theta_K^c - \theta_0^c$
  $\nabla_\phi F_i \leftarrow (-\delta\theta^c)^\top \nabla_\phi \mathcal{H}_\phi(z_t)$
  $\phi' \leftarrow \phi - \alpha \nabla_\phi F_i$ `// Virtual update`
  $\nabla_\phi l(\phi') \leftarrow \nabla_\phi \sum_{z_j \in \mathcal{Z} \setminus \{z_t\}} \|\mathcal{H}_{\phi'}(z_j) - \mathcal{H}_{\phi^*}(z_j)\|_2^2$
  $\phi \leftarrow \phi - \alpha \left(\nabla_\phi F_i + \gamma_{cl} \nabla_\phi l(\phi')\right)$
  $\theta^f \leftarrow \theta_K^f$
  return ϕ, θ^f

Function *UL*$(\phi, \phi^*, z_k, z'_k, y, C_k^{\text{ret}}, R, \mathcal{Z})$
  $\theta_{\text{base}}^c \leftarrow \mathcal{H}_{\phi^*}(z_k)$
  $\mathcal{N}_{\text{proxy}} \leftarrow \text{FeatureSynthesis}(\theta_{\text{base}}^c, y, C_k^{\text{ret}}, R)$
  $\theta_{\text{target}}^c \leftarrow \text{LocalUpdate}(\theta_{\text{base}}^c, \mathcal{N}_{\text{proxy}})$
  for *alignment iterations* **do**
    $\nabla_\phi F_{\text{UL}} \leftarrow \left[\mathcal{H}_\phi(z'_k) - \theta_{\text{target}}^c\right]^\top \nabla_\phi \mathcal{H}_\phi(z'_k)$
    $\phi' \leftarrow \phi - \eta \nabla_\phi F_{\text{UL}}$ `// Virtual update`
    $\nabla_\phi l(\phi') \leftarrow \nabla_\phi \sum_{z_j \in \mathcal{Z} \setminus \{z_k\}} \|\mathcal{H}_{\phi'}(z_j) - \mathcal{H}_{\phi^*}(z_j)\|_2^2$
    $\phi \leftarrow \phi - \eta \left(\nabla_\phi F_{\text{UL}} + \gamma_{ul} \nabla_\phi l(\phi')\right)$
  return ϕ

B Convergence Analysis

B.1 Assumptions and Auxiliary Results

Assumption 1 (Lipschitz gradient) *For every task $t \in [T]$ and agent $i \in \mathcal{V}$, there exist constants $L_1, L_2 > 0$ such that*

$$\|\nabla_\theta g_{t,i}(\theta) - \nabla_\theta g_{t,i}(\theta')\| \le L_1 \|\theta - \theta'\|, \tag{18}$$

$$\|\nabla_\phi \mathcal{H}_\phi(z_t) - \nabla_{\phi'} \mathcal{H}_{\phi'}(z_t)\| \le L_2 \|\phi - \phi'\|. \tag{19}$$

This assumption supplies the smoothness needed for the descent argument and for controlling changes in the generated classifier head.

Assumption 2 (Dynamic topology and spectral gap) *For every task and round q, the communication matrix W^q is real, symmetric, and doubly stochastic: $W^q_{ij} \in [0,1]$, $W^q_{ij} = W^q_{ji}$, $\sum_j W^q_{ij} = 1$, and $W^q \mathbf{1}_n = \mathbf{1}_n$. Define*

$$\rho_q := \left(\max\{|\lambda_2(W^q)|, |\lambda_n(W^q)|\}\right)^2, \tag{20}$$

where $\lambda_l(\cdot)$ is the l-th largest eigenvalue, and let $\rho = \max_q \rho_q$.

Here ρ is the squared non-consensus eigenvalue, and hence a contraction parameter: as ρ approaches 1, the spectral gap $1 - \sqrt{\rho}$ shrinks and the disagreement bound weakens.

Assumption 3 (Bounded intra-task gradient difference) *For every task $t \in [T]$, agent $i \in \mathcal{V}$, and communication round q, there exists a constant σ_g such that*

$$\|\nabla_\phi F_{t,i}(\phi_{t,i,q}) - \nabla_\phi F_t(\phi_{t,i,q})\|^2 \le \sigma_g^2. \tag{21}$$

This assumption quantifies agent-to-agent data heterogeneity within a single task.

Assumption 4 (Bounded inter-task gradient difference) *For every pair of tasks $t, t' \in [T]$, there exists a constant σ_t such that*

$$\|\nabla_\phi F_t(\phi) - \nabla_\phi F_{t'}(\phi)\|^2 \le \sigma_t^2. \tag{22}$$

This assumption isolates interference caused by differences between task-level objectives.

Assumption 5 (Bounded variance) *There exist constants $\delta, \sigma_2, \sigma_3$ such that, for every $t \in [T]$, $i \in \mathcal{V}$, and sample $\xi \sim D_t^{(i)}$,*

$$\mathbb{E}\|\nabla_\theta g_{t,i}(\theta) - \nabla_\theta g_{t,i}(\theta, \xi)\|^2 \le \delta^2. \tag{23}$$

This assumption controls the stochastic-gradient noise introduced by local sampling.

Assumption 6 (Bounded model gradients) *There exist constants b_1, b_2 such that, for all $i \in \mathcal{V}$,*

$$\|\nabla_\theta g_i(\theta)\| \le b_1, \qquad \|\nabla_\phi \mathcal{H}_\phi(z)\| \le b_2. \tag{24}$$

This assumption bounds the factors that appear when local target-model gradients are propagated through the hypernetwork.

Lemma 1 *Under Assumptions 1 and 6, for every task $t \in [T]$ and agent $i \in \mathcal{V}$, there exists a constant L_ϕ such that*

$$\|\nabla_\phi F_{t,i}(\phi) - \nabla_{\phi'} F_{t,i}(\phi')\| \le L_\phi \|\phi - \phi'\|. \tag{25}$$

Lemma 2 *For communication round q on agent i and $1 \le l \le K$, the following inequality holds:*

$$\mathbb{E}\left\| \nabla_\phi F_{t,i}(\phi_q) - \nabla_{\theta^c} g_{t,i} \left(\mathcal{H}_{\phi_q}(z_t) - \beta \sum_{j=0}^{l-1} \nabla_{\theta^c} g_{t,i}\big(\mathcal{H}_{\phi_q}(z_t)\big) \right) \nabla_\phi \mathcal{H}_{\phi_q}(z_t)^T \right\|^2$$

$$\le 27 L_1^2 b_2^2 \beta^2 l^2 (b_1^2 + \delta^2) + 3 b_1^2 \sigma_g^2 + 6 L_1^2 b_2^2 \sigma_l^2 := \Gamma. \tag{26}$$

B.2 PROOFS

Proof of Theorem 1 Here $\partial F_T(\Phi_{T,q})$ is the exact gradient stack, while the two-argument notation $\partial F_T(\Phi_{T,q}, z_T)$ denotes the stochastic local update direction used in the proof. We start from $F_{1:T}(\frac{\Phi_{T,q+1}\mathbf{1}_n}{n})$:

$$\begin{aligned}
&\mathbb{E}F_{1:T}\left(\frac{\Phi_{T,q+1}\mathbf{1}_n}{n}\right)\\
&=\mathbb{E}F_{1:T}\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}-\alpha\frac{\partial F_T(\Phi_{T,q},z_T)\mathbf{1}_n}{n}\right)\\
&\le \mathbb{E}F_{1:T}\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)+\underbrace{\mathbb{E}\langle\nabla F_{1:T}\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right),\frac{-\alpha\partial F_T(\Phi_{T,q},z_T)\mathbf{1}_n}{n}\rangle}_{T_1}\\
&\quad+\underbrace{\frac{\alpha^2L_\phi}{2}\mathbb{E}\left\|\frac{\partial F_T(\Phi_{T,q},z_T)\mathbf{1}_n}{n}\right\|^2}_{T_2}
\end{aligned}$$

To bound T_1:

$$\begin{aligned}
T_1&=\mathbb{E}\langle\nabla F_{1:T}\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right),\frac{-\alpha\partial F_T(\Phi_{T,q},z_T)\mathbf{1}_n}{n}\rangle\\
&=\mathbb{E}\langle\nabla F_{1:T}\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right),-\frac{\alpha\partial F_T(\Phi_{T,q},z_T)\mathbf{1}_n}{n}+\alpha\nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\rangle\\
&\quad+\mathbb{E}\langle\nabla F_{1:T}\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right),-\alpha\nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\rangle\\
&=\underbrace{\mathbb{E}\langle\nabla F_{1:T}\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right),-\frac{\alpha\partial F_T(\Phi_{T,q},z_T)\mathbf{1}_n}{n}+\alpha\nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\rangle}_{T_{11}}\\
&\quad+\underbrace{\mathbb{E}\langle\nabla F_{1:T-1}\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right),-\alpha\nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\rangle}_{T_{12}}-\alpha\mathbb{E}\left\|\nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\right\|^2
\end{aligned}$$

To bound T_{11}, we substitute $\nabla F_{1:T}=\sum_{t=1}^T\nabla F_t=T\nabla F_T+\sum_{t=1}^T(\nabla F_t-\nabla F_T)$.

$$\begin{aligned}
T_{11}&=\alpha\mathbb{E}\langle\nabla F_{1:T}\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right),\nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)-\frac{\partial F_T(\Phi_{T,q},z_T)\mathbf{1}_n}{n}\rangle\\
&=\alpha\mathbb{E}\langle T\nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)+\sum_{t=1}^T\left(\nabla F_t\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)-\nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\right),\\
&\quad\nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)-\frac{\partial F_T(\Phi_{T,q},z_T)\mathbf{1}_n}{n}\rangle\\
&=\alpha T\mathbb{E}\langle\nabla F_T,\nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)-\frac{\partial F_T(\Phi_{T,q},z_T)\mathbf{1}_n}{n}\rangle\\
&\quad+\alpha\mathbb{E}\langle\sum_{t=1}^T(\nabla F_t-\nabla F_T),\nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)-\frac{\partial F_T(\Phi_{T,q},z_T)\mathbf{1}_n}{n}\rangle\\
&\le\frac{\alpha T}{2}\mathbb{E}\|\nabla F_T\|^2+\frac{\alpha T}{2}\mathbb{E}\left\|\nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)-\frac{\partial F_T(\Phi_{T,q},z_T)\mathbf{1}_n}{n}\right\|^2\\
&\quad+\frac{\alpha}{2}\mathbb{E}\left\|\sum_{t=1}^T(\nabla F_t-\nabla F_T)\right\|^2+\frac{\alpha}{2}\mathbb{E}\left\|\nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)-\frac{\partial F_T(\Phi_{T,q},z_T)\mathbf{1}_n}{n}\right\|^2
\end{aligned}$$

$$\leq \frac{\alpha T}{2}\mathbb{E}\|\nabla F_T\|^2 + \frac{\alpha(T+1)}{2}\underbrace{\mathbb{E}\left\|\nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right) - \frac{\partial F_T(\Phi_{T,q}, z_T)\mathbf{1}_n}{n}\right\|^2}_{T_1'}$$

$$+ \frac{\alpha T}{2}\sum_{t=1}^{T}\mathbb{E}\|\nabla F_t - \nabla F_T\|^2$$

$$\leq \frac{\alpha T}{2}\mathbb{E}\left\|\nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\right\|^2 + \frac{\alpha(T+1)}{2}T_{21} + \frac{\alpha T^2}{2}\sigma_t^2$$

To bound T_1':

$$\begin{aligned}
T_1' &= \mathbb{E}\left\|\nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right) - \frac{\partial F_T(\Phi_{T,q}, z_T)\mathbf{1}_n}{n}\right\|^2 \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla F_{T,i}\left(\sum_{i'=1}^{n}\frac{\phi_{T,i',q}}{n}\right) - \nabla F_{T,i}(\phi_{T,i,q})\right\|^2 \\
&\leq \frac{L_\phi^2}{n}\sum_{i=1}^{n}\underbrace{\mathbb{E}\left\|\frac{\sum_{i'=1}^{n}\phi_{T,i',q}}{n} - \phi_{T,i,q}\right\|^2}_{:=R_q^{(i)}}
\end{aligned}$$

The first inequality above replaces the stochastic local update direction by the exact agent-gradient stack without displaying the required error bridge; we retain the line as part of the provisional source derivation rather than supplying an unproved step. Here $R_q^{(i)}$ denotes the distance between the agent-averaged hypernetwork parameter and agent i's local hypernetwork parameter.

$$\begin{aligned}
R_q^{(i)} &= \mathbb{E}\|\frac{\sum_{i=1}^{n}\phi_{T,q}^{(i')}}{n} - \phi_{T,q}^{(i')}\|^2 \\
&= \mathbb{E}\|\frac{\Phi_{T,q}\mathbf{1}_n}{n} - \Phi_{T,q}\cdot e_i\|^2 \\
&= \mathbb{E}\|\frac{\Phi_{T,q-1}W^{q-1}\mathbf{1}_n - \alpha\partial F_T(\Phi_{T,q-1}, z_T)\mathbf{1}_n}{n} \\
&\quad - (\Phi_{T,q-1}W^{q-1}e_i - \alpha\partial F_T(\Phi_{T,q-1}, z_T)e_i)\|^2 \\
&= \mathbb{E}\|\frac{\Phi_{T,q-1}\mathbf{1}_n - \alpha\partial F_T(\Phi_{T,q-1}, z_T)\mathbf{1}_n}{n} \\
&\quad - (\Phi_{T,q-1}W^{q-1}e_i - \alpha\partial F_T(\Phi_{T,q-1}, z_T)e_i)\|^2 \\
&= \mathbb{E}\|\frac{\Phi_{T,0}\mathbf{1}_n - \alpha\sum_{i=0}^{q-1}\partial F_T(\Phi_{T,i}, z_T)\mathbf{1}_n}{n} \\
&\quad - (\Phi_{T,0}W^{q-1}\cdots W^0 e_i - \sum_{j=0}^{q-1}\alpha\partial F_T(\Phi_{T,j}, z_T)W^{q-2}\cdots W^j e_i)\|^2 \\
&= \mathbb{E}\|\Phi_{T,0}\left(\frac{\mathbf{1}_n}{n} - W^{q-1}\cdots W^0 e_i\right) \\
&\quad - \sum_{j=0}^{q-1}\alpha\cdot\partial F_T\left(\Phi_{T,j}, z_T\right)\cdot\left(\frac{\mathbf{1}_n}{n} - W^{q-2}\cdots W^j e_i\right)\|^2 \\
&= \mathbb{E}\|\sum_{j=0}^{q-1}\alpha\cdot\partial F_T\left(\Phi_{T,j}, z_T\right)\left(\frac{\mathbf{1}_n}{n} - W^{q-2}\cdots W^j e_i\right)\|^2 \\
&= \alpha^2\mathbb{E}\|\sum_{j=0}^{q-1}\partial F_T\left(\Phi_{T,j}, z_T\right)\left(\frac{\mathbf{1}_n}{n} - W^{q-2}\cdots W^j e_i\right)\|^2
\end{aligned}$$

$$
\begin{aligned}
&= \alpha^2 \mathbb{E}\| \sum_{j=0}^{q-1} \partial F_T\left(\Phi_{T,j}, z_T\right) \left(\frac{\mathbf{1}_n}{n} - W^{q-2} \cdots W^j e_i \right) \\
&\quad - \sum_{j=0}^{q-1} \partial F_T\left(\Phi_{T,j}\right) \left(\frac{\mathbf{1}_n}{n} - W^{q-2} \cdots W^j e_i \right) \\
&\quad + \sum_{j=0}^{q-1} \partial F_T\left(\Phi_{T,j}\right) \left(\frac{\mathbf{1}_n}{n} - W^{q-2} \cdots W^j e_i \right) \|^2 \\
&\leq 3n\alpha^2 \sum_{j=0}^{q-1} K^2 \Gamma \mathbb{E}\| \left(\frac{\mathbf{1}_n}{n} - W^{q-2} \cdots W^j e_i \right) \|^2 \\
&\quad + 3\alpha^2 \underbrace{\mathbb{E}\| \sum_{j=0}^{q-1} \partial F_T\left(\Phi_{T,j}\right) \left(\frac{\mathbf{1}_n}{n} - W^{q-2} \cdots W^j e_i \right) \|^2}_{R_1} \\
&\leq \frac{3n\alpha^2 K^2 \Gamma}{1-\rho} + 3\alpha^2 R_1
\end{aligned}
$$

To bound R_1:

$$
\begin{aligned}
&\mathbb{E}\| \sum_{j=0}^{q-1} \partial F_T\left(\Phi_{T,j}\right) \left(\frac{\mathbf{1}_n}{n} - W^{q-2} \cdots W^j e_i \right) \|^2 \\
&= \underbrace{\sum_{j=0}^{q-1} \mathbb{E}\| \partial F_T\left(\Phi_{T,j}\right) \left(\frac{\mathbf{1}_n}{n} - W^{q-2} \cdots W^j e_i \right) \|^2}_{R_2} \\
&\quad + \underbrace{\sum_{j \neq j'} \mathbb{E}\langle \partial F_T\left(\Phi_{T,j}\right) (\frac{\mathbf{1}_n}{n} - W^{q-2} \cdots W^j e_i), \partial F_T\left(\Phi_{j'}\right) (\frac{\mathbf{1}_n}{n} - W^{q-2} \cdots W^{j'} e_i) \rangle}_{R_3}
\end{aligned}
$$

To bound R_2:

$$
\begin{aligned}
R_2 &\leq \sum_{j=0}^{q-1} \mathbb{E}\| \partial F_T\left(\Phi_{T,j}\right) \| \left\| \frac{\mathbf{1}_n}{n} - W^{q-2} \cdots W^j e_i \right\|^2 \\
&\leq \sum_{j=0}^{q-1} 3\mathbb{E}\| \partial F_T\left(\Phi_{T,j}\right) - \partial F_T\left(\frac{\Phi_{T,j}\mathbf{1}_n}{n} \mathbf{1}_n^T \right) \|^2 \left\| \frac{\mathbf{1}_n}{n} - W^{q-2} \cdots W^j e_i \right\|^2 \\
&\quad + \sum_{j=0}^{q-1} 3\mathbb{E}\| \partial F_T\left(\frac{\Phi_{T,j}\mathbf{1}_n}{n} \mathbf{1}_n^T \right) - \nabla F_T\left(\frac{\Phi_{T,j}\mathbf{1}_n}{n} \right) \mathbf{1}_n^T \|^2 \left\| \frac{\mathbf{1}_n}{n} - W^{q-2} \cdots W^j e_i \right\|^2 \\
&\quad + \sum_{j=0}^{q-1} 3\mathbb{E}\| \nabla F_T\left(\frac{\Phi_{T,j}\mathbf{1}_n}{n} \right) \mathbf{1}_n^T \|^2 \left\| \frac{\mathbf{1}_n}{n} - W^{q-2} \cdots W^j e_i \right\|^2 \\
&\leq 3L_\phi^2 \sum_{j=0}^{q-1} \sum_{h=1}^{n} \mathbb{E} R_j^{(h)} \left\| \frac{\mathbf{1}_n}{n} - W^{q-2} \cdots W^j e_i \right\|^2 + 3n\sigma_g^2 \cdot \frac{1}{1-\rho} \\
&\quad + 3 \sum_{j=0}^{q-1} \mathbb{E} \left\| \nabla F_T\left(\frac{\Phi_{T,j}\mathbf{1}_n}{n} \right) \mathbf{1}_n^T \right\|^2 \left\| \frac{\mathbf{1}_n}{n} - W^{q-2} \cdots W^j e_i \right\|^2
\end{aligned}
$$

To bound R_3:

$$
R_3 \leq \sum_{j \neq j'}^{q-1} \mathbb{E}\| \partial F_T(\Phi_j) (\frac{\mathbf{1}_n}{n} - W^{q-2} \cdots W^j e_i) \|
$$

$$
\begin{aligned}
&\cdot \|\partial F_T(\Phi_{j'})(\frac{\mathbf{1}_n}{n} - W^{q-2}\cdots W^{j'}e_i)\| \\
&\leq \sum_{j\neq j'}^{q-1} \mathbb{E}\|\partial F_T(\Phi_j)\|\|\frac{\mathbf{1}_n}{n} - W^{q-2}\cdots W^{j}e_i\| \\
&\quad \cdot \|\partial F_T(\Phi_{j'})\|\|\frac{\mathbf{1}_n}{n} - W^{q-2}\cdots W^{j'}e_i\| \\
&\leq \sum_{j\neq j'}^{q-1} \mathbb{E}\frac{\|\partial F_T(\Phi_j)\|^2}{2}\|\frac{\mathbf{1}_n}{n} - W^{q-2}\cdots W^{j}e_i\|\|\frac{\mathbf{1}_n}{n} - W^{q-2}\cdots W^{j'}e_i\| \\
&\quad + \sum_{j\neq j'}^{q-1} \mathbb{E}\frac{\|\partial F_T(\Phi_{j'})\|^2}{2}\|\frac{\mathbf{1}_n}{n} - W^{q-2}\cdots W^{j}e_i\|\|\frac{\mathbf{1}_n}{n} - W^{q-2}\cdots W^{j'}e_i\| \\
&\leq \sum_{j\neq j'}^{q-1} \mathbb{E}\left(\frac{\|\partial F_T(\Phi_j)\|}{2} + \frac{\|\partial F_T(\Phi_{j'})\|^2}{2}\right)\cdot \rho^{q-\frac{j+j'}{2}-1} \\
&\leq \sum_{j\neq j'}^{q-1} \mathbb{E}\left(\|\partial F_T(\Phi_j)\|^2\right)\cdot \rho^{q-\frac{j+j'}{2}-1} \\
&\leq \underbrace{3\sum_{j\neq j'}^{q-1}\left(\sum_{h=1}^{n}\mathbb{E}L_\phi^2\mathbb{E}R_j^{(h)} + \mathbb{E}\|\nabla F_T(\frac{\Phi_j\mathbf{1}_n}{n})\mathbf{1}_n^T\|\right)\cdot\rho^{q-\frac{j+j'}{2}-1}}_{R_4} \\
&\quad + \underbrace{\sum_{j\neq j'}^{q-1} 3n\sigma_g^2\rho^{q-1-\frac{j+j'}{2}}}_{R_5}
\end{aligned}
$$

To bound R_4:

$$
\begin{aligned}
R_4 &= 6\sum_{j=0}^{q-1}\left(\sum_{h=1}^{n}\mathbb{E}L_\phi^2\mathbb{E}R_j^{(h)} + \mathbb{E}\|F_T(\frac{\Phi_j\mathbf{1}_n}{n})\cdot\mathbf{1}_n^T\|\right)\sum_{j'=j+1}^{q-1}\sqrt{\rho}^{2q-j-j'-2} \\
&\leq 6\sum_{j=0}^{q-1}\left(\sum_{h=1}^{n}\mathbb{E}L_\phi^2\mathbb{E}R_j^{(h)} + \mathbb{E}\|F_T(\frac{\Phi_j\mathbf{1}_n}{n})\cdot\mathbf{1}_n^T\|\right)\cdot\frac{\sqrt{\rho}^{q-j-1}}{1-\sqrt{\rho}}
\end{aligned}
$$

To bound R_5:

$$
\begin{aligned}
R_5 &\leq 6n\sigma_g^2\sum_{j>j'}^{q-1}\rho^{q-1-\frac{j+j'}{2}} \\
&= 6n\sigma_g^2\frac{(\rho^{q/2}-1)(\rho^{q/2}-\sqrt{\rho})}{(\sqrt{\rho}-1)^2(\sqrt{\rho}+1)} \\
&\leq 6n\sigma_g^2\frac{1}{(1-\sqrt{\rho})^2}
\end{aligned}
$$

Substituting R_4 and R_5 into R_3 gives

$$
T_5 \leq 6\sum_{j=0}^{q-1}\left(\sum_{h=1}^{n}\mathbb{E}L_\phi^2R_j^{(h)} + \mathbb{E}\|F_T(\frac{\Phi_j\mathbf{1}_n}{n})\cdot\mathbf{1}_n^T\|\right)\cdot\frac{\sqrt{\rho}^{q-j-1}}{1-\sqrt{\rho}} + 6n\sigma_g^2\frac{1}{(1-\sqrt{\rho})^2}
$$

Substituting R_2 and R_3 into R_1 gives

$$
R_1 \leq 3L_\phi^2\sum_{j=0}^{q-1}\sum_{h=1}^{n}\mathbb{E}R_j^{(h)}\left\|\frac{\mathbf{1}_n}{n} - W^{q-2}\cdots W^{j}e_i\right\|^2
$$

$$
\begin{aligned}
&+ 3\sum_{j=0}^{q-1}\mathbb{E}\left\|\nabla F_T\left(\frac{\Phi_{T,j}\mathbf{1}_n}{n}\right)\mathbf{1}_n^T\right\|^2\left\|\frac{\mathbf{1}_n}{n}-W^{q-2}\cdots W^j e_i\right\|^2\\
&+ 6\sum_{j=0}^{q-1}\left(\sum_{h=1}^{n}\mathbb{E}L_\phi^2 R_j^{(h)}+\mathbb{E}\|F_T(\frac{\Phi_j\mathbf{1}_n}{n})\cdot\mathbf{1}_n^T\|\right)\cdot\frac{\sqrt{\rho}^{q-j-1}}{1-\sqrt{\rho}}\\
&+ 3n\sigma_g^2\cdot\frac{1}{1-\rho}+6n\sigma_g^2\frac{1}{(1-\sqrt{\rho})^2}\\
\le\ & 3L_\phi^2\sum_{j=0}^{q-1}\sum_{h=1}^{n}\mathbb{E}R_j^{(h)}\left\|\frac{\mathbf{1}_n}{n}-W^{q-2}\cdots W^j e_i\right\|^2\\
&+ 3\sum_{j=0}^{q-1}\mathbb{E}\left\|\nabla F_T\left(\frac{\Phi_{T,j}\mathbf{1}_n}{n}\right)\mathbf{1}_n^T\right\|^2\left\|\frac{\mathbf{1}_n}{n}-W^{q-2}\cdots W^j e_i\right\|^2\\
&+ 6\sum_{j=0}^{q-1}\left(\sum_{h=1}^{n}\mathbb{E}L_\phi^2 R_j^{(h)}+\mathbb{E}\|F(\frac{\Phi_j\mathbf{1}_n}{n})\cdot\mathbf{1}_n^T\|\right)\cdot\frac{\sqrt{\rho}^{q-j-1}}{1-\sqrt{\rho}}+\frac{9n\sigma_g^2}{(1-\sqrt{\rho})^2}
\end{aligned}
$$

Substituting the bound for R_1 into $R_q^{(i)}$ gives

$$
\begin{aligned}
R_q^{(i)} \le\ & \frac{3n\alpha^2K^2\Gamma}{1-\rho}\\
&+ 9\alpha^2L_\phi^2\sum_{j=0}^{q-1}\sum_{h=1}^{n}\mathbb{E}R_j^{(h)}\|\frac{\mathbf{1}_n}{n}-W^{q-2}\cdots W^j e_i\|^2\\
&+ 9\alpha^2\sum_{j=0}^{q-1}\mathbb{E}\|\nabla F_T(\frac{\Phi_j\mathbf{1}_n}{n})\cdot\mathbf{1}_n^T\|^2\|\frac{\mathbf{1}_n}{n}-W^{q-2}\cdots W^j e_i\|^2\\
&+ 18\alpha^2\sum_{j=0}^{q-1}\left(\sum_{h=1}^{n}\mathbb{E}L_\phi^2 R_j^{(h)}+\mathbb{E}\|\nabla F_T(\frac{\Phi_j\mathbf{1}_n}{n})\cdot\mathbf{1}_n^T\|\right)\cdot\frac{\sqrt{\rho}^{q-j-1}}{1-\sqrt{\rho}}\\
&+ \frac{27\alpha^2n\sigma_g^2}{(1-\sqrt{\rho})^2}\\
\le\ & \frac{3n\alpha^2K^2\Gamma}{1-\rho}+9\alpha^2L_\phi^2\sum_{j=0}^{q-1}\sum_{h=1}^{n}\mathbb{E}R_j^{(h)}\rho^{q-j-1}\\
&+ 9\alpha^2\sum_{j=0}^{q-1}\mathbb{E}\|\nabla F_T(\frac{\Phi_j\mathbf{1}_n}{n})\cdot\mathbf{1}_n^T\|^2\cdot\rho^{q-j-1}\\
&+ 18\alpha^2\sum_{j=0}^{q-1}\left(\sum_{h=1}^{n}\mathbb{E}L_\phi^2 R_j^{(h)}+\mathbb{E}\|\nabla F_T(\frac{\Phi_j\mathbf{1}_n}{n})\cdot\mathbf{1}_n^T\|\right)\cdot\frac{\sqrt{\rho}^{q-j-1}}{1-\sqrt{\rho}}\\
&+ \frac{27\alpha^2n\sigma_g^2}{(1-\sqrt{\rho})^2}\\
=\ & \frac{3n\alpha^2K^2\Gamma}{1-\rho}+\frac{27\alpha^2n\sigma_g^2}{(1-\sqrt{\rho})^2}\\
&+ 9\alpha^2\sum_{j=0}^{q-1}\mathbb{E}\|\nabla F_T(\frac{\Phi_j\mathbf{1}_n}{n})\cdot\mathbf{1}_n^T\|^2\left(\rho^{q-j-1}+\frac{2\sqrt{\rho}^{q-j-1}}{1-\sqrt{\rho}}\right)\\
&+ 9\alpha^2L_\phi^2\sum_{j=0}^{q-1}\sum_{h=1}^{n}\mathbb{E}R_j^{(h)}\left(\rho^{q-j-1}+\frac{2\sqrt{\rho}^{q-j-1}}{1-\sqrt{\rho}}\right)
\end{aligned}
$$

Averaging over all agents gives

$$\begin{aligned}\mathbb{E}M_q &= \frac{\mathbb{E}\sum_{i=1}^n R_q^{(i)}}{n} \\ &\le \frac{3n\alpha^2K^2\Gamma}{1-\rho} + \frac{27\alpha^2 n\sigma_g^2}{(1-\sqrt{\rho})^2} \\ &\quad + 9\alpha^2\sum_{j=0}^{q-1}\mathbb{E}\|\nabla F_T(\frac{\Phi_j\mathbf{1}_n}{n})\cdot\mathbf{1}_n^T\|^2\left(\rho^{q-j-1}+\frac{2\sqrt{\rho}^{q-j-1}}{1-\sqrt{\rho}}\right) \\ &\quad + 9\alpha^2 nL_\phi^2\sum_{j=0}^{q-1}\mathbb{E}M_j\left(\rho^{q-j-1}+\frac{2\sqrt{\rho}^{q-j-1}}{1-\sqrt{\rho}}\right)\end{aligned}$$

Summing from $q=0$ to $Q-1$ gives

$$\begin{aligned}\sum_{q=0}^{Q-1}\mathbb{E}M_q &\le \frac{3n\alpha^2QK^2\Gamma}{1-\rho} + \frac{27\alpha^2 n\sigma_g^2 Q}{(1-\sqrt{\rho})^2} \\ &\quad + 9\alpha^2\sum_{q=0}^{Q-1}\sum_{j=0}^{q-1}\mathbb{E}\|\nabla F_T(\frac{\Phi_j\mathbf{1}_n}{n})\cdot\mathbf{1}_n^T\|^2\left(\rho^{q-j-1}+\frac{2\sqrt{\rho}^{q-j-1}}{1-\sqrt{\rho}}\right) \\ &\quad + 9\alpha^2 nL_\phi^2\sum_{q=0}^{Q-1}\sum_{j=0}^{q-1}\mathbb{E}M_j\left(\rho^{q-j-1}+\frac{2\sqrt{\rho}^{q-j-1}}{1-\sqrt{\rho}}\right) \\ &\le \frac{3n\alpha^2QK^2\Gamma}{1-\rho} + \frac{27\alpha^2 n\sigma_g^2 Q}{(1-\sqrt{\rho})^2} \\ &\quad + 9\alpha^2\sum_{q=0}^{Q-1}\mathbb{E}\|\nabla F_T(\frac{\Phi_j\mathbf{1}_n}{n})\cdot\mathbf{1}_n^T\|^2\left(\sum_{i=0}^{\infty}\rho^i+\frac{2\sum_{i=0}^{\infty}\sqrt{\rho}^i}{1-\sqrt{\rho}}\right) \\ &\quad + 9\alpha^2 nL_\phi^2\sum_{q=0}^{Q-1}\mathbb{E}M_q\left(\sum_{i=0}^{\infty}\rho^i+\frac{2\sum_{i=0}^{\infty}\sqrt{\rho}^i}{1-\sqrt{\rho}}\right) \\ &\le \frac{3n\alpha^2QK^2\Gamma}{1-\rho} + \frac{27\alpha^2 n\sigma_g^2 Q}{(1-\sqrt{\rho})^2} \\ &\quad + \frac{27\alpha^2}{(1-\sqrt{\rho})^2}\sum_{q=0}^{Q-1}\mathbb{E}\|\nabla F_T(\frac{\Phi_j\mathbf{1}_n}{n})\cdot\mathbf{1}_n^T\|^2 + \frac{27\alpha^2 nL_\phi^2}{(1-\sqrt{\rho})^2}\sum_{q=0}^{Q-1}\mathbb{E}M_q\end{aligned}$$

Rearranging the terms gives

$$\begin{aligned}&\left(1-\frac{27\alpha^2 nL_\phi^2}{(1-\sqrt{\rho})^2}\right)\sum_{q=0}^{Q-1}\mathbb{E}M_q \\ &\le \frac{3n\alpha^2QK^2\Gamma}{1-\rho} + \frac{27\alpha^2 n\sigma_g^2 Q}{(1-\sqrt{\rho})^2} \\ &\quad + \frac{27\alpha^2}{(1-\sqrt{\rho})^2}\sum_{q=0}^{Q-1}\mathbb{E}\left\|\nabla F_T\left(\frac{\Phi_j\mathbf{1}_n}{n}\right)\mathbf{1}_n^T\right\|^2\end{aligned}$$

Let $A_1 = 1-\frac{27\alpha^2 nL_\phi^2}{(1-\sqrt{\rho})^2}$. Then

$$\sum_{q=0}^{Q-1}\mathbb{E}M_q \le \frac{3n\alpha^2QK^2\Gamma}{A_1(1-\rho)} + \frac{27\alpha^2 n\sigma_g^2 Q}{A_1(1-\sqrt{\rho})^2} + \frac{27\alpha^2}{A_1(1-\sqrt{\rho})^2}\sum_{q=0}^{Q-1}\mathbb{E}\|\nabla F_T(\frac{\Phi_j\mathbf{1}_n}{n})\cdot\mathbf{1}_n^T\|^2$$

Returning to T_1' gives

$$T_1' \le \frac{L_\phi^2}{n}\sum_{i=1}^n R_q^{(i)} = L_\phi^2\mathbb{E}M_q$$

To bound T_{12}, use $\|a+b\|^2 - \|a\|^2 - \|b\|^2 = 2\langle a, b\rangle$:

$$
\begin{aligned}
T_{12} &= -\alpha\mathbb{E}\langle \nabla F_{1:T-1}\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right), \nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\rangle \\
&= -\frac{\alpha}{2}\mathbb{E}\left(\left\|\nabla F_{1:T}\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\right\|^2 - \left\|\nabla F_{1:T-1}\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\right\|^2 - \left\|\nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\right\|^2\right) \\
&= -\frac{\alpha}{2}\mathbb{E}\left\|\nabla F_{1:T}\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\right\|^2 + \frac{\alpha}{2}\mathbb{E}\left\|\nabla F_{1:T-1}\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\right\|^2 \\
&\quad + \frac{\alpha}{2}\mathbb{E}\left\|\nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\right\|^2 \\
&= -\frac{\alpha}{2}\mathbb{E}\left\|\nabla F_{1:T}\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\right\|^2 + \frac{\alpha}{2}\mathbb{E}\|\nabla F_{1:T-1}\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right) \\
&\quad - \nabla F_{1:T-1}\left(\frac{\Phi_{T,0}\mathbf{1}_n}{n}\right) + \nabla F_{1:T-1}\left(\frac{\Phi_{T,0}\mathbf{1}_n}{n}\right)\|^2 + \frac{\alpha}{2}\mathbb{E}\left\|\nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\right\|^2 \\
&= -\frac{\alpha}{2}\mathbb{E}\left\|\nabla F_{1:T}\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\right\|^2 + \frac{3\alpha^3 L_\phi^2 (T-1)^2}{2}\frac{K^2\Gamma}{n} \\
&\quad + \frac{3\alpha}{2}\mathbb{E}\left\|\nabla F_{1:T-1}\left(\frac{\Phi_{T,0}\mathbf{1}_n}{n}\right)\right\|^2 + \frac{\alpha}{2}\mathbb{E}\left\|\nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\right\|^2
\end{aligned}
$$

Substituting T_1' into T_{11} gives

$$
T_{11} \le \frac{\alpha T}{2}\mathbb{E}\left\|\nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\right\|^2 + \frac{\alpha(T+1)}{2}L_\phi^2\mathbb{E}M_q + \frac{\alpha T^2}{2}\sigma_t^2
$$

Substituting T_{11} and T_{12} into T_1 gives

$$
\begin{aligned}
T_1 &\le T_{11} + T_{12} - \alpha\mathbb{E}\left\|\nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\right\|^2 \\
&\le \frac{\alpha T}{2}\mathbb{E}\left\|\nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\right\|^2 + \frac{\alpha(T+1)}{2}L_\phi^2\mathbb{E}M_q + \frac{\alpha T^2}{2}\sigma_t^2 \\
&\quad - \alpha\mathbb{E}\left\|\nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\right\|^2 \\
&\quad + \frac{3\alpha^3 L_\phi^2 (T-1)^2}{2}\frac{K^2\Gamma}{n} \\
&\quad + \frac{3\alpha}{2}\mathbb{E}\left\|\nabla F_{1:T-1}\left(\frac{\Phi_{T,0}\mathbf{1}_n}{n}\right)\right\|^2 - \frac{\alpha}{2}\mathbb{E}\left\|\nabla F_{1:T}\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\right\|^2 \\
&\quad + \frac{\alpha}{2}\mathbb{E}\left\|\nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\right\|^2 \\
&\le -\frac{\alpha}{2}\mathbb{E}\left\|\nabla F_{1:T}\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\right\|^2 + \frac{3\alpha^3 L_\phi^2 (T-1)^2}{2}\frac{K^2\Gamma}{n} \\
&\quad + \frac{3\alpha}{2}\mathbb{E}\left\|\nabla F_{1:T-1}\left(\frac{\Phi_{T,0}\mathbf{1}_n}{n}\right)\right\|^2 \\
&\quad + \frac{\alpha(T-1)}{2}\mathbb{E}\left\|\nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\right\|^2 + \frac{\alpha(T+1)}{2}L_\phi^2\mathbb{E}M_q + \frac{\alpha T^2}{2}\sigma_t^2
\end{aligned}
$$

To bound T_2:

T_2 **proof status.** The following legacy formal expansion is not justified by the current definitions: it identifies the stochastic local update direction with $\nabla F_{T,i}(\phi_q^{(i)}, z_T)$ and then evaluates $\nabla_\phi F_{T,i}$ on

a generated-head vector outside its declared domain. The ensuing T_2 bound is therefore provisional pending an author-supplied estimator identity.

$$
\begin{aligned}
T_2 &= \frac{\alpha^2 L_\phi}{2}\mathbb{E}\left\|\frac{\partial F_T(\Phi_{T,q}, z_T)\mathbf{1}_n}{n}\right\|^2 \\
&= \frac{\alpha^2 L_\phi}{2}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla F_{T,i}\left(\phi_q^{(i)}, z_T\right)\right\|^2 \\
&= \frac{\alpha^2 L_\phi}{2}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{K}\nabla_{\theta^c} g_{q,l}^i\left(\mathcal{H}_{\phi_q^{(i)}}(z_T)\right)\nabla_\phi \mathcal{H}_{\phi_q^{(i)}}(z_T)^T\right\|^2 \\
&= \frac{\alpha^2 L_\phi}{2}\mathbb{E}\Bigg\|\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{K}\nabla_{\theta^c} g_{q,l}^i\left(\mathcal{H}_{\phi_q^{(i)}}(z_T)\right)\nabla_\phi \mathcal{H}_{\phi_q^{(i)}}(z_T)^T \\
&\quad - \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{K}\nabla_\phi F_{T,i}\left[\mathcal{H}_{\phi_q^{(i)}}(z_T) - \beta\sum_{j=0}^{l-1}\nabla_{\theta^c} g_{q,j}^i\left(\mathcal{H}_{\phi_q^{(i)}}(z_T)\right)\right]\nabla_\phi \mathcal{H}_{\phi_q^{(i)}}(z_T)^T \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{K}\nabla_\phi F_{T,i}\left[\mathcal{H}_{\phi_q^{(i)}}(z_T) - \beta\sum_{j=0}^{l-1}\nabla_{\theta^c} g_{q,j}^i\left(\mathcal{H}_{\phi_q^{(i)}}(z_T)\right)\right]\nabla_\phi \mathcal{H}_{\phi_q^{(i)}}(z_T)^T \\
&\quad - \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{K}\nabla_\phi F_{T,i}(\phi_q^{(i)}) + \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{K}\nabla_\phi F_{T,i}(\phi_q^{(i)})\|^2 \\
&\le \frac{3\alpha^2 L_\phi}{2}\mathbb{E}\Bigg\|\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{K}\nabla_{\theta^c} g_{q,l}^i\left(\mathcal{H}_{\phi_q^{(i)}}(z_T)\right)\nabla_\phi \mathcal{H}_{\phi_q^{(i)}}(z_T)^T \\
&\quad - \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{K}\nabla_\phi F_{T,i}\left[\mathcal{H}_{\phi_q^{(i)}}(z_T) - \beta\sum_{j=0}^{l-1}\nabla_{\theta^c} g_{q,j}^i\left(\mathcal{H}_{\phi_q^{(i)}}(z_T)\right)\right]\nabla_\phi \mathcal{H}_{\phi_q^{(i)}}(z_T)^T\Bigg\|^2 \\
&\quad + \frac{3\alpha^2 L_\phi}{2}\mathbb{E}\Bigg\|\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{K}\nabla_\phi F_{T,i}\left[\mathcal{H}_{\phi_q^{(i)}}(z_T) - \beta\sum_{j=0}^{l-1}\nabla_{\theta^c} g_{q,j}^i\left(\mathcal{H}_{\phi_q^{(i)}}(z_T)\right)\right]\nabla_\phi \mathcal{H}_{\phi_q^{(i)}}(z_T)^T \\
&\quad - \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{K}\nabla_\phi F_i(\phi_q^{(i)})\Bigg\|^2 + \frac{3\alpha^2 L_\phi}{2}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{K}\nabla_\phi F_{T,i}(\phi_q^{(i)})\right\|^2 \\
&\le \frac{3\alpha^2 L_\phi K^2\beta^2 b_2^2\delta^2}{2n} + \frac{3\alpha^2 L_\phi K^2}{2}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_\phi F_{T,i}\left(\phi_q^{(i)}\right)\right\|^2 + \frac{3K^2\alpha^2 L_\phi \Gamma}{2n}
\end{aligned}
$$

Combining these bounds gives

$$
\begin{aligned}
&\mathbb{E}F_{1:T}\left(\frac{\Phi_{T,q+1}\mathbf{1}_n}{n}\right) \\
\le&\mathbb{E}F_{1:T}\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right) + T_1 + \frac{3\alpha^2 L_\phi K^2\beta^2 b_2^2\delta^2}{2n} \\
&\quad + \frac{3\alpha^2 L_\phi K^2}{2}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_\phi F_{T,i}\left(\phi_q^{(i)}\right)\right\|^2 + \frac{3K^2\alpha^2 L_\phi \Gamma}{2n} \\
\le&\mathbb{E}F_{1:T}\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right) - \frac{\alpha}{2}\mathbb{E}\left\|\nabla F_{1:T}\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\right\|^2 \\
&\quad + \frac{3\alpha}{2}\mathbb{E}\left\|\nabla F_{1:T-1}\left(\frac{\Phi_{T,0}\mathbf{1}_n}{n}\right)\right\|^2 + \frac{\alpha(T-1)}{2}\mathbb{E}\left\|\nabla F_T\left(\frac{\Phi_{T,g}\mathbf{1}_n}{n}\right)\right\|^2
\end{aligned}
$$

$$+\frac{3\alpha^2 L_\phi K^2}{2}\mathbb{E}\left\|\frac{\partial F_T(\Phi_{T,q})\mathbf{1}_n}{n}\right\|^2+\frac{\alpha(T+1)L_\phi^2}{2}\mathbb{E}M_q+\frac{\alpha T^2}{2}\sigma_t^2$$
$$+\frac{3\alpha^3 L_\phi^2(T-1)^2}{2n}(K^2\Gamma)+\frac{3\alpha^2 L_\phi K^2\beta^2 b_2^2\delta^2}{2n}+\frac{3K^2\alpha^2 L_\phi\Gamma}{2n}$$

Summing from $q=0$ to $q=Q-1$ gives

$$\begin{aligned}
&\frac{\alpha}{2}\sum_{q=0}^{Q-1}\mathbb{E}\left\|\nabla F_{1:T}\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\right\|^2\\
&\leq \underbrace{\mathbb{E}F_{1:T}\left(\frac{\Phi_{T,0}\mathbf{1}_n}{n}\right)-\mathbb{E}F_{1:T}\left(\frac{\Phi_{T,Q}\mathbf{1}_n}{n}\right)}_{\Delta_f}\\
&\quad+\frac{3\alpha Q}{2}\mathbb{E}\left\|\nabla F_{1:T-1}\left(\frac{\Phi_{T,0}\mathbf{1}_n}{n}\right)\right\|^2\\
&\quad+\frac{\alpha(T-1)}{2}\sum_{q=0}^{Q-1}\mathbb{E}\left\|\nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\right\|^2\\
&\quad+\frac{3\alpha^2 L_\phi K^2}{2}\sum_{q=0}^{Q-1}\mathbb{E}\left\|\frac{\partial F_T(\Phi_{T,q})\mathbf{1}_n}{n}\right\|^2\\
&\quad+Q(\frac{\alpha T^2}{2}\sigma_t^2+\frac{3\alpha^3 L_\phi^2(T-1)^2K^2\Gamma}{2n}\\
&\quad+\frac{3\alpha^2 L_\phi K^2\beta^2 b_2^2\delta^2}{2n}+\frac{3K^2\alpha^2 L_\phi\Gamma}{2n})\\
&\quad+\frac{\alpha(T+1)L_\phi^2}{2}\sum_{q=0}^{Q-1}\mathbb{E}M_q
\end{aligned}$$

Substituting the bound for $\sum_{q=0}^{Q-1}\mathbb{E}M_q$ gives

$$\begin{aligned}
&\frac{\alpha(T+1)L_\phi^2}{2}\sum_{q=0}^{Q-1}\mathbb{E}M_q\\
&\leq\frac{\alpha(T+1)L_\phi^2}{2}[\frac{3nQ\alpha^2K^2\Gamma}{(1-\rho)A_1}+\frac{27\alpha^2 n\sigma_g^2 Q}{(1-\sqrt{\rho})^2A_1}\\
&\quad+\frac{27\alpha^2 n}{(1-\sqrt{\rho})^2A_1}\sum_{q=0}^{Q-1}\mathbb{E}\left\|\nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\right\|^2]\\
&=Q[\frac{3\alpha^3(T+1)L_\phi^2nK^2\Gamma}{2(1-\rho)A_1}+\frac{27\alpha^3(T+1)L_\phi^2n}{2(1-\sqrt{\rho})^2A_1}\sigma_g^2]\\
&\quad+\frac{27\alpha^3(T+1)L_\phi^2n}{2(1-\sqrt{\rho})^2A_1}\sum_{q=0}^{Q-1}\mathbb{E}\left\|\nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\right\|^2
\end{aligned}$$

Combining the preceding inequalities gives

$$\begin{aligned}
&\frac{\alpha}{2}\sum_{q=0}^{Q-1}\mathbb{E}\left\|\nabla F_{1:T}\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\right\|^2\\
&\leq\Delta_f+\frac{3\alpha Q}{2}\mathbb{E}\left\|\nabla F_{1:T-1}\left(\frac{\Phi_{T,0}\mathbf{1}_n}{n}\right)\right\|^2
\end{aligned}$$

$$
\begin{aligned}
&+\left(\frac{\alpha(T-1)}{2}+\frac{27\alpha^3(T+1)L_\phi^2 n}{2(1-\sqrt{\rho})^2A_1}\right)\sum_{q=0}^{Q-1}\mathbb{E}\left\|\nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\right\|^2\\
&+\frac{3\alpha^2L_\phi K^2}{2}\sum_{q=0}^{Q-1}\mathbb{E}\left\|\frac{\partial F_T(\Phi_{T,q})\mathbf{1}_n}{n}\right\|^2+Q(\frac{\alpha T^2}{2}\sigma_t^2\\
&+\frac{3\alpha^3L_\phi^2(T-1)^2K^2\Gamma}{2n}+\frac{3\alpha^2L_\phi K^2\beta^2b_2^2\delta^2}{2n}+\frac{3K^2\alpha^2L_\phi\Gamma}{2n}\\
&+\frac{3\alpha^3(T+1)L_\phi^2nK^2\Gamma}{2(1-\rho)A_1}+\frac{27\alpha^3(T+1)L_\phi^2n}{2(1-\sqrt{\rho})^2A_1}\sigma_g^2)
\end{aligned}
$$

Dividing both sides of the inequality by $\frac{\alpha Q}{2}$ gives

$$
\begin{aligned}
&\frac{1}{Q}\sum_{q=0}^{Q-1}\mathbb{E}\left\|\nabla F_{1:T}\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\right\|^2\\
&\le\frac{2\Delta_f}{\alpha Q}+3\mathbb{E}\left\|\nabla F_{1:T-1}\left(\frac{\Phi_{T,0}\mathbf{1}_n}{n}\right)\right\|^2\\
&\quad+\left((T-1)+\frac{27\alpha^2(T+1)L_\phi^2n}{(1-\sqrt{\rho})^2A_1}\right)\frac{1}{Q}\sum_{q=0}^{Q-1}\mathbb{E}\left\|\nabla F_T\left(\frac{\Phi_{T,q}\mathbf{1}_n}{n}\right)\right\|^2\\
&\quad+3\alpha L_\phi K^2\frac{1}{Q}\sum_{q=0}^{Q-1}\mathbb{E}\left\|\frac{\partial F_T(\Phi_{T,q})\mathbf{1}_n}{n}\right\|^2+T^2\sigma_t^2\\
&\quad+\frac{3\alpha^2L_\phi^2(T-1)^2K^2\Gamma}{n}+\frac{3\alpha L_\phi K^2\beta^2b_2^2\delta^2}{n}+\frac{3K^2\alpha L_\phi\Gamma}{n}\\
&\quad+\frac{3\alpha^2(T+1)L_\phi^2nK^2\Gamma}{(1-\rho)A_1}+\frac{27\alpha^2(T+1)L_\phi^2n}{(1-\sqrt{\rho})^2A_1}\sigma_g^2.
\end{aligned}
$$

Let $\alpha=\sqrt{n/Q}$. If $Q\ge 54L_\phi^2n^2/(1-\sqrt{\rho})^2$, then

$$
A_1=1-\frac{27n^2L_\phi^2}{Q(1-\sqrt{\rho})^2}\ge\frac{1}{2}>0.
$$

Under Q-uniform bounds on the quantities multiplying the explicitly Q-dependent terms, and for fixed n, T, K, and ρ, those remainder terms in the endpoint inequality scale as $\mathcal{O}(Q^{-1/2}+Q^{-1})$. The full endpoint remains a residual-neighborhood bound: it still contains the historical-gradient term, the current-task-gradient average with its $(T-1)$ coefficient, and $T^2\sigma_t^2$. Consequently, this derivation does not establish that the full stationarity measure converges to zero.

C Experimental Details

C.1 Baselines

Class-Incremental Learning (CL) These methods primarily focus on mitigating catastrophic forgetting during the acquisition of new tasks:

- **Fine-Tuning (Lian et al., 2018):** A straightforward approach that updates the model on new data without specific constraints, serving as a lower bound for forgetting.
- **LwF (Li & Hoiem, 2017):** A distillation-based method that preserves previous knowledge by using the old model's outputs as soft labels for the new task.
- **EWC (Kirkpatrick et al., 2017):** A regularization-based approach that penalizes changes to weights important to previous tasks via the Fisher Information Matrix.

- **HNET (von Oswald et al., 2020):** A task-conditioned hypernetwork baseline that generates task-specific parameters and uses output regularization to mitigate forgetting.

Machine Unlearning (UL) These baselines are designed for unlearning specific target classes. We integrate these UL baselines with EWC or LwF to endow the model with both unlearning and continual learning capabilities:

- **FedAU (Gu et al., 2024):** An efficient federated machine unlearning framework that performs unlearning via a lightweight auxiliary module. By applying a linear transformation to the model weights based on the target data distribution, it facilitates rapid unlearning without exhaustive retraining or recovery.
- **QuickDropSGA (Dhasade et al., 2024):** A method that applies gradient ascent to target-class data for class unlearning.
- **BadTeacher (Chundawat et al., 2023):** An approach that uses a student–teacher framework for selective unlearning. A competent teacher maintains retained-set performance, while an incompetent teacher guides the model away from the target class without retraining from scratch.

Unified Class-Incremental and Unlearning Frameworks (CLU) We also compare with state-of-the-art unified frameworks designed for simultaneous incremental learning and unlearning:

- **FedAF (Li et al., 2025):** A neuro-inspired federated unlearning framework that achieves class-wise data removal through active forgetting. FedAF utilizes a teacher-student memory generation mechanism, where new memories are generated to overwrite old ones. This approach is independent of model architectures and provides a precise, fine-grained unlearning solution without the need for historical data storage.
- **LSF (Shibata et al., 2021)**: A proactive lifelong learning framework that achieves selective forgetting through class-specific synthetic "mnemonic codes." By watermarking these codes onto training samples during the learning phase, the model can later remove specific class knowledge using only the codes, bypassing the need for original data. In our work, we include it to evaluate the efficacy of signal-based selective unlearning.
- **CLPU-DER++ (Liu et al., 2022)**: An exactly private baseline designed for the combined challenge of continual learning and private unlearning. Based on the DER++ replay framework, it incorporates strict lifecycle management for the memory buffer to support task-level unlearning upon request.
- **CLMUL (Chatterjee et al., 2024):** As the inaugural framework to integrate continual learning with unlearning, CLMUL originally relies on a centralized architecture and buffer-based replay. To ensure a rigorous comparison with our decentralized setting, we transition its implementation into a distributed formulation and maintain a local buffer size of 200 samples per agent.
- **UG-CLU (Huang et al., 2025):** A unified gradient-based framework for task-agnostic learning–unlearning sequences. Its objective combines acquisition, unlearning, preservation, and modulation with a retained-manifold constraint for the stability–plasticity trade-off.

Provenance and buffers. UniCLUN is the label in the experimental record, but the available sources do not establish its bibliographic or implementation mapping. The records separately name a local *CLMUL* implementation with a 200-sample buffer per agent and do not establish that it maps to tabled UniCLUN. Methods requiring retained or target-class historical data use a common buffer whose size, composition, and sampling rule are otherwise unspecified; we therefore do not transfer the 200-sample value, procedure, or citation to another baseline.

C.2 Topology-aware Analysis

As in the main experiments, communication uses a symmetric doubly stochastic mixing matrix. To examine the relationship between communication topology and empirical performance, we evaluate Dec-UnHNet with Random, Ring, and Star graphs. The Random graph follows the Erdős–Rényi

construction used in the main experiments. In the Ring topology, each agent communicates with its two immediate neighbors, whereas the Star topology routes communication through a central agent. Table 3 shows that the Random topology attains the highest retained-class accuracy at all reported stages. The lower accuracies under Ring and Star are consistent with topology-dependent consensus drift, although these results do not isolate connectivity as the sole cause.

Table 3: Topology-aware analysis on CIFAR-100 with $Dir(\lambda = 0.5)$.

Method	$cl_{0:9}$ $r\%$	ul_2 $r/u\%$	$cl_{10:19}$ $r\%$	ul_0 $r/u\%$	$cl_{20:29}$ $r\%$	ul_{29} $r/u\%$
Star	41.92	31.07/ 0.00	34.39	33.76/0.00	30.66	25.82 / 0.00
Ring	49.36	36.43/ 0.00	33.29	30.01/0.00	27.12	27.09 / 0.00
Random	69.68	60.97/0.00	40.88	39.22/0.00	38.35	36.29/0.00

C.3 Legacy Task-Level Diagnostic

We further evaluate task-level unlearning to demonstrate the flexibility of *Dec-UnHNet*across class- and task-level targets. We add LSF and CLPU-DER++, which are designed for task-level unlearning; UniCLUN and UG-CLU also support this setting. As shown in Table 4, UniCLUN retains target-task accuracies of 16.83% and 14.34%, indicating incomplete unlearning in these runs. In contrast, *Dec-UnHNet*achieves 0.00% target-task accuracy at both unlearning steps while maintaining higher retained-task accuracy. For task-level unlearning, *Dec-UnHNet*discards the corresponding task descriptor and maps the task-specific parameters to random weights while retaining the other task mappings.

Table 4: Task-level accuracy on CIFAR-100 with $Dir(\lambda = 0.5)$ for ① $cl_{0:19} \rightarrow$ ② $cl_{20:39} \rightarrow$ ③ $ul_{0:19} \rightarrow$ ④ $cl_{40:59} \rightarrow$ ⑤ $cl_{60:79} \rightarrow$ ⑥ $ul_{40:59}$. Unlearning columns report r/u (%).

Method	$cl_{0:19}$ $r\%$	$cl_{20:39}$ $r\%$	$ul_{0:19}$ $r/u\%$	$cl_{40:59}$ $r\%$	$cl_{60:79}$ $r\%$	$ul_{40:59}$ $r/u\%$
LSF	33.41	25.97	8.86/0.01	16.29	7.92	2.04/0.06
CLPU-DER++	52.20	28.05	28.29/0.26	27.89	21.34	15.02/0.01
UniCLUN	52.20	44.33	23.50/16.83	35.89	39.45	25.76/14.34
UG-CLU	55.09	37.82	12.19/0.00	34.00	34.87	27.34/0.00
Ours	**55.42**	**48.33**	**48.23/0.00**	**42.23**	**39.97**	**39.57/0.00**